Text–to–Video: Image Semantics and NLP

Katharina Schwarz

Dissertation • Tübingen 2018

Text–to–Video: Image Semantics and NLP

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Katharina Schwarz
aus Augsburg

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	23.07.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Hendrik P. A. Lensch
2. Berichterstatter:	Prof. Dr. Andreas Schilling

"Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?"

J. R. R. Tolkien, The Hobbit

"Stories live to be told to others."

Dan P. McAdams

Abstract

When aiming at automatically translating an arbitrary text into a visual story, the main challenge consists in finding a semantically close visual representation whereby the displayed meaning should remain the same as in the given text. Besides, the appearance of an image itself largely influences how its meaningful information is transported towards an observer. This thesis now demonstrates that investigating in both, *image semantics* as well as the *semantic relatedness between visual and textual sources* enables us to tackle the challenging semantic gap and to find a semantically close translation from natural language to a corresponding visual representation.

Within the last years, social networking became of high interest leading to an enormous and still increasing amount of online available data. Photo sharing sites like Flickr allow users to associate textual information with their uploaded imagery. Thus, this thesis exploits this huge knowledge source of user generated data providing initial links between images and words, and other meaningful data.

In order to approach visual semantics, this work presents various methods to analyze the visual structure as well as the appearance of images in terms of meaningful similarities, aesthetic appeal, and emotional effect towards an observer. In detail, our GPU-based approach efficiently finds visual similarities between images in large datasets across visual domains and identifies various meanings for ambiguous words exploring similarity in online search results. Further, we investigate in the highly subjective aesthetic appeal of images and make use of deep learning to directly learn aesthetic rankings from a broad diversity of user reactions in social online behavior. To gain even deeper insights into the influence of visual appearance towards an observer, we explore how simple image processing is capable of actually changing the emotional perception and derive a simple but effective image filter.

To identify meaningful connections between written text and visual representations, we employ methods from Natural Language Processing (NLP). Extensive textual processing allows us to create semantically relevant illustrations for simple text elements as well as complete storylines. More precisely, we present an approach that resolves dependencies in textual descriptions to arrange 3D models correctly. Further, we develop a method that finds semantically relevant illustrations to texts of different types based on a novel hierarchical querying algorithm. Finally, we present an optimization based framework that is capable of not only generating semantically relevant but also visually coherent picture stories in different styles.

Kurzfassung

Bei der automatischen Umwandlung eines beliebigen Textes in eine visuelle Geschichte, besteht die größte Herausforderung darin eine semantisch passende visuelle Darstellung zu finden. Dabei sollte die Bedeutung der Darstellung dem vorgegebenen Text entsprechen. Darüber hinaus hat die Erscheinung eines Bildes einen großen Einfluß darauf, wie seine bedeutungsvollen Inhalte auf einen Betrachter übertragen werden. Diese Dissertation zeigt, dass die Erforschung sowohl der *Bildsemantik* als auch der *semantischen Verbindung zwischen visuellen und textuellen Quellen* es ermöglicht, die anspruchsvolle semantische Lücke zu schließen und eine semantisch nahe Übersetzung von natürlicher Sprache in eine entsprechend sinngemäße visuelle Darstellung zu finden.

Des Weiteren gewann die soziale Vernetzung in den letzten Jahren zunehmend an Bedeutung, was zu einer enormen und immer noch wachsenden Menge an online verfügbaren Daten geführt hat. Foto-Sharing-Websites wie Flickr ermöglichen es Benutzern, Textinformationen mit ihren hochgeladenen Bildern zu verknüpfen. Die vorliegende Arbeit nutzt die enorme Wissensquelle von benutzergenerierten Daten welche erste Verbindungen zwischen Bildern und Wörtern sowie anderen aussagekräftigen Daten zur Verfügung stellt.

Zur Erforschung der visuellen Semantik stellt diese Arbeit unterschiedliche Methoden vor, um die visuelle Struktur sowie die Wirkung von Bildern in Bezug auf bedeutungsvolle Ähnlichkeiten, ästhetische Erscheinung und emotionalem Einfluss auf einen Beobachter zu analysieren. Genauer gesagt, findet unser GPU-basierter Ansatz effizient visuelle Ähnlichkeiten zwischen Bildern in großen Datenmengen quer über visuelle Domänen hinweg und identifiziert verschiedene Bedeutungen für mehrdeutige Wörter durch die Erforschung von Ähnlichkeiten in Online-Suchergebnissen. Des Weiteren wird die höchst subjektive ästhetische Anziehungskraft von Bildern untersucht und "deep learning" genutzt, um direkt ästhetische Einordnungen aus einer breiten Vielfalt von Benutzerreaktionen im sozialen Online-Verhalten zu lernen. Um noch tiefere Erkenntnisse über den Einfluss des visuellen Erscheinungsbildes auf einen Betrachter zu gewinnen, wird erforscht, wie alleinig einfache Bildverarbeitung in der Lage ist, tatsächlich die emotionale Wahrnehmung zu verändern und ein einfacher aber wirkungsvoller Bildfilter davon abgeleitet werden kann.

Um bedeutungserhaltende Verbindungen zwischen geschriebenem Text und visueller Darstellung zu ermitteln, werden Methoden des "Natural Language Processing (NLP)" verwendet, die der Verarbeitung natürlicher Sprache dienen. Der Einsatz umfangreicher Textverarbeitung ermöglicht es, semantisch relevante Illustrationen für einfache Textteile sowie für komplette Handlungsstränge zu erzeugen. Im Detail wird ein Ansatz vorgestellt, der Abhängigkeiten in Textbeschreibungen auflöst, um 3D-Modelle korrekt anzuordnen. Des Weiteren wird eine Methode entwickelt die, basierend auf einem neuen hierarchischen Such-Anfrage Algorithmus, semantisch relevante Illustrationen zu Texten verschiedener Art findet. Schließlich wird ein optimierungsbasiertes Framework vorgestellt, das nicht nur semantisch relevante, sondern auch visuell kohärente Bildgeschichten in verschiedenen Bildstilen erzeugen kann.

Acknowledgments

First of all, I kindly thank my supervisor Prof. Dr. Hendrik P. A. Lensch for his enthusiasm, his guidance, and his valuable advice. As such an open-minded person with a passion for visionary topics, he has not only made this work possible but, above all, an intriguing experience. Further, I want to express my sincere gratitude to Prof. Dr. Andreas Schilling for supporting this thesis as a second supervisor and for his always friendly attitude.

Many thanks to the entire graphics group at the University of Tübingen for inspiring conversations, intensive discussions, and times of great fun. Even in most stressful times, there has always been a place for an enjoyable cup of coffee. Thanks to the current members, in particular, Fabian Groh, Patrick Wieschollek, Dennis Bukenberger, Benjamin Resch, Jieen Chen, Sebastian Herholz, and Raphael Braun. Thanks also to the former members of the graphics group, especially, Boris Ajdin, Beatriz Trinchão Andrade, Manuel Finckh, Christoph Rößing, Jian Wei, Takahiro Okabe, as well as people close-by the group, in particular, Manuel Lange and Andreas Karge. Special thanks to Christian Fuchs for thorough proofreading of this thesis as well as many fun drives between Ulm and Tübingen. Warmest thanks to the short and long time secretaries, especially, Manuela di Paolo for many pleasant chats. Grateful thanks to all my collaborators for having provided an important contribution to the development of several projects, in particular, apart from the ones already mentioned, Tobias Häußler and Christian Spika. I also want to thank Prof. Dr. Michael Goesele for supporting my first research project with very helpful advice. Very special thanks to Jan Fröhlich for providing in-depth understanding on color and Andreas Minuth for supporting us with his expertise in color grading for the EmoTune project. Thanks also to all the great people I have met on conferences, especially, Iraklis Varlamis who was a great pleasure to meet on my very first conference.

I deeply want to express an enormous thanks to Tamara L. Berg for being such a great mentor during my research stay at the University of North Carolina as well as while finishing our joint project. It was a great pleasure to work with her and I gained a lot out of this collaboration. Many thanks also to all the wonderful people I met at the UNC for giving me such a warmish welcome and a great research experience in Chapel Hill, among others, Alex Berg, Jan-Michael Frahm, Hadi Kiapour, Vicente Ordonez, Johannes Schönberger, True Price, Sirion Vittayakorn, Xufeng Han, Hyo Jin Kim, Jared Heinly, Wei Liu, and Yipin Zhou.

With great gratitude I would like to thank all the Flickr users for supporting this work with their photographs and the AMT workers for participating in our user studies. This work could not have been done without them.

Most heartfelt thanks to my family and friends for steady support, listening, balancing, fun, and all the distraction. Especially, I want to thank my parents Elisabeth and Gerd for letting me grow up in such a creative environment surrounded by music and art, for all the caring, for always believing in me, and for continuously supporting me through my life.

My deepest thanks go to Manuel for his availability and encouragement at all times, for dealing with my ups and downs, and for supporting me with all his strength. His great patience as well as many intriguing discussions have provided highly motivational support during the development of this work. Thanks also for thorough proofreading and a lot of “thumb pressing”.

These last years have been a pretty awesome and extremely valuable time!

Contents

List of Figures

List of Tables

1 Introduction

Images are the core element in *computer graphics* and *computer vision*. Whereas the graphics community aims for synthesizing demanding visual representations, the field of computer vision focuses on analyzing images to gain deeper knowledge and derive relevant models. In both fields, the semantics of an image plays an important role as it is capable of transporting features or models to a higher level of understanding and provides a stronger relation between images and their meaning in the real world.

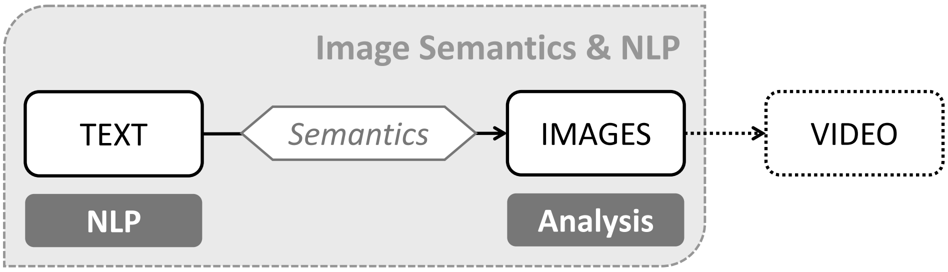


Figure 1.1: *Text–to–Video (TTV). We propose to automatically generate a video based on natural language input. Thereby, the focus of this thesis lies in building a connection from an arbitrary written text to a semantically meaningful set of images that tell the given story visually. Employing methods from NLP and closely analyzing the visual appearance of images to exploit the strength of visual semantics support this aim.*

This thesis explores *image semantics* in terms of analyzing visual appearance and its connection to human beings and, therefore, employs methods from both fields, i.e., computer graphics and computer vision. Additionally, this work incorporates techniques from *Natural Language Processing (NLP)* to assemble relevant visual representations to a given textual description connecting text and images on a strong semantic level. In the long term, we propose the automatic language based generation of dynamic videos, more precisely, a standalone *Text–to–Video (TTV)* system. Figure 1.1 indicates the primary focus of the present thesis within the TTV pipeline. Overall, this chapter motivates the idea behind this thesis and outlines the structure as well as the main contributions of this work.

1.1 Motivation

In all times, storytelling has been an essential way to connect people by sharing their thoughts or personal experiences, to communicate their knowledge, or simply to entertain each other. At the same time, storytelling influences the narrator's own personality. From a psychological point of view, McAdams identified a strong relationship between the stories a person tells and the development of this person's identity and argued that people actually define their personalities through their narratives [McA93]. Further, McAdams outlined that the telling of stories as well as their comprehension are primary intentions of humans whereby the development of the according mental conditions starts from early childhood [McA01]. Overall, one can say that storytelling is highly valuable for the personality of individuals as well as the social relationships between people. Thus, telling a story by natural language is a process everybody gets trained for all his life.

However, telling a story by images is significantly harder. The generation of images, picture stories, or even videos is a time consuming process and the outcome largely depends on skill. Knowledge about certain artistic or compositional rules can aid in the process of creating such a visual story but, typically, a lot of experience and talent are necessary to depict the intended content and, especially, to establish a certain atmosphere or transport desired feelings. Artists often even exaggerate their stylistic means to express their intentions. Certainly, already a single picture can tell a big and intense story. Overall, pictures have a similar expressive range as natural language with regard to describing objects, actions, or evoking specific emotions.

Aiming to connect textual storytelling with the expressive range of pictures by automatic means and to explore the intriguing task of visual storytelling, several aspects and challenges have inspired the development of this thesis.

Images and Natural Language. Images and language are naturally connected. Whenever we see an image, it triggers thoughts, inner ratings, or even complete stories in our minds. The other way around, any text we read or listen to produces a mental image to visually reflect the words. Further, words can actually aid in understanding the content, idea, or intention of an image, for example, providing an additional description along with a painting or a photograph. Similarly, images can aid in understanding the story beyond words, for example, as visualizations of scene descriptions in the form of a storyboard or to clarify instructions in an operation manual. Thereby, such visualizations can resolve textual ambiguities.

Social Online Data. Nowadays, huge amounts of data connecting images with words exist online. With their strong intention to tell their stories, people largely exploit the immense capabilities of social media arisen during the last years to communicate information instantly and everywhere. Social online systems together

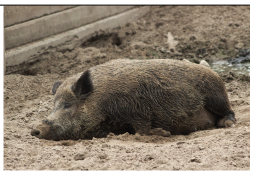
wanakatree, NewZealand, **landscape**, tree, lake	isaac, leão, fotos, flickr	**landscape**, Gates of the Arctic National Park and Preserve, public, domain	fireworks, paysage, paisaje, **landscape**
sky	blackandwhite, monochrome, sea, water, outdoor, **landscape**, shore, seaside, ocean	outdoor, mountain, sky, hill	minimalism, black background, monochrome, field, outdoor
			
dmc, falls, lumix, lx7, panasonic, tews, water, Hamilton, Ontario, Canada, sns, hdr, panasonic lumix dmc-lx7, **landscape**, waterfall	Panasonic DMC-FZ200, mount sefton, Mount cook national park, Snow, Mountains, Southern Alps, G, geo tagged, free photos	Badlands National Park, weather, **landscapes**, scenery, rainbow, sky, storm	
outdoor	outdoor, **landscape**, mountain, road, mountain peak, hill, mountainside	serene, outdoor, **landscape**, mountain, sunset, texture, sea	animal, bear, **landscape**, mammal

Figure 1.2: *Flickr images with user associated tags (black) and machine tags generated by Flickr (gray). The images were retrieved for the query "landscape" (bold font) and establish different atmospheres. Noisy human or machine tags can lead to wrong image results (last column).*

with mobile applications allow people to share personal data like thoughts, stories, but also images with the world. The resulting high interest in social media led to an enormous and still increasing amount of image data in the web.

Meanwhile, the photo sharing site Flickr [1] states to be "home to 13 billion photos and 2 million groups" (as of April 20th 2017). The photos in such online collections are taken by a wide variety of people with different skills and intentions. Thus, this tremendous amount of pictures comes along with a huge range of displayed content, a broad diversity with regard to the appearance of pictures, and certainly also differences in terms of image quality, e.g., low-level quality like resolution, or aesthetic quality. Further, Flickr allows users to attach meta information like a title, a short description, camera details, or tags to their uploaded photographs and also provides internal mechanisms to associate machine generated tags to images. Such associated meta information enables to search images in Flickr by a textual query.

In particular, the idea of tags is to provide concise additional information about the picture allowing for a deeper understanding of the image itself, the situation

[1] www.flickr.com

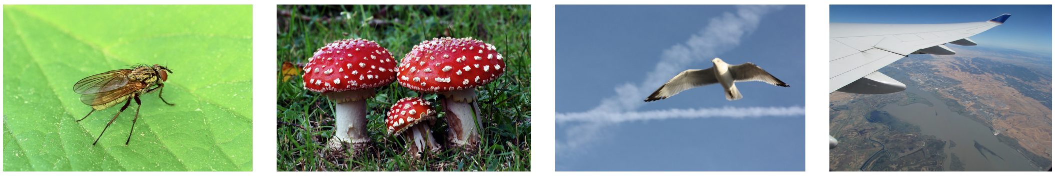


Figure 1.3: *Flickr images for the query "fly" picturing various meanings of this ambiguous word. Left to right: Insect, fly agarics, a flying bird, and the action of flying in an airplane.*

in which it was taken, or the location. Figure 1.2 shows some outdoor images we retrieved from Flickr by querying for the word "landscape" and the tags associated by users as well as machine tags attached by automatic means. As can be seen, human tags often comprise tags about the displayed scene as well as additional background information, e.g., about the location where the picture was taken like "New Zealand" or "Hamilton, Ontario, Canada" (first column). Machine tags rather focus on a broader kind of information about the visual content like the tag "outdoor" which has been automatically created for most of these images. However, machine tags can enhance search mechanisms, as in this example, where the tag "landscape" was attached automatically for the two images in the second column. However, tags are not always reliable. Often, there is a certain amount of noise within the tags leading to wrongly returned images for a textual query. Thereby, wrongly added machine tags as well as the high subjectivity of the human tagging process itself can cause such noise which consequently reveals unexpected image results. For example, the images of the fireworks and the pig are not what we actually consider a landscape (last column). Overall, the images in Figure 1.2 indicate the broad diversity inherent in online data. The correctly returned pictures present different types of landscapes in varying appearances and establish different atmospheres. For example, in the third column a rather calm and sunny landscape with blue sky and a couple of clouds (top) is contrasted to an image establishing a stronger tension with the rainbow under the clouded sky (bottom). Additionally, the images depict various styles, for example, the first image in the top row is mainly kept in gray tones, the second one is a black and white picture, or the first picture in the lower row was taken under long exposure.

Overall, the attached tags might not always be completely reliable and quality differs largely, both making the data rather noisy. However, images from online photo collections together with their associated meta-data provide a big source of knowledge connecting words and images while offering a broad visual diversity.

Ambiguous Words. However, when querying for a specific word, one has to take into account that words can have different meanings. Figure 1.3 illustrates several senses of the word "fly". Typically, when such ambiguous words are considered in isolation, their actual meaning can hardly be grasped. Thus, search results showing different meanings are returned, as in the example which comprises a picture of

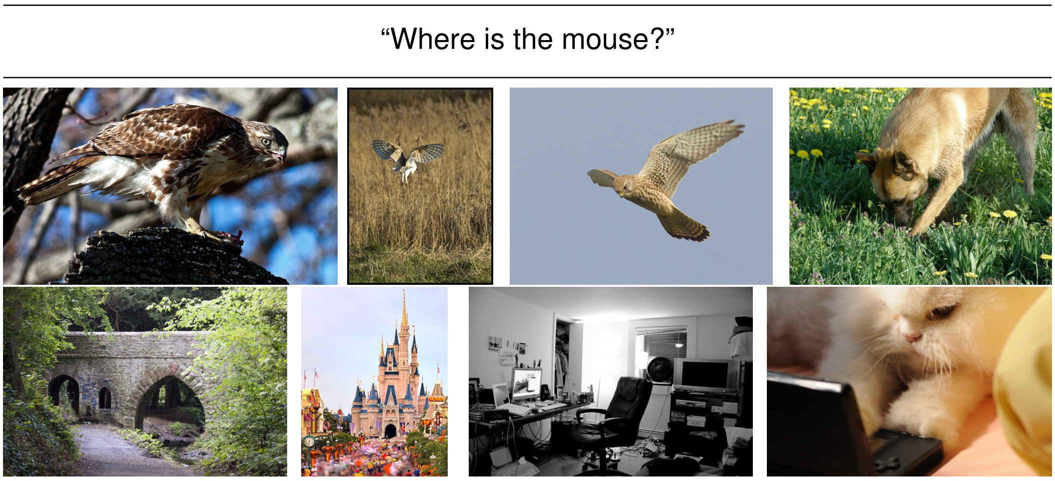


Figure 1.4: *Example showing the complexity of ambiguity on the textual and the visual level. For the question (top) various valid situations are visualized by the images (bottom). However, whereas the question does not provide enough context to identify which kind of “mouse” could be meant, e.g., a real mouse, mickey mouse, or a PC mouse, it is rather obvious in most of the images. Only the last two images (last row) leave a visual ambiguity whether a real mouse or a computer related one is meant.*

a fly as an insect, an image of fly agarics, and two pictures of the action of flying. Thus, when searching for a specific visualization, more information can help to identify the particular meaning of such a word, e.g., “fly, insect”.

Semantic Gap. The mentioned ambiguity of words makes it even more challenging to find a perfectly matching image to a textual description. Typically, for such words, their particular meaning needs to be grasped from the context. Unfortunately, it is sometimes hardly possible to derive the correct meaning from a sentence or a question if not enough additional information is provided. This complexity of meaning is indicated in Figure 1.4. The images show a diversity of possible visualizations for the question “Where is the mouse?”. Even without actually picturing a mouse, most of the images clearly indicate what kind of mouse is meant, e.g., the animals rather search for a real mouse, Cinderellas castle links to mickey mouse and the office guides one’s thoughts towards a computer mouse. However, the full office could also hide a real mouse and, contrarily, the cat could also search for the mouse pointer on the display. Anyway, from the textual question alone, the correct meaning of the word “mouse” itself can not be grasped. In this case, more context would be needed to identify for what kind of mouse the questioner is looking for. Overall, a lot of analysis needs to be done on both, the textual and the visual side, to bridge the semantic gap and tackle the complexity of semantics. More details on semantics and ambiguity will be discussed in Chapter 2.

Mapping Text to a Visualization – Semantic Challenge. As previously mentioned, the generation of a visual story, e.g., a single image, a picture story, or even a complete video, is very difficult as the process is time consuming and the outcome largely depends on skill. At the same time, images provide an enormous expressive range. Thus, to tackle the challenge of visual storytelling we propose to start from a textual basis and consider the creation of an according visual representation not as a generative process, but as a retrieval problem. Therefore, we aim at exploiting the mentioned tremendous amount of online available visual data associated with text. In other words, we investigate in finding a semantically close translation from natural language to a visually meaningful representation.

However, as soon as we read a sentence or a small piece of text, the mental image we build in our mind is influenced by our personal experience and reflects our subjective interpretation. The perfect mapping from our mental image to a visual representation would result in an image that shows the identical situation as the mental image and triggers the same feelings. Additionally, such mental images differ largely between people making it impossible to picture everyone's imagination. When it comes to longer texts or even complete stories, the problem is even more challenging as we build complex visual stories in our minds and, thus, a generated picture story or video should reflect all the content and atmosphere, in other words, transport all the semantics of the initial text.

Overall, when aiming to automatically translate natural language into a corresponding visual representation, there exist several main issues:

- *Semantic gap:* Textual description and visual representation should comprise the same content and transport the same meaning. As it is hardly possible for both modalities to present the completely identical meaning, a semantically close translation needs to be found to bridge this semantic gap.
- *Textual ambiguities:* For words with different senses, their particular meaning needs to be identified within its context to enable a correct translation making it even more challenging to find a suitable image to a textual description.
- *Visual semantics:* Images themselves depict meaningful representations telling their own story. Together with their visual appearance, images are capable of establishing a certain atmosphere and trigger a specific feeling.
- *Subjectivity:* The semantic interpretation of meaning is highly subjective and can differ largely between people. Thus, humans need to be integrated into the process and data as well as ratings from a broad and highly diverse crowd are necessary.

Idea of Thesis. The present thesis now addresses those mentioned aspects and challenges and investigates in the intriguing semantic connection between a textual description and a visual representation as well as the expressive visual semantics of

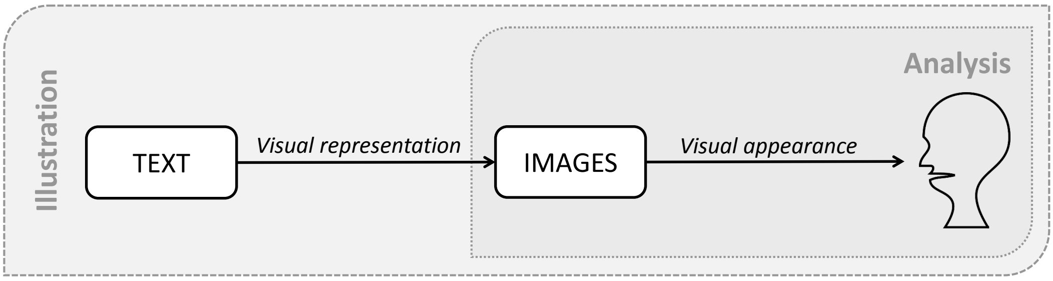


Figure 1.5: *Connection of two main parts. Analyzing the visual appearance of images and their effect towards the perception of an observer leads to insights about image semantics. Based on those insights, pleasing visual representations to a given text can be found in the illustration process.*

images. Generating such an automatic translation of an arbitrary text into a visual representation is especially challenging as it is often not possible to completely bridge the semantic gap. At the same time, the appearance of an image itself largely influences how its meaningful information is transported towards the observer. As each image is capable of telling its own story, when combining multiple pictures into a visual story, their visual coherency should be considered to ensure the flow and transport a coherent mood. Overall, the highly subjective nature of semantics increases the challenge of this task. Thus, we use methods of Natural Language Processing and work on closely analyzing the visual semantics of images. Thereby, a key element is including multiple knowledge sources both in form of texts as well as online collected image databases, but also incorporating human beings to approach the subjective nature of semantics. An overview of the methods developed and presented during this work is given in the following section.

1.2 Overview of Core Elements

As we aim at finding a suitable visual illustration to a given textual description, we identify three core elements: the written input *text*, reasonable *images*, and the *observer* to evaluate the result. To approach the connection between them, this thesis is divided into two main parts, the *analysis* of images with a strong connection to human beings (Chapter 5) and the *illustration* of a text with visual data (Chapter 6). The connection between the mentioned terms is illustrated in Figure 1.5. Before focusing on the actual illustration of textual sources, we investigate in analyzing the visual appearance of images and their effect towards the perception of an observer (Chapter 5). Thereby, we gain insights on the connection between the visual appearance and the semantics of images. Based on those insights, we are not only able to find semantically close visual representations to a given textual description, but also visually pleasing ones (Chapter 6).

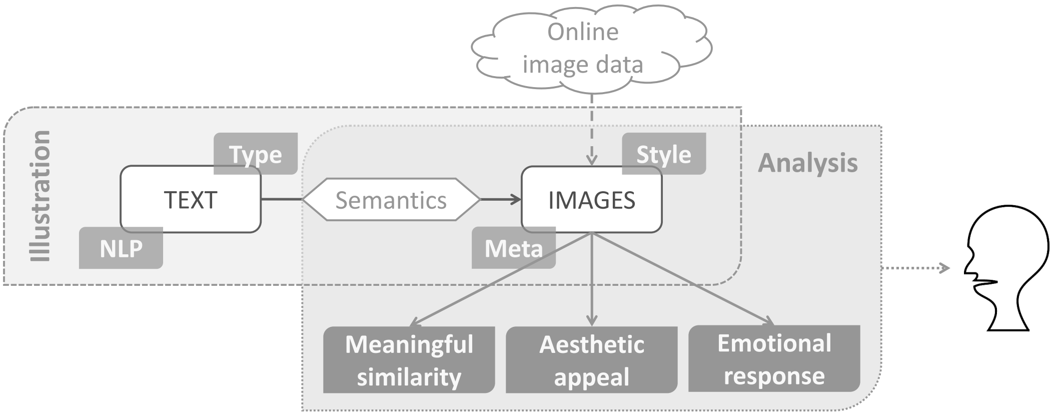


Figure 1.6: *Overview of the core elements. The first part analyzes the appearance of images to identify meanings from visual similarity and explore aesthetic appeal as well as emotional response of an observer. The second part mainly investigates in finding a semantically meaningful illustration to a written text using methods of NLP. We consider different text types and various styles of illustration. Both parts rely on images associated with suitable meta data which we retrieve from online photo collections.*

In detail, Figure 1.6 gives an overview over the core elements that are explored within the mentioned two main parts of this thesis. Both parts build on visual data which is retrieved from online repositories or online photo collections and contains associated meta information. First, in Chapter 5, we *analyze the visual appearance of images* as, additionally to specific content displayed in an image, its visual appearance plays an important role on how its meaningful visual information is transported to a person. We raise the following questions:

- Can visual similarity aid in resolving textual ambiguity?
- Do people favor images due to their aesthetic appeal?
- Can global modifications of the visual appearance influence the emotional response of an observer?

In the first part, we will address those issues. We analyze the visual structure of images to identify similar meanings based on visual similarities as well as user-associated meta data (Section 5.1). Further, we analyze the global appearance of images. Therefore, we explore aesthetic appeal from a broad diversity of user ratings (Section 5.2) and directly investigate in the emotional effect of the global visual appearance of an image towards an observer (Section 5.3).

The second part is described in Chapter 6 and addresses the task of *illustrating text with images* or, in other words, finding a semantically meaningful visual representation to a given written text. We will present work on different types of texts, different visual representations, and consider various visual styles for the illustration task. Regardless of their type, the given input texts are analyzed using methods of NLP. As previously mentioned, expressing mental images visually is a time-consuming and challenging process. Thus, we demonstrate the effectiveness of employing natural language to build a visual representation in the form of a virtual environment (Section 6.1). Then, aiming at simplifying the process of telling a picture story but without the need of creating the images, we shift the focus to images as a visual representation. To approach the task of illustrating text with meaningful pictures, we first present a system to illustrate short texts with relevant images (Section 6.2). To even illustrate a complete story, it is crucial to ensure visual coherence along the storyline to maintain the mood of the story. Thus, we finally present a framework that optimizes over semantic relevance as well as visual coherence along the storyline (Section 6.3).

1.3 Main Contributions

The main objective of this work is the exploration of *image semantics* as well as the *semantic relatedness between visual and textual sources*. Thereby, this thesis presents various ways to illustrate different types of written text with an adequate visual representation leading to a final illustration that is semantically relevant as well as visually coherent. As the semantic expression of an image largely depends on its visual appearance, this thesis further analyzes images in terms of similarity, aesthetic appeal and evoked affect. In detail, the main contributions are:

- *Semantics*. A detailed discussion on semantics in the context of natural language, images, and online media as well as related issues on subjectivity and evaluation is given in Chapter 2. Further, throughout this thesis, a 3-fold exploration of semantics is provided: Text–based, image–based, and connection in between as well as perception-based evaluation of the latter two.

- *Efficient visual similarities.* A GPU-based approach is presented to efficiently find visual similarities between images across modalities in large image data bases. Its application to image responses from online search engines enables the identification of various senses of ambiguous words. [SHL12]

- *Learning aesthetics from diverse crowd.* Predicting the aesthetic appeal of an image is approached employing deep learning. To learn from a huge diversity of people, social online behavior is examined and images with adequate meta-data are collected to derive a meaningful score of aesthetics. [SWL18]

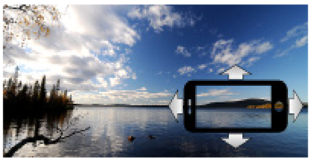
- *Emotional image modification.* A simple image filter composed of basic global image modifications that can change a viewer's emotional perception is derived from collecting empirical data on images associated with emotion labels and analyzing the valence ratings of different modifications. [SFFL17]

- *Textual analysis.* A two-sided textual analysis (text and image-tags) with methods of NLP is performed to bridge the semantic gap. [SRC$^+$10, SBL17] Simple descriptive up to highly abstract creative texts are tackled with different kinds of visualizations. Text decomposition is performed based on the type of the text and the target type of the visualization. [SRC$^+$10, SSDL11, SBL17]

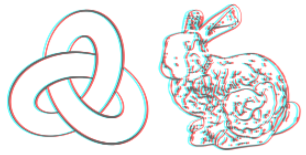
- *Image search for text snippets with high precision.* A hierarchical querying approach to construct meaningful textual search terms and retrieve images from online photo collections with high semantic precision is developed. [SRC$^+$10]

- *Auto-Illustration in visual styles.* A novel framework combining textual semantic search, content similarity, style classification, and discrete optimization to generate picture story illustrations with controllable visual style is presented. Our approach allows us to identify the dominant style of a text. [SBL17]

The following list comprises the publications that are relevant for this thesis and that contain large parts of the above mentioned contributions [SRC$^+$10, SSDL11, SHL12, SBL17, SFFL17, SWL18] as well as additional ones that have arisen from collaboration [BSGL15, BSL17]:

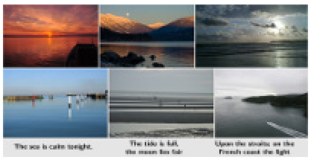


Will People Like Your Image? Learning the Aesthetic Space. Katharina Schwarz, Patrick Wieschollek, Hendrik P. A. Lensch. *Winter Conference on Applications of Computer Vision (WACV), 2018.* [SWL18]



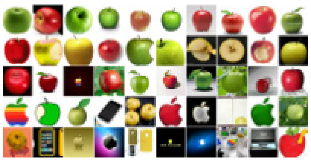
Stereo-consistent Contours in Object Space. Dennis R. Bukenberger, Katharina Schwarz, Hendrik P. A. Lensch. *Computer Graphics Forum (CGF), 2017.* [BSL17]

EmoTune - Changing Emotional Response to Images. Katharina Schwarz, Christian Fuchs, Manuel Finckh, Hendrik P. A. Lensch. *Color and Imaging Conference (CIC), 2017.* **Best Student Paper Award.** [SFFL17]


The sea is calm tonight.
The tide is full, the moon lies fair
Upon the straits; on the French coast the light

Auto-Illustrating Poems and Songs with Style. Katharina Schwarz, Tamara L. Berg, Hendrik P. A. Lensch. *Asian Conference on Computer Vision (ACCV), 2016.* [SBL17]

Rotoscoping on Stereoscopic Images and Videos. Dennis R. Bukenberger, Katharina Schwarz, Fabian Groh, Hendrik P. A. Lensch. *Vision, Modeling and Visualization (VMV), 2015.* [BSGL15]



An Efficient Parallel Strategy for Matching Visual Self-Similarities in Large Image Databases. Katharina Schwarz, Tobias Häußler, Hendrik P. A. Lensch. *European Conference on Computer Vision (ECCV) Ws, 2012.* [SHL12]

AVDT - Automatic Visualization of Descriptive Texts. Christian Spika, Katharina Schwarz, Holger Dammertz, Hendrik P. A. Lensch. *Vision, Modeling and Visualization (VMV), 2011.* [SSDL11]

Text-to-Video: Story Illustration from Online Photo Collections. Katharina Schwarz, Pavel Rojtberg, Joachim Caspar, Iryna Gurevych, Michael Goesele, Hendrik P. A. Lensch. *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), 2010.* **Best Paper Award.** [SRC$^+$10]

1.4 Shared Contributions

As mentioned, several publications are relevant for this present thesis. They are co-authored by external contributors. Their contribution is listed in the following:

- The pipeline in the AVDT project [SSDL11] has been implemented by Christian Spika as part of his MSc thesis.
- The efficient version of the self-similarity descriptor on GPU [SHL12] has been implemented by Tobias Häußler.
- Several of the modifiers in the EmoTune project [SFFL17] have been implemented by Christian Fuchs. The implementation of the edge-avoiding wavelets has been provided by Manuel Finckh.
- The deep learning network for the aesthetics prediction [SWL18] has been provided by Patrick Wieschollek.

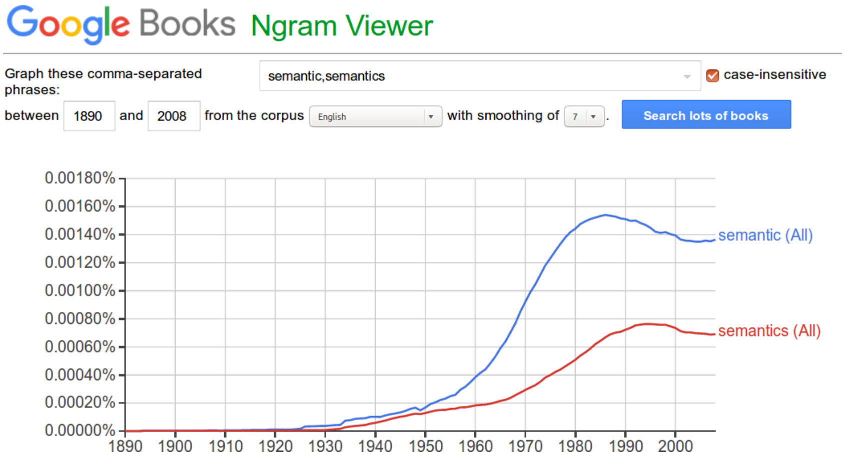
Some of the ideas were developed in shared discussions. Apart from this, the work presented in this thesis is sole work from Katharina Schwarz.

1.5 Chapter Outline

The present thesis begins with a fundamental discussion of the term “semantics” in Chapter 2 exposing its relevance in the area of linguistics as well as the challenge of image semantics as a highly complex and subjective topic. As large parts of this work build on a textual basis, Chapter 3 provides background information on basics in linguistics focusing on Natural Language Processing as well as the handling of information in terms of extraction and retrieval. It is followed by a wide-ranging discussion of related work reviewing the benefits of large online photo collections in computer graphics and computer vision as well as previous work exploring the direct connection between images and text in Chapter 4. Then, the main work of this thesis is split into two parts: First, Chapter 5 presents work on image analysis comprising meaningful similarities, aesthetic appeal, and emotional response. Second, Chapter 6 approaches the challenging task of meaningful text illustration building upon the basics described in Chapter 3 as well as incorporating insights derived from the visual analysis in Chapter 5. Finally, Chapter 7 concludes this thesis summarizing the most important outcomes as well as relevant aspects on semantics and gives an outlook on possible extensions to this work.

2 Semantics

With his “Essai de sémantique” published in 1897, Bréal [Bré97] is widely considered as an important pioneer to our modern understanding of *semantics*. In his work, he motivated science to study language not only from the grammar side but to explore it “from the side on which it appeals to the mind” [BCP00] arguing for a meaningful view on language. For example, he investigated into the reasons that potentially change the meaning of words [Bré97]. However, during the last half century, the terms “semantic” and “semantics” became of increasing interest. Figure 2.1 indicates this trend as presented by the Google Books Ngram Viewer [1] which provides information about the occurrences of terms or phrases within a large amount of books over a specified period of time. As indicated by the Encyclopædia Britannica [Lep17], the term “semantics” is derived from the Greek verb sēmainō (“to mean” or “to signify”). Moreover, a general definition of semantics is similarly given by several other dictionaries as:



Graph these comma-separated phrases: semantic,semantics ☑ case-insensitive

between 1890 and 2008 from the corpus English with smoothing of 7 .

Figure 2.1: *Occurrence of terms “semantic” and “semantics” since 1890 as displayed by the Google Books Ngram Viewer*[1]*. Increasing interest can be recorded during the last half century.*

> *Semantics:* “Study of meaning” [Semndc, Semnda, Semnde]

Overall, semantics can describe the meaning of a variety of different sources, e.g., natural language, artificial language, images, and basically anything that transports

[1] `http://books.google.com/ngrams`

a higher meaning. However, comparing the meaning of such different sources directly leads to the challenging *semantic gap*. According to Hein [Hei10], the semantic gap can be defined as follows:

> *Semantic gap:* "the difference in meaning between constructs formed within different representation systems"

The different representation systems relating to semantics and considered during this thesis are mainly text, images, and the observer. Thus, this chapter outlines the relevance of semantics in language (Section 2.1) and images (Section 2.2). Thereby, data bases are presented, that have been created to connect words or images on a semantic level and support approaching semantics from a computational perspective. Further, the high subjectivity of semantics and, thus, the necessity of integrating humans during development as well as evaluation of machine-based algorithms is also discussed (Section 2.3).

2.1 Semantics in Linguistics

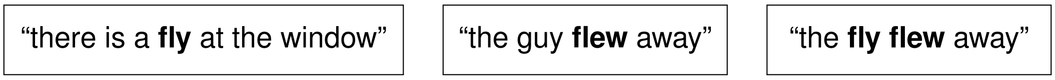
Since a long time, huge interest has been shown in studying language and its meaning. In particular, the field of linguistics is concerned with the scientific study of language [Cry90]. Crystal states that "languages have a great deal in common in the way they produce sounds, organize their grammars, and construct systems of meaning in words" [Cry01]. Thus, exploring the identification of such universal principals that define human language is part of the research of linguists. Thereby, semantics is of high relevance and, in context of linguistics, has been termed as the "branch of linguistics and logic concerned with meaning". As language is based on words and structures between them, corresponding definitions of semantics can be found:

- "Semantics is the study of meaning of words, constructions, and utterances." [MS01]
- "Semantics deals with the meaning of individual words and entire texts." [BG04]

Further, several dictionaries provide similar definitions when searching for the term "semantics". A few are itemized in the following:

- "The meaning of a word, phrase, or text." [Semndd]
- "the study of the meanings of words and phrases in language" [Semndb]
- "The meaning or the interpretation of a word, sentence, or other language form" [Semnde]

Overall, it is obvious that semantics in linguistics comprises the study of the meaning of language and its components. The according linguistic branch that deals with the study and analysis of word meanings as well as relations between them is



“there is a **fly** at the window” “the guy **flew** away” “the **fly flew** away”

Figure 2.2: *Ambiguity in context. Example sentences providing different context around the word “fly” as well as its conjugation “flew” leading to various meanings.*

called *lexical semantics* [JM08, Semndd]. Typically, a further division is made into studying the meaning of single words and studying how their individual meanings are combined into larger meaningful connections, e.g., sentences [MS01]. However, understanding the meaning of a certain word or phrase is strongly correlated to resolving its ambiguity as words may have a variety of meanings.

2.1.1 Ambiguity

In general, according to the Encyclopædia Britannica, ambiguity can be seen as the “use of words that allow alternative interpretations” [oEB98]. Typically, as ambiguity provides room for various meanings with a need to be resolved, it is commonly understood as a huge challenge. Further, such possibilities for alternative interpretations can make communication more difficult.

At the same time, ambiguity can provide a certain richness in language. Especially in poetry, ambiguity of language can increase the richness of a poem [oEB98]. In the context of poetry, William Empson [Emp57] motivated that ambiguity as carried by a poem can even enhance the reader’s understanding of the work when being forced to think of an individual interpretation.

However, the two previously mentioned divisions of studying semantics [MS01] can be transferred to the ambiguity of individual words as well as considering word ambiguity within a certain context which typically aids in resolving the individual meanings.

- *Ambiguous Words.* Single words can have various meanings. For example, “fly” can mean the insect, the action of an animal moving in the air, a person traveling by airplane, or moving quickly away from a place. However, when only considering the individual word, the specific meaning can hardly be identified.

- *Ambiguity in Context.* Context is needed to grasp the particular meaning of an ambiguous word. Some example sentences are given in Figure 2.2. Take for example “there is a fly at the window”. This statement clearly identifies “fly” as the insect. However, “the guy flew away” still allows different interpretations. Although considered in context, “flew” could still mean that the guy flew away with the airplane or that he ran away quickly. Considering the statement “the fly flew away”, people might easily connect “flew” with *moving in the air* and interpret the sentence as the insect moving away in the air.

Ambiguity of words or statements can already make communication between people rather complicated. However, for machines it is still a huge challenge to automatically resolve and understand such ambiguity in language. Thus, when processing natural language with a machine, most tasks somehow resolve a certain kind of ambiguity.

2.1.2 WordNet – A Semantic Hierarchy

In order to list such ambiguities or meanings of single words and provide semantic connections between them, *WordNet*, a lexical database for the English language, has been introduced by Miller [Mil95] and Fellbaum [Fel98]. In WordNet, nouns, verbs, adjectives, and adverbs are organized into *synsets* (synonym sets) and hierarchically connected by their semantic and lexical relations. Thus, WordNet provides a structure in which words are grouped by their meanings. Table 2.1 shows the synsets S provided by the online version of WordNet [Wor10][2] for the word “mouse”. Thereby, four meanings as a noun (n) and two senses as a verb (v) are found in the WordNet database. Overall, WordNet has been designed for the purpose of providing a lexical database that can be accessed by machines [Mil95] and has become a useful tool to process natural language, e.g., to identify similarities between words or to automatically grasp the meaning of a particular word.

Noun

- S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
- S: (n) shiner, black eye, **mouse** (a swollen bruise caused by a blow to the eye)
- S: (n) **mouse** (person who is quiet or timid)
- S: (n) **mouse**, computer mouse (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad) "a mouse takes much more room than a trackball"

Verb

- S: (v) sneak, **mouse**, creep, pussyfoot (to go stealthily or furtively) "..stead of sneaking around spying on the neighbor’s house"
- S: (v) **mouse** (manipulate the mouse of a computer)

Table 2.1: *WordNet results when browsing for “mouse”. Different synsets (S) are provided by the online version of WordNet [Wor10] [2] for “mouse” whereby meanings as noun (n) and verb (v) are listed (accessed August 3, 2017).*

[2] `http://wordnetweb.princeton.edu/perl/webwn`

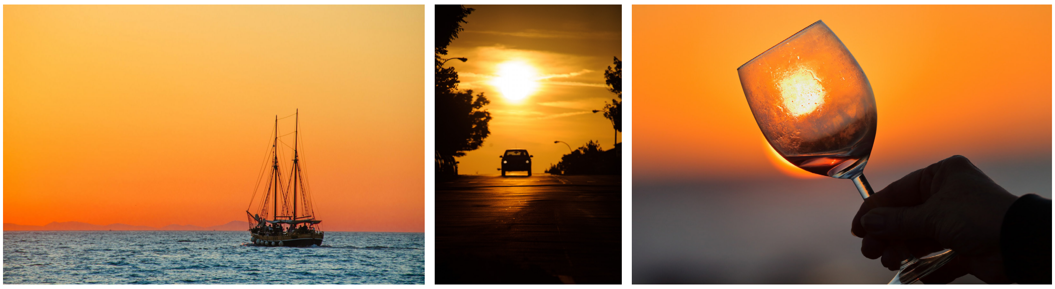


***Figure 2.3:** Images showing different content (boat, car, glass) in similar warmish sunset appearance. Thereby, the visual semantics depends on both, the depicted content and the established appearance.*

As large parts of this thesis build upon textual descriptions, identifying the semantics of text is a fundamental part of this work. Especially for the illustration task (Chapter 6), extracting relevant information from the textual description forms the basis to realize a semantically close translation from natural language to a visual representation. Therefore, basics on processing natural language will be presented in Chapter 3. As soon as the meaning of the text is identified, the semantics of images needs to be considered which will be described in the next section.

2.2 Image Semantics

When viewing an image, we initially process its displayed *content* based on factors like our personal experience. At the same time, the *appearance* of the image also has a large influence on how we feel about it and if we might like the image or rather not like it. Further, each image is capable of telling its own meaningful story. Overall, the semantics transported by an image can be highly complex. The previously mentioned definitions of semantics can be easily transferred to the particular case of images. Thus, *image semantics* can be generally denoted as follows:

Image Semantics: Meaning of an image.

Further, in linguistics (Section 2.1), semantics is more precisely specified as the meaning arising from the language components, e.g., words, phrases, or sentences. Therefrom, we can derive a more detailed definition of image semantics comprising the global picture as well as its visual components:

*Image semantics deals with the meaning of the visual components of an image
as well as with the meaning of the overall picture.*

Typically, an image shows a specific scene or situation, the content, and transports a certain mood with its appearance. Thereby, visual components can be content-based elements like objects or people, but can also be appearance related ones like color or the visual style of a picture. For example, in Figure 2.3, all images establish a similar warmish atmosphere while depicting different content, i.e., different items like a boat, a car, or a glass in front of a sunset. Besides, a human observer can still have a personal interpretation of the visual meaning. Thus, from a computational perspective, there exists a visual semantic gap between the feature-based representation of the image and its observer (Section 2.2.1).

However, based on the huge amount of visual online data, the generation of large-scale structured data supports the design of new algorithms that provide a better understanding of the pictured content (Section 2.2.2). As already mentioned, image semantics is not just about the content an image displays. Moreover, the appearance is at least similarly relevant for the visual semantics (Section 2.2.3). Thus, for an image I, its semantics $I_{sem} = semantics(I)$ can be described as composition from its displayed content $I_{cont} = content(I)$ and its appearance $I_{appear} = appearance(I)$:

$$semantics(I) \approx content(I) \circ appearance(I) \tag{2.1}$$

Both, the visual content I_{cont} as well as the overall appearance I_{appear} are highly relevant to represent the meaning of an image. Their particular contribution might differ between images but can also depend on each other (Section 2.2.3). Furthermore, the interpretation of the meaning of an image can differ largely between people making visual semantics a highly subjective topic (Section 2.2.4).

2.2.1 Visual Semantic Gap

Generally, in order to enable machine-based processing and understanding of images, the visual data needs to be quantified and transferred into a mathematical representation. Suitable features need to be extracted to describe the visual information. Thus, from a computational perspective, images can, for example, be represented by low-level image features or global image descriptors that describe that pictured visual data. In this context of feature extraction, the *semantic gap* has been denoted as:

- "gap between image features and the user" [SWS+00]
- "lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation" [DJLW08]
- "difficulty of determining a set of image features that correspond to a certain semantic meaning." [FK15]

In other words, these notations describe the semantic gap as the lack of correspondence between extracted visual information and the user. Thereby, the challenge of bridging the semantic gap is to identify those corresponding meaningful image features that correlate with what a user would see in an image. This problem occurs, for example, in the task of *content-based image retrieval (CBIR)* which deals with the retrieval of images based on their visual content. To further explore what people consider as relevant when they look at an image, previous work has investigated in observing what people describe about an image, in particular, what kind of tags people associate with certain imagery. An overview of related work dealing with CBIR and image tagging will be presented in Section 4.1.1.

Overall, in the particular case of images, the *visual semantic gap* can be considered as the lack of correspondence between the *visual representation* of an image and the *view of the observer*. Thereby, the huge amount of social online available data serves as an immense data source to analyze the relation between images and human behavior, e.g., by exploiting the textual meta data people attach to describe their pictures. However, if textual tags are considered in isolation, e.g., in text-based image search, word ambiguity can lead to broad and noisy visual search results. Thus, in this thesis, we will present work to refine online search results and identify various meanings based on visual similarities (Section 5.1).

Generally, grouping similar images into clusters that display a particular meaning enables the extraction of visual features that represent this specific meaning, e.g., the content pictured within a set. Thus, based on the enormous online available data, a hierarchically structured image data base has been generated that provides image groups of specific visual content as well as semantic connections between those sets. This large-scale visual database, called “ImageNet”, supports approaches in understanding visual content and will be presented in the following section.

2.2.2 ImageNet – A Semantic Hierarchy of Visual Content

Certainly, the pictured *content* plays a major role for the meaning of an image. In general, under the assumption that an image is not too abstract or blurry and the content is clearly visible on the picture, humans can easily identify and understand the displayed scene. However, as denoted in the previous section, recognizing the displayed content is more challenging for machines. Within the last years, computer vision algorithms witnessed great success in exploiting the huge amount of available visual data to grasp and understand the content a picture displays (Section 4.1.3). Thereby, the recognition of people, animals, or popular objects became reliable, and research on naming them and associating properties or actions has evolved. Methods have been developed that, e.g., generate textual descriptions based on the extracted visual content (Section 4.2.1). Overall, the generation of the *ImageNet* has been a major support for the design of such machine-based algorithms as it provides a semantic structure of image sets that visualize the same content as well as meaningful links between those sets.

Generation. The *ImageNet* [3] has been created by Deng et al. [DDS$^+$09]. In order to provide a large structured data set of visual content, they made use of the huge but unstructured amount of online available image data (Section 1.1) and exploited the hierarchical structure of the previously described WordNet (Section 2.1.2). Based on a part of the noun synsets of WordNet, they queried various search engines to collect candidate images for those nouns. Further, to provide a clean image collection, they performed a human-based verification step on the assembled images. Therefore, they made use of Amazon Mechanical Turk (Section 2.3.1) to ask humans if a particular image matches the initial query. This allowed them to provide, on average, between 500 and 1000 verified images per node. In its initial state, ImageNet was built upon 12 subtrees of WordNet and consisted of around 5K synsets and 3.2 million images [DDS$^+$09]. Meanwhile, according to the statistics on the ImageNet website [3] (accessed September 28, 2017), the data set has grown to a considerable size of around 21K synsets with over 14M images.

Visual Synsets. The ImageNet provides image sets that visualize a part of the noun subtrees and synsets of WordNet. Thereby, those visual sets are semantically connected according to the semantic hierarchical structure of WordNet and can provide visual representations of different meanings of a particular noun. This allows the meaningful exploration of a word on a visual level. This means that, similar to searching for the different meanings of a word in WordNet, the ImageNet can also be browsed not only for visual representations of a specific word, but also for visualizations of various meanings of this particular word, i.e., visual representations of the according synsets. For example, as previously shown in Table 2.1, when searching for a word in WordNet, e.g., the word "mouse", different synsets describing the various meanings are returned. Similarly, the ImageNet returns image sets for some of those meanings. Figure 2.4 shows an example of the results returned by the ImageNet online explorer [4] when searching for "mouse". Thereby, various meanings are depicted by the provided image sets, e.g., a mouse as an animal, or a computer mouse. Additional information such as the name of the synset, its popularity (composed of information from Google's text search results and the British National Corpus), or the actual depth in the WordNet hierarchy are also included.

With its high diversity of image content over various categories, the ImageNet supports the identification of the pictured visual content due to the underlying textual structure. Therefore, this data base has been widely exploited to develop, e.g., methods for image classification or object recognition and will be contrasted to other structured image data bases for visual recognition tasks in Section 4.1.3. Further, this large-scale structured and clean visual data supports the successful development of deep learning networks as it allows machines to learn and predict

[3] `http://image-net.org`

[4] `http://image-net.org/explore`

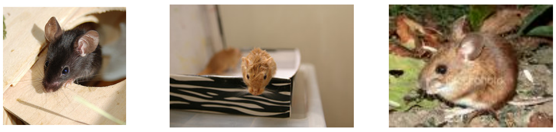
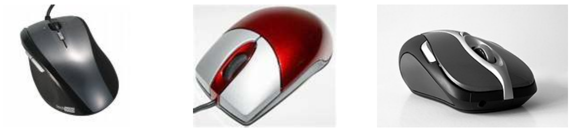
Thumbnails	Synset	Popularity percentile	Depth in WordNet
	mouse	89%	11
	mouse, computer mouse	87%	7

Figure 2.4: *Some image set results with additional information returned by the ImageNet online explorer* [4] *when browsing for "mouse". The different image sets visualize underlying synsets in WordNet and depict various meanings, e.g., an animal or a pc mouse. (accessed October 4, 2017)*

the visual content pictured on images. Relevant work in this area will be presented in Section 4.1.3.

This thesis indirectly makes use of this data set to the extent that CNNs, which have been trained and evaluated on the ImageNet, are exploited to approach visual aesthetics (Section 5.2) and to predict the visual content depicted by images for the task of visual storytelling (Section 6.3). However, although the ImageNet provides a large amount of visual data that certainly depicts specific visual content and is attached with textual information like the name of the synset and a corresponding textual definition, it is not directly suitable for the text illustration task presented within this thesis (Section 6). First of all, the data is restricted to nouns. Although nouns are very important in texts, other word types, e.g., adjectives or adverbs, are also relevant to maintain the meaning when translating a text into a semantically close visual representation (Sections 6.2.3, 6.3.3). Furthermore, the images mostly only depict the specific content denoted by their noun class, e.g., a certain object, rather than a complete scene which is required for the illustration task. Thus, this thesis focuses on images and associated meta data from Flickr to approach the mentioned task of illustrating arbitrary texts with pictures.

2.2.3 Visual Appearance

Although the content plays a major role for the meaning of an image, the *appearance* of the picture is at least of equal relevance for its displayed semantics. A specific appearance like a dominant *color theme* or a particular *visual style* establish a certain atmosphere and transport a specific mood with an image. For example, as previously mentioned, the pictures shown in Figure 1.2 were retrieved for the query "landscape" and display outdoor scenes in different visual appearances which are capable of establishing various atmospheres. Thus, the different appearances largely influence the transported meaning of the single pictures whereas the actual content remains the same.

Meaningful Color Filtering. Color is a powerful visual feature that typically relates to certain moods and largely influences the transported meaning of an image. For example, in Figure 1.2, the dominance of black and white in the second image (top row) might lead to a rather sad feeling. However, selecting a particular color along with a textual search query can even refine the meaning of the queried content. Figure 2.5 shows such refinements when filtering the image search results of the query "landscape" by a specific color. Thereby, the finally displayed type of landscape changes based on the selected color. For example, *green* leads to grass and field scenes, *blue* to sea and a big amount of sky in the image, and filtering by *red* tones mainly results in pictures of sunset landscapes, red rocks or red flowers. Thus, filtering search results by color can present various meanings of the same initial textual query. Obviously, there exists a strong correlation between the dominant color theme and the semantics of an image.

Related work on color manipulation to change the visual appearance of an image and research investigating in the effect of color on the emotions of human beings will be outlined in Section 5.3.2. Further, this thesis explores the connection between the visual appearance of an image and the emotional response of the observer with a strong focus on image color and the aim to actually change the perceived emotion through global image modifications (Section 5.3). Additionally, as color plays a significant role for the visual appearance of pictures, we incorporate a color feature in our text illustration methods (Sections 6.2, 6.3) to support the visual coherence of neighboring images within a stream.

Visual Style. Images can appear in a variety of different *visual styles* which function as semantic representations. Typically, artists utilize particular styles as meaningful devices to strengthen their intentions and trigger certain thoughts or feelings in an observer. Generally, such a style can be defined as:

> *Style:* "A distinctive appearance, typically determined by the principles according to which something is designed." [Stynd]

In other words, a visual style is a particular type of appearance that follows specific rules. Visual styles comprise a huge diversity of different types. For example, Figure 1.2 shows an image taken under "long exposure" which can be seen as a photographic style. But also art genres like "baroque" can be considered as visual styles. As noted by Karayev et al. [KTH$^+$14], the visual style strongly influences how the image is looked at and, in many cases, is an important part of the meaning of the image. Thus, they explored a variety of different art genres, e.g., "impressionism" or "cubism", and visual styles, e.g., "vintage" or "noir", and presented work on

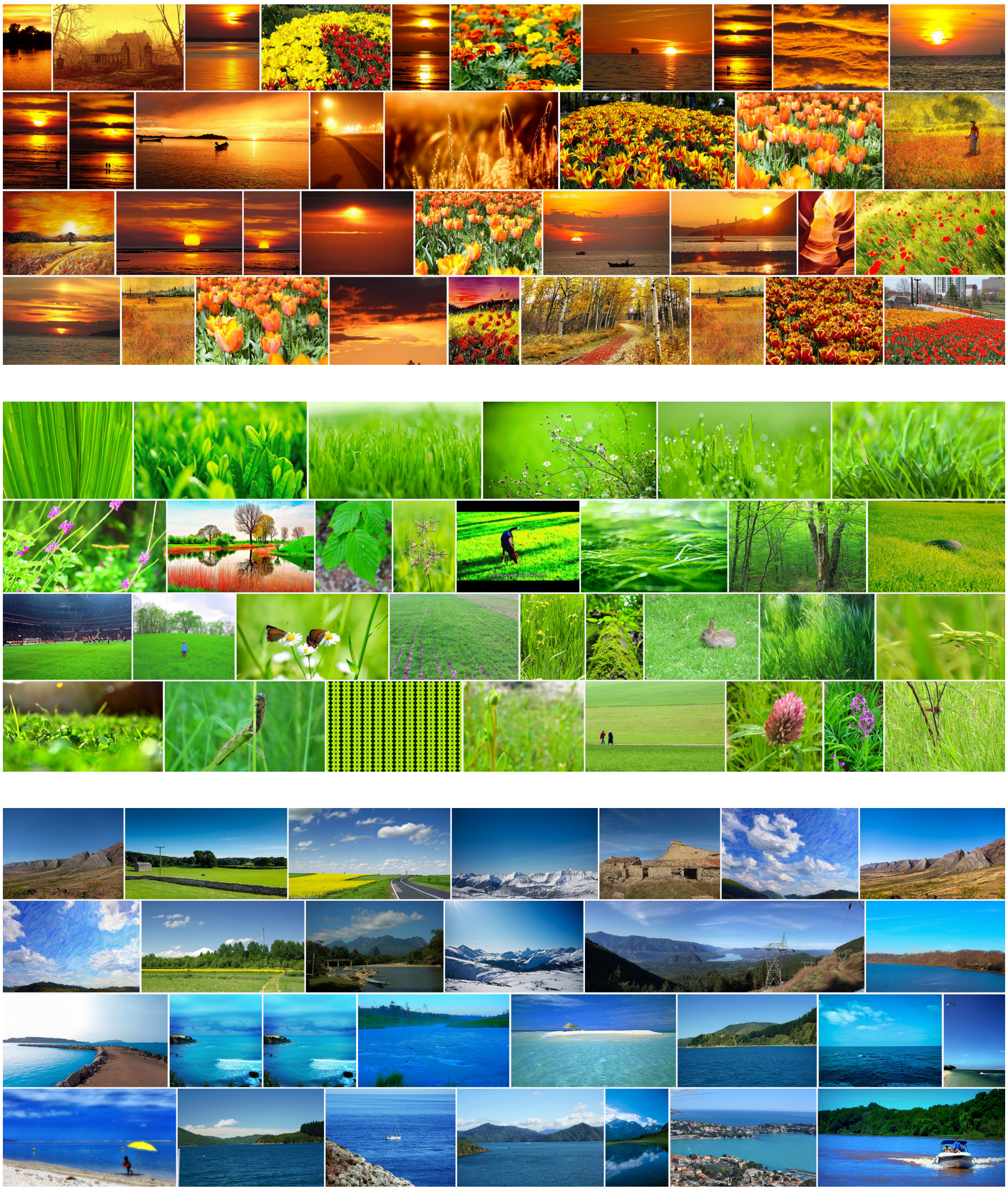


__Figure 2.5:__ Flickr results for query "landscape". Filtering the images by color displays different types of landscapes. Top to bottom: "Red" leads to sunset and flowers, "green" to field and grass, and "blue" to sea or a big amount of sky in the image. Date of retrieval: April 26, 2017. Image results licensed under "No known copyright restrictions".

recognizing such styles in images. An overview of related work on visual style will be given in Section 6.3.2. The present thesis makes use of the power of visual styles to enable meaningful visual storytelling in various styles as well as to strengthen visual coherence along an image stream when generating such picture stories, and explores which styles are preferred by observers for the illustration task (Section 6.3).

2.2.4 Image and Observer

Overall, visual semantics is a highly subjective topic. Person related factors like, e.g., cultural aspects or personal experience largely influence how an *observer* feels when looking at an *image*. Besides, as internal ratings between people differ largely, the feelings between observers typically also vary. Thus, not only the image itself, but also personal aspects have a strong impact on the way a person sees a picture. For example, an appealing look of an image has a large influence on how much the picture is liked by the observer. Further, if such an image is perceived as pleasing, it can still evoke a variety of different feelings. In both cases, the opinions and feelings can differ largely between individuals. Hence the *aesthetic appeal* of an image as well as the *emotion* of the observer are relevant subjective components. In the Oxford Dictionaries[5], those terms are defined as:

> *Aesthetics:* "A set of principles concerned with the nature and appreciation of beauty." [Aesnd]
>
> *Emotion:* "A strong feeling deriving from one's circumstances, mood, or relationships with others." [Emond]

Thus, aesthetics is generally related to beauty. In the particular case of images, a picture is typically considered as aesthetically pleasing, if it displays some beautiful content, presents a beautiful appearance, or also, both together can be meant. On the side of the observer, an emotion typically arises from the inside of a person but, at the same time, also depends on the persons' environment. Thereby, also pictures can have a strong impact on the emotional perception of an observer.

An overview of the computational aspects of aesthetics and emotion in images along with related aspects in, e.g., photography, paintings, or visual arts, has been given by Joshi et al. [JDF+11]. Their survey highlights the challenging semantic gap between low-level image features and high-level person related semantics as well as the relevance of human subjectivity when dealing with the perception of aesthetics and emotion. Previous work on visual aesthetics in images as well as its formulation as a deep learning problem will be discussed in Section 5.2.2. Further, research that has investigated in emotion categorization will be outlined in Section 5.3.2.

As both, the aesthetic appeal of an image as well as the evoked affect of an observer are of high relevance to approach visual semantics, both components are explored

[5] https://en.oxforddictionaries.com/

in this thesis. In order to tackle the subjective nature of aesthetics, we will analyze the aesthetic appeal of images. In particular, we will exploit a huge diversity of user ratings from social online behavior to learn rankings and predict aesthetic appeal on arbitrary input images (Section 5.2). Further, this thesis investigates into the direct connection between the appearance of images and the emotional response of the observer and explores to which extent global image modifications are actually capable of changing the emotional perception (Section 5.3).

Overall, the high subjectivity of semantics requires the incorporation of human beings when developing methods that deal with the meaning of images. Thus, this thesis exploits the recent interest in social online behavior and makes use of the huge amount of images with associated textual information provided by a broad diversity of people through Flickr. Additionally, in order to evaluate the developed techniques on a highly diverse level, we employ crowdsourcing which will be described in the following section.

2.3 Crowdsourcing Visual Semantics

With the growth of the internet and the huge interest in online social behavior, *crowdsourcing* became a popular method for large-scale data collection. The Oxford Dictionary [5] provides the following definition:

> *Crowdsourcing:* "The practice of obtaining information or input into a task or project by enlisting the services of a large number of people, either paid or unpaid, typically via the Internet." [Crond]

Thus, as crowdsourcing tasks are typically spread over the web and, thus, the particular tasks are carried out world wide, it offers the opportunity to address a huge and highly diverse crowd.

As previously mentioned, the highly subjective nature of semantics implies a variety of interpretation possibilities that can differ largely between individuals. Especially, the semantics of images involves a wide range of meaningful interpretations as content as well as appearance can be perceived differently and trigger various complex feelings. Already two persons might have a different opinion about whether a particular picture has, e.g., an appealing look. For example, all the images in Figure 2.3 show a sunset scene with different content elements like a boat, a car, or glass. Although all pictures establish a similarly appealing warmish atmosphere, people might favor one image over the other depending on their personally preferred content.

Therefore, when approaching this highly subjective task of visual semantics from a computational perspective, humans need to be integrated into the process. More specifically, to evaluate machine-based methods developed with the objective to approach visual semantics, a large enough and highly diverse crowd is needed to overcome isolated opinions as well as cultural aspects. Crowdsourcing can be employed to evaluate tasks that involve visual semantics. Thereby, Amazon Mechanical Turk became a popular web-based platform to perform crowdsourcing.

2.3.1 Web-based Experiments with AMT

Crowdsourcing platforms like *Amazon Mechanical Turk (AMT)* [6] allow to perform web-based experiments and collect data online. Thereby, AMT became a popular marketplace and is widely used for research. For example, Buhrmester et al. [BKG11] explored the capability of AMT to contribute to social sciences like psychology. In AMT, the web workers are called "Turkers" or "MTurkers" and can earn money by performing tasks online. Those tasks are referred to as "HITs" (Human Intelligence Tasks) which are typically rather small tasks at very low costs. AMT has been found to be a promising tool for conducting research yielding reliable data on a demographically diverse level [BKG11]. This diversity on Internet samples for data collection has also been discovered by Gosling et al. [GVSJ04]. They have explored several preconceptions about web-based studies in general and have conducted a large scale comparison to traditional samples. Overall, they have found that Internet samples provide diversity in several terms like gender, geographic region, or age while yielding similar findings as traditional samples. In addition, as described by Buhrmester et al. [BKG11], they have demonstrated that web-based studies are able to reduce biases inherent in traditional studies. Overall, conducting experiments on the Internet and, in particular, using AMT has become of interest in various research fields.

2.3.2 AMT for Visual Experiments

With the aim to validate if crowdsourcing is a suitable method to perform graphical perception experiments, Heer et al. [HB10] contrasted web-based experiments with previously performed laboratory ones. Their experiments comprise the exploration of variables for data visualizations, in particular spatial encodings like area or shape, and luminance contrast. In detail, they carried out several experiments on the Amazon Mechanical Turk (AMT) platform, and identified several benefits over laboratory experiments:

- "up to an order of magnitude cost reduction",
- "a faster time to completion",
- and "access to wider populations" [HB10].

However, they mention that additional tools could further aid to obtain better control and, e.g., avoid overlap of test persons participating across experiments.

[6] `www.mturk.com/mturk/`

Anyway, compared to laboratory experiments, they found that crowdsourcing is suitable to evaluate visualization design, especially, as the scalability of web-based experiments allows for much higher participation at the same cost.

Overall, crowdsourcing with AMT has been successfully exploited to conduct research in fields like social sciences but also for graphical perception experiments. In summary, the main advantages are the *high demographic diversity*, the *large participation at low cost*, and the typically *fast completion time*. Therefrom, especially the large participation of a broad and diversified crowd makes AMT well suited to evaluate the highly subjective task of visual semantics and validate trends. Thus, several parts of this thesis utilize crowdsourcing to support the development of techniques that approach visual semantics and to evaluate the corresponding results in appropriate user studies. In detail, we employ AMT to evaluate our derived model to rank aesthetic appeal (Section 5.2), to measure the emotional response on visual stimuli (Section 5.3), and to rate the quality of the semantic translation from textual descriptions into meaningful visual representations as well as the coherence of the visual appearance along a set of images (Section 6.3).

2.4 Conclusion

In summary, semantics can relate to the meaning of any source. Semantics has been widely explored in linguistics and increasing interest has been shown in exploring semantics within visual tasks. However, especially for machines it is a huge challenge to grasp the evolving high-level meaning or, even, to bridge the semantic gap between different sources. The high subjectivity of semantics claims for humans to be integrated into the process of developing and evaluating techniques that approach semantics. It is essential to access a huge diversity of people to overcome, e.g., cultural aspects. Therefore, crowdsourcing and, in particular, AMT has successfully been exploited to support research in different fields.

The present thesis focuses on the semantics of images and the semantic relatedness between visual and textual sources. When considering image semantics, i.e., the meaning of an image, both, the displayed content as well as the appearance of an image are of great importance. Further, already the meaning of a single image can differ largely between observers. To evaluate the highly subjective task of visual semantics as well as the meaningful connection between text and images, ratings from a wide variety of people on a high demographically diverse level are necessary. Therefore, additionally to employing manual annotations and ratings from a huge amount of people exploiting their social online behavior in Flickr, we also perform web-based experiments. In particular, we employ AMT in several projects (Sections 5.2, 5.3, 6.3).

Further, as the present thesis aims to automatically illustrate text with pictures (Chapter 6), the semantics of the text needs to be identified to realize a semantically close translation from natural language to a visual representation. Already single words can transport a diversity of meanings and are often only understood within their context. For machines, it is even more challenging to identify the particular meaning. However, especially the WordNet database enables the design of algorithms that allow machines to identify the semantics of natural language, e.g., a text passage, a sentence, or the single words. As processing the textual description and extracting relevant information forms a fundamental part of this thesis, basics on natural language processing will be given in the following chapter.

3 Basics on NLP

Throughout this thesis, natural language plays an important role. Several parts build on a textual basis either to submit queries to online photo collections for retrieving images, translate short texts into visual representations, or even illustrate complete stories. In order to enable the usage of given natural language in a machine based environment, the text needs to be analyzed and prepared for further processing. Therefore, we employ techniques from the field of *Natural Language Processing (NLP)* to analyze written text and extract relevant information to our needs. This chapter outlines the fundamental concepts of NLP which are relevant for this work.

After clarification of basic terminology (Section 3.1), relevant techniques of NLP, in particular, fundamental concepts to process written language are introduced (Section 3.2). Then, several relevant models to represent text are described (Section 3.3). Finally, different concepts on handling information like extracting information from text, or retrieving information which often employs NLP as preprocessing step and relies on mentioned representation models are outlined (Section 3.4).

3.1 Terminology

Whereas diverse forms of language exist, e.g., birds communicate via sounds, or humans use speech to talk directly to each other, we will focus on written text as basic type of language. In the following, important elements related to language and text processing are itemized to clarify basic terminology that is commonly used in language processing and relevant for this thesis.

3.1.1 Language

In order to successfully process and understand *natural language*, text processing systems require a certain amount of *knowledge* about language.

Natural Language. The field of Natural Language Processing (NLP) follows the idea of allowing computers to process and understand unrestricted human language with the aim of enabling computers to perform useful tasks involving human language, e.g., communication between humans and machines [JM08].

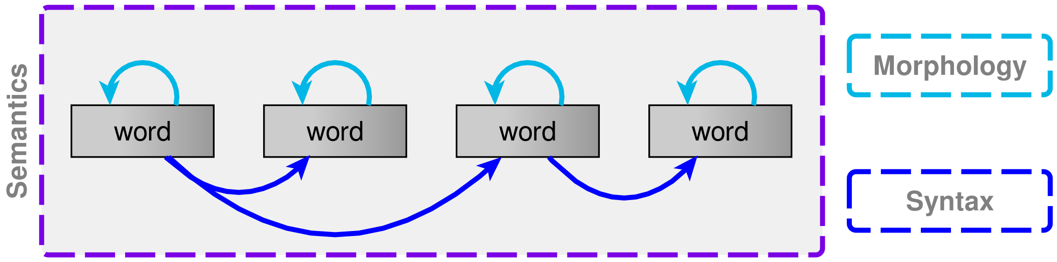


Figure 3.1: *Various kinds of knowledge are required to tackle language behavior.*

Thereby, the human *natural language* is contrasted to the formal *computer language* to distinguish between these sources. Several parts of this thesis allow for such natural language written by humans as input to translate it into a corresponding visual representation.

Knowledge. Generally, a lot of *knowledge* is required to tackle complex language behavior. Jurafsky and Martin [JM08] have listed several types between which they differ, whereof the following ones are relevant to this thesis:

- *Morphology*: “knowledge of the meaningful components of words”,
- *Syntax*: “knowledge of the structural relationships between words”,
- *Semantics*: “knowledge of meaning” [JM08].

An overview of those knowledge components is given in Figure 3.1. They form important parts in the field of computational linguistics. In detail, *morphology* is concerned with the inner structure of single words and their creation from pieces (morphs), *syntax* deals with the way single words are connected within sentences, and *semantics* comprises the meaning of single words as well as entire texts [BG04].

In general, language processing systems differ from data processing tools in such that they have a certain *knowledge* about language or, in other words, “knowledge about what it means to be a word” [JM08]. As already indicated in Section 2.1, understanding the meaning of words is strongly correlated with the task of resolving their ambiguities as a single word can have multiple meanings. This thesis explores the meanings of words to derive corresponding visual representations and, thereby, relies on the mentioned types of knowledge.

3.1.2 Text

As work in the field of NLP cannot be done through large-scale observation of language usage in a real world scenario, texts are used instead and serve as a substitute [MS01]. A text typically consists of *words* as the most basic tokens that are combined into *sentences* by a suitable structure. Further, *documents* comprise multiple sentences and a collection of texts is referred to as a *corpus* in NLP. A *treebank* is a large corpus in which the sentences are syntactically annotated.

Words. A *word* is a sequence of one or more letters, provides at least one *meaning* or *sense*, and forms the basic element or *token* within natural language (Section 3.2.1). Typically, multiple words are separated by spaces or punctuation marks. As already mentioned, words can have different meanings whereby the particular meaning often needs to be grasped from the context. Further, words can be grouped into different categories or *types*, e.g., nouns or verbs. At the same time, a single word can have various types, e.g., the word "run" can be a noun or a verb depending on its position within a sentence. The process of *part-of-speech (POS) tagging* (Section 3.2.2) defines those types.

Dictionaries. The meanings of words can further be looked up in *dictionaries*, i.e., lexical resources that also provide various information, e.g., about the forms and types of individual words. The electronic WordNet dictionary introduced in Section 2.1.2 even provides links between synsets that are semantically or lexically related. An example is shown in Table 2.1.

Sentences. *Sentences* connect multiple words with a suitable grammatical structure, i.e., the syntax. A sentence provides context around single words and can assign a specific meaning to ambiguous ones. In NLP, sentences are *parsed* into their syntactic structure (Section 3.2.2) and represented as *parse trees* (Figure 3.5) that display the grammatical structure and indicate the role of words within a sentence.

Documents. A *document* is a unit of written text and comprises multiple words which are typically structured into sentences. The notation of a piece of text as a document is commonly used in information retrieval (Section 3.4.2). Thereby, frequency counts, i.e., counting the words or the types of different words are employed to classify documents [MS01].

Corpora. A collection of texts or documents is defined as a body of texts and termed *corpus* (Latin for "body") in NLP whereby several such collections are called *corpora*. Corpora are relevant to many statistical NLP techniques as they support such algorithms to learn both lexical and structural preferences from large data [MS01]. A popular and pioneering example is the *Brown corpus* (Section 3.2.2).

Treebanks. A *treebank* is a large parsed corpus which provides syntactic annotations, commonly in the form of parse trees that are generated for every sentence [JM08]. Such treebanks have become popular in linguistic research as they allow to train stochastic parsers. A large-scale and well-known treebank is the *Penn Treebank* which will be presented in Section 3.2.2.

3.2 NLP Techniques

In order to identify basic linguistic elements like words within a text, their meaning, or the structure between them, several concepts have been developed. This sections outlines fundamental techniques of NLP to process written natural language.

3.2.1 Words

As mentioned in Section 3.1.2, words are the most important entities in natural language. Thereby, *morphology* deals with the meaningful components within a word. Breaking down a word into its *morphemes*, e.g., the stem and the ending, enables to form new words like a plural from a singular form.

NLP tasks presented in this section comprise detecting the inner structure of words to create new ones or recognizing the basic stem to identify derived ones, but also collecting words that are less meaningful than others for specific tasks. However, before words or their inner structure can be analyzed, the words, more precisely, the tokens need to be detected in the text.

Tokenization

When processing natural language, one of the first steps is to divide the input text into its individual units called *tokens*. A token can be a word but can also be, e.g., a number or a punctuation mark. The process of identifying those tokens, more precisely, their ranges within a text is called *tokenization*. The key task that needs to be handled can be formulated as "what counts as a word" [MS01]. Straightforward, tokenization can be thought of as simply separating words by their surrounding white space. However, there are several issues with this procedure. [JM08]

A common problem is the treatment of punctuation marks and whether they belong to a word or not. For example, if tokens are solely divided by whitespace, punctuation marks are not split from a word if they directly follow one another even if they do not belong together as, e.g., the punctuation indicates the end of a sentence. On the other hand, punctuation marks sometimes belong to a word and should remain as a part of it. For example, to distinguish between the abbreviation for the state of Washington "Wash." and the capitalized form of the verb "wash", the period associated with the abbreviation should stay there. [MS01] This problem of identifying punctuation marks that indicate the end of a sentence will be outlined in Section 3.2.2.

Another issue is dealing with contractions like "I'll" or "isn't". Overall, this problem is solved differently. Some methods treat such contractions as two words and, e.g., resolve the mentioned examples as "I will" and "is not". Other approaches seek to maintain the initially mentioned form and consider them as so-called *graphic words*. This notion introduced by Kučera and Francis [KF67] is defined as

> "a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks". [KF67]

In this context, numbers and monetary amounts like "€13,05" are also viewed as words. Following this definition, tokenization often relies on the occurrence of whitespace around words even if it is not always well-defined. For example, words like "data base" or city names like "New York" contain whitespace inbetween but should probably be considered as a single word. [MS01]

Stemming and Lemmatization

As already indicated, obtaining morphological knowledge about the inner structure of a word is relevant for various approaches. A common task is to identify the *stem* of a word to find out which words are only derivations and actually come from the same. The process of simply stripping off the endings of words is known as *stemming*. It is, for example, applied in retrieval systems (Section 3.4.2) to find matches between similar words that occur in different forms, e.g., singular and plural, whereby the particular form is irrelevant for the retrieval task. Other linguistic tasks need to know whether several words have the same root and, thus, are different forms of the same word. For example, "sang", "sung", and "sings" all come from the same word "sing". This base form is called a "lemma" and the process of mapping the other forms to this central one is known as "lemmatization". [JM08]

Overall, it can be distinguished between two types:

- *Inflectional forms* of a word are different forms such as "organize", "organizes", and "organizing".
- *Derivationally related words* are families of words with similar meaning such as "democracy", "democratic", and "democratization".

Then both, stemming and lemmatization, aim at finding the basic form. However, the two methods differ in such that

- *stemming* chops off suffixes to identify the stem in a rather crude heuristic process, whereas
- *lemmatization* removes inflectional endings and returns the base form, the lemma, making use of a vocabulary as well as proper morphological analysis.

For example considering the word "saw", stemming might simply output "s", whereas lemmatization would probably return "see" or "saw" depending on the role in which the word was used, i.e., as a verb or a noun. [MRS08] Overall, depending on the task, one or the other may be more helpful. The probably most common algorithm for stemming is briefly introduced in the following.

Porter Stemmer. A popular and empirically effective approach for stemming English language is *Porter's stemming algorithm* [Por80]. The algorithm consists of several phases, in particular, of 5 steps which are applied sequentially to enable the removal of longer and more complex suffixes. The rules to remove a suffix are provided in the following form:

$$\text{(condition) S1} \rightarrow \text{S2}$$

In particular, in the case that a word ends with S1 and the stem preceding S1 fulfills the given condition, S1 is replaced by S2. The purpose of this condition is to verify if a word is long enough that the matching part S1 can be considered a suffix and not a part of the stem. The following rule is an example from step 4 of the rule set to chop off "ement" from the end of a word by replacing it with a space:

$$(m > 1) \text{ EMENT} \rightarrow$$

Thereby, the measure m checks the number of syllables in the word to verify its length. For example, the word "replacement" would be mapped to "replac" which consists of two syllables ($m = 2$) and, thus, fulfills the condition $m > 1$. Contrarily, for the word "cement", no mapping would be performed as cutting off "ement" would result in the token "c" containing no syllables ($m = 0$) and, thus, not fulfilling the condition. Such words with zero syllables are also called *null words*. [Por80]

Stop Word Removal

Some tasks require the removal of complete words from a text or a set prior to further processing. Individual words that occur extremely often but are typically of little value for a particular task are called *stop words*. Such words like "the", "a", or "to" usually do not convey the meaning of a text and, therefore, can be filtered out. This process is know as *stop word removal*. A common approach to identify stop words in a text is to use a pre-compiled stop word list which can be found in

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your',
'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her',
'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs',
'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',
'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how',
'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can',
'will', 'just', 'don', 'should', 'now']
```

Figure 3.2: *Stop word list from the python Natural Language Toolkit (NLTK)[1] [BKL09].*

different lengths. Figure 3.2 shows an example of such a list of stop words from the python Natural Language Toolkit (NLTK)[1] [BKL09].

Stop word removal excludes such words from a set or a text and is often applied for reasons of efficiency or effectiveness, depending on the task. In general, the goal of *text mining* is to mine useful information from a textual document. As stop words like articles, prepositions, etc. do usually not represent meaningful information in a text, stop word removal is often used as a pre-processing step for various approaches in the field of text mining [VIN15]. In information retrieval (Section 3.4.2), stop words are often removed from the vocabulary to increase efficiency of the index creation as such words are typically not useful when matching a query and a document [MS01]. Often, they do not help to distinguish one document from another. On the other hand, some well known sentences, e.g., the famous phrase

"To be, or not to be" in W. Shakespeare's play "Hamlet",

only consist of common stop words and might not be retrieved [MRS08]. As IR systems need to deal with such phrases, the trend goes towards smaller or even no stop word lists or focus more precisely on the statistics of such common words. More details are given by Manning et al. [MRS08] Further, Saif et al. [SFHA14] have explored if stop word removal supports the effectiveness of sentiment classification over Twitter and have compared different techniques for stop word identification. For that particular task, they found that removing singleton words results in a good trade-off between performance and processing time.

Overall, this thesis makes use of the presented techniques, e.g., to tokenize the input texts or stories (Chapter 6), which is a fundamental step to enable any further textual processing. Further, approaches of stemming and lemmatization are employed to

[1] `http://www.nltk.org/`

identify singular forms of nouns enabling to connect named objects with 3D models (Section 6.1) or to facilitate tag-based image search (Sections 6.2, 6.3). Finally, stop word removal is applied to support relevant matching between text lines and image tags from a large corpus (Section 6.3). In particular, we filter such stop words from our creative text corpus (Section 6.3.4) to create a relevant word list (Appendix A)

3.2.2 Syntax

The last section has focused on individual words and their inner structure. However, as indicated in Section 3.1.2, single word tokens are typically combined into sentences with a suitable grammatical structure, also called *syntax*. Techniques comprise detection of sentences or identification of the role of words within sentences and are developed using large corpora.

Segmentation

In addition to identifying words in language, segmenting a text into its sentences is another essential first step in NLP and is referred to as *sentence segmentation*. Generally, the recognition of sentences within a text is based on punctuation marks like ".", "?", or "!", as they typically mark the boundaries of a sentence. However, whereas question marks or exclamation points quite clearly determine sentence boundaries, periods can also mark abbreviations, e.g., "Mr.". [JM08] In other words, as mentioned in Section 3.2.1, punctuation marks can also belong to a word token, wherefore it is important to find out which punctuation marks indicate the end of a sentence and which indicate abbreviations.

Various approaches for sentence segmentation exist. A key task thereby is period disambiguation. Typically, a binary classifier is built following a sequence of rules like regular expressions, or employing machine learning. For the classifier to decide whether a period belongs to an abbreviation or not, abbreviation dictionaries can be utilized. [JM08]

An example of a typical basic heuristic algorithm for recognizing sentence boundaries has been outlined by Manning and Schütze [MS01]. The main steps are summarized in Figure 3.3. Assume a boundary-marker is a putative sentence boundary. Then, boundary-markers are placed after all punctuation marks and are iteratively discarded based on several cases. The remaining ones are considered as sentence boundaries. Overall, the main steps are listed in Figure 3.3. Variations of this algorithm have been successfully used in many systems. However, such solutions tend to be rather domain-specific and require a certain amount of hand-coding. [MS01] Thus, more general solutions have been developed, for example, by Palmer and Hearst [PH94, PH97] which present an efficient machine learning approach and

- Place boundary-markers after punctuation marks (". ? ! ; : -")
- If quotation marks follow boundary-marker, move marker behind them
- Remove period boundaries if preceded by known abbreviation
- Remove "?" and "!" boundary if followed by lowercase (or known name)
- Consider remaining boundary-markers as sentence boundaries

Figure 3.3: *Heuristic algorithm for sentence boundary detection by Manning and Schütze. A boundary-marker is a putative sentence boundary. (Summarized from [MS01], Fig. 4.1).*

avoid labeling of large training data sets by utilizing the part-of-speech distribution of words surrounding the ambiguous punctuation mark.

Part-of-Speech (POS) Tagging

The words of a language can be classified according to their syntactic behavior. These word classes are then referred to as *part-of-speech*, or POS, and provide a great deal of information about the word itself as well as its neighbors. Parts-of-speech can be split into two classes, namely *closed class* and *open class* types. An example for a closed class is the set of prepositions as there is a fixed amount in English to which new ones are rarely added. In contrast, open classes like nouns or verbs still grow as they, e.g., borrow words from other languages. The most important open classes are

- *nouns*, e.g., referring to people, animals, things,
- *verbs*, expressing the action in a sentence,
- *adjectives*, describing properties of nouns, and
- *adverbs*, referring to any other part of language than a noun.

However, more fine-grained classifications of word classes are used to distinguish, e.g., between a singular or a plural form. [JM08] In general, words and their part-of-speech are listed in a *dictionary* or *lexicon*. Well-established sets of abbreviations, called *POS tags*, exist to name those word classes. For example, a noun is normally associated with the tag NN or its plural form NNS. [MS01] The NLP process of labeling a word in a text (or corpus) with its corresponding POS tag is then referred to as *part-of-speech tagging*. An example of a tagged sentence is given in Figure 3.4.

```
"The   children   ate   sweet   candy"
 DT      NNS      VBD    JJ       NN
```

Figure 3.4: *Short example sentence associated with part-of-speech tags from the popular Penn Treebank POS set.*

Table 3.1: *Part of the Penn Treebank POS tagset relevant for this thesis. In total, the Penn Treebank tagset consists of 48 tags whereby 36 are POS tags and 12 other tags relate to punctuation and currency symbols [MMS93].*

Tag	Part-of-Speech
CC	Coordinating conjunction
CD	Cardinal number
IN	Preposition / subordinating conjunction
JJ	Adjective
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
RB	Adverb
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund/ present participle
VBN	Verb, past participle

In this example sentence, part-of-speech tagging has been performed using abbreviations from the popular Penn Treebank POS tagset (Table 3.1). Thereby, "the" is tagged with DT, "children" with NNS, the plural form of a noun, "candy" with NN, "ate" with VBD, for the past tense verb, and "sweet" with JJ.

Overall, several tagsets have been developed. However, the probably most popular ones for English are the 87-tags set employed for the Brown corpus from which many other tagsets have evolved and the smaller but widely used Penn Treebank tagset. [JM08] The Penn Treebank comprises 48 tags [MMS93], whereof a part that is relevant for this thesis is listed in Table 3.1. The Brown corpus and the Penn Treebank are introduced in the following.

Brown corpus. The first major corpus of English is the *Brown corpus* [FK64] which has been assembled at the Brown University and published by Kučera and Francis [KF67, FK82]. Being considered a representative sample of American English, it serves as a balanced collection of 500 written text samples representing various genres and comprises around 1 million tagged words [MS01].

Penn Treebank. The *Penn Treebank* is a large-scale and well-known treebank which has been constructed during multiple phases [MMS93, MKM^{+}94] The first release already provided over 4.5M English words annotated with their POS tags whereby the set of POS tags is a reduced version of the Brown Corpus tagset to eliminate redundancy. A part of the Penn Treebank POS tagset is listed in Table 3.1. Additionally, more than half of this text is associated with annotations in the skeletal bracket form depicting the syntactic structure [MMS93].

Taggers. In order to perform POS tagging, a variety of taggers have been developed. Overall, tagging approaches can be split into *rule-based taggers* and *stochastic taggers*. Rule-based taggers rely on a large set of hand-written rules. Those disambiguation rules determine, for example, that if an ambiguous word follows a determiner, it is rather a noun than a verb. Stochastic taggers resolve such ambiguities based on a training corpus and compute the probability of a particular word having a given tag in a certain context. Further, a tagger that shares properties from both tagging approaches is referred to as *transformation-based tagger*, or the *Brill tagger* which is named after Brill [Bri95]. It is based on rules that specify in which cases an ambiguous word should be associated with which tag whereby the rules are learned from a tagged training corpus. [JM08]

Parsing

In general, *parsing* in NLP means that given some input, a linguistic structure is produced as output. Analyzing the syntax of a sentence provides information on how the meaning of the sentence can be identified based on the meaning of the words, e.g., who performs an action to whom in a described situation like in the following example:

"The girl gave a book to the boy." – "The boy gave a book to the girl."

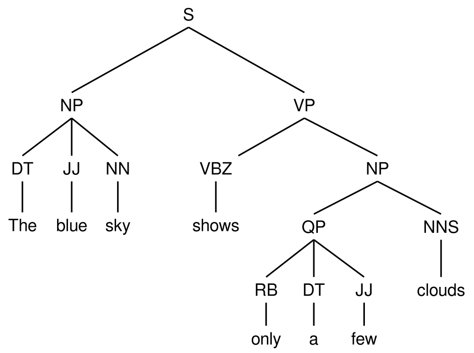
Although these sentences consist of the same words, they carry a different meaning resulting from the ordering of the words. [MS01] In order to identify the structure of a sentence, it is parsed into its so-called *constituents*. A constituent consists of group of words that are assumed to behave as a single unit or phrase. [JM08] The major constituents or phrase types are:

- *Noun phrase (NP)*: A noun phrase usually embeds a noun and gathers information about the noun which is the head of the phrase. It is typically composed of a potential determiner, optional adjective phrases, the noun head, and possibly post-modifiers like, e.g., prepositional phrases.
- *Verb phrase (VP)*: A verb phrase is the part of a sentence that contains the verb as its head and comprises all elements of the sentence that are syntactically dependent on the verb.
- *Prepositional phrase (PP)*: Starting with a preposition and containing a noun phrase complement, prepositional phrases often appear within the other mentioned types of phrases expressing attributes like spatial or temporal locations. [MS01]

Typically, the identified phrases are arranged in hierarchical order called a *parse tree*. In other words, a *parser* uses the syntax or, more precisely, the structure of sentences and the way individual words are connected, and generates a parse tree from the sentence [BG04]. An example for a parse tree of the sentence "The blue sky shows

```
(ROOT
(S
(NP (DT The) (JJ blue) (NN sky))
(VP (VBZ shows)
(NP
(QP (RB only) (DT a) (JJ few))
(NNS clouds)))))
```

(a) Bracket structure



(b) Parse tree

Figure 3.5: Parse tree for sentence "The blue sky shows only a few clouds" with POS tags from Penn Treebank tagset. The parse tree (b) illustrates the parse of the Stanford Parser [KM03] returned as bracket structure by the online version[2] (a).

only a few clouds" is given in Figure 3.5. The bracket structure (Fig. 3.5(a)) is the output of the online version of the Stanford Parser [2] and is visualized by the parse tree representation (Fig. 3.5(b)). The sentence S is split into its noun phrase (NP) and its verb phrase (VP) as the next node hierarchy and so on. The leaf nodes contain the single tokens as well as their POS tags from the Penn Treebank tagset.

Context-free Grammars (CGF). A basic approach to parse sentences into a syntactic representation relies on modeling *context-free grammars (CGF)* to express relationships between the words in a sentence. A CGF consists of a set of rules that defines how symbols can be grouped and ordered, and a lexicon that connects words and symbols. An example for a rule set defining, e.g., possible compositions for a *noun phrase (NP)* is given in the following:

$$NP \rightarrow Det\ Nominal$$
$$NP \rightarrow ProperNoun$$
$$Nominal \rightarrow Noun \mid NominalNoun$$

According to these rules, a noun phrase may consist of a *ProperNoun* or a *determiner (DET)* followed by a *Nominal*, and a *Nominal* may consist of one or more *Nouns*. An example of a lexicon entry linking the *Noun* symbol with words would be:

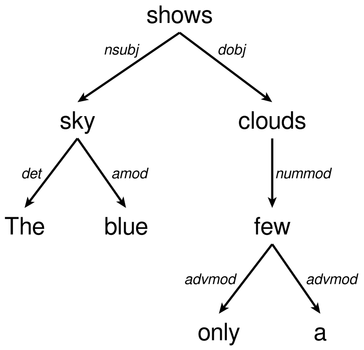
$$Noun \rightarrow sky \mid sun \mid candy \mid salad$$

Overall, context-free grammars serve as the basis for other models. For example, a *statistical parsing* approach is to add probabilities to such rules supporting

[2] http://nlp.stanford.edu:8080/parser

```
det(sky-3, The-1)
amod(sky-3, blue-2)
nsubj(shows-4, sky-3)
root(ROOT-0, has-4)
advmod(few-7, only-5)
advmod(few-7, a-6)
nummod(clouds-8, few-7)
dobj(shows-4, clouds-8)
```

(a) Bracket structure



(b) Dependency tree

Figure 3.6: *Typed dependency parse of same sentence as in Fig. 3.5. The Stanford Dependencies are returned in bracket structure (a) by the mentioned online version of the Stanford Parser. The dependency tree (b) visualizes the output.*

disambiguation and being closer to human parsing. Such a simple augmentation is known as *probabilistic context-free grammar (PCFG)*. [JM08]

Partial Parsing. Instead of building complete and often complex parse trees, a *partial parse* aims at rather identifying the text segments with valuable and content-bearing information which typically leads to trees with a rather flat structure gathering only the major constituents of sentences, e.g., noun phrases [JM08]. This kind of partial parsing is also known as *chunking* and typically used in IE systems (Section 3.4.1) as they often do not require all possibly available information but the valuable one [JM08]. Due to the missing hierarchical structure, the following representation is typical:

[$_{NP}$ `The man`] [$_{VP}$ `ate the salad.`]

This simple bracketing notation denotes the basic flat structure of non-overlapping chunks and is enough to locate the text segments and their types, e.g., the noun phrase (NP) and the verb phrase (VP), as identified by the chunking task [JM08].

Dependency Parsing. The syntax trees mentioned above are common representation forms depicting the syntactic structure of sentences by their constituents. Contrarily, *dependency grammars* describe the sentence structure solely through binary semantic or syntactic relationships between the given words. [JM08]

A popular model are the *Stanford Dependencies* presented by de Marneffe et al. [dMMM06]. They extract dependencies between words in a sentence and, additionally, associate labels with those grammatical relations. Such labeled dependencies are called *typed dependency parses*. Labels used by the Stanford Dependencies to identify grammatical relations are, e.g., *nsubj* (nominal subject), *dobj* (direct object), or *det* (determiner). An example of a typed dependency parse of the same sentence as in Figure 3.5 is shown in Figure 3.6. The Stanford Parser returns the Stanford Dependencies in bracket structure (Fig. 3.6(a)). The dependency tree (Fig. 3.6(b)) provides a visualization of the identified Stanford Dependencies.

Overall, the techniques presented in this section are relevant to several parts of this thesis, e.g., to label words with their according part-of-speech tags (Chapter 6), to identify sentences within a text (Sections 6.1, 6.2), or to extract dependencies between objects as a basis to resolve spatial relations in a virtual environment (Section 6.1).

3.3 Representation Models

The way text is represented has a large influence on the performance of applications. For example, systems for information retrieval (Section 3.4.2) typically consider text as bag-of-words ignoring syntactic ordering and widely make use of the vector space model to efficiently score similarity between documents in the ranking approach.

3.3.1 N-Grams

One of the most important models in language processing is the *N-gram model*. An *N-gram* is a sequence of N words or tokens. For example, a 2-gram, or bigram, consists of 2 words. The idea of the N-gram model is then to predict the last word of a sequence of N words from the previous $N-1$ terms. Overall, N-gram models are essential, for example, to identify words in handwriting recognition, or to support spelling correction while predicting potentially correct options. [JM08]

In detail, assuming an N-gram consists of N words w, then the N-gram can also be written as a sequence

$$s = (w_1, w_2, ..., w_n) = (h, w_n),$$

with $h = (w_1, w_2, ..., w_{n-1})$ consisting of a sequence of $N-1$ words and being the preceding history of the N^{th} word w_n. The goal of an N-gram model is to calculate the probability $P(w_n|h)$, i.e., the probability of the word w_n given its preceding history h.

In order to predict a particular word in a sequence or, more precisely, to calculate the probability for a word to appear at the end of a sequence, a statistical N-gram model needs to compute the probability of sequences. Thus, large corpora as well as knowledge about the tokens within the corresponding corpus are required for being able to count words and compute probabilities in NLP. The knowledge about existing tokens is gained from tokenization which has been described in Section 3.2.1. Then, to calculate the mentioned probability $P(w_n|h)$ that a word follows a sequence h of $N-1$ words, given a large tokenized corpus, a straightforward approach would be to count the occurrences $C(s)$ and $C(h)$ of the sequences s and h respectively and calculate the probability as

$$P(w_n|h) = \frac{C(s)}{C(h)} = \frac{C(w_1, w_2, ..., w_n)}{C(w_1, w_2, ..., w_{n-1})}. \tag{3.1}$$

In probability theory, computing the probability as indicated in Equation 3.1 would provide an answer to the question "Out of the times we saw the history h, how many times was it followed by the word w"[JM08]. Overall, for longer sequences, this is not a suitable estimate. One reason is, that sequences with slightly different orderings of the words will fall out of the counting, although they would provide an adequate contribution to the prediction, e.g., if the ordering only differs in the beginning of the sentence. Thus, bigram models are typically used to approximate the history following the Markov assumption that the probability of a particular word is only depending on its preceding word. [JM08]

3.3.2 Bag-of-words

Approaches that simply consider text as a set of words while ignoring syntactic information or linear ordering are typically referred to as *bag-of-words*. Thereby, it is assumed that the meaning of a text or document comes exclusively from the individual words whereby neither the ordering nor the constituency of the terms play a role. [JM08] This means that, for example, the meaning of the following phrases

"*I see what I eat* and *I eat what I see*" [JM08]

is considered to be the same. However, while the exact ordering of the words or their location within the text is ignored, the number of occurrences of each term within the document is substantial [MRS08]. Thus, the occurrences are counted, more specifically, extracted as features and represented as vectors in a vector space model which will be described in the following.

Overall, the bag-of-words model represents a basic view of information retrieval systems (Section 3.4.2) and has also been used in computer vision as the *bag-of-visual-words* approach, e.g., for content-based image retrieval to represent local image features as "visual words" (see Section 4.1.1).

3.3.3 Vector Space Model

A model that maps words, sentences, or documents into a high-dimensional vector space is known as the *vector space model*. Thereby, each word, e.g., from a corpus or a collection, is represented by a dimension in the space. The model then allows to easily relate words to each other whereby the spatial proximity correlates with semantic proximity between the words [MS01]. This correlation as well as the simplicity of the vector space model make it a widely used model for ranking systems in ad hoc retrieval (Section 3.4.2). Thereby, documents and queries are represented as feature vectors with *term weights* in the vector space.

In detail, assuming that N is the number of terms in a collection, then N is also the number of dimensions in the vector space. Then, for example, to compare a query q with a document d, both need to be represented in the vector space. Thus, a vector for the query q is represented as

$$\vec{q} = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

and a vector for a document $d_j = j$ as

$$\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{n,j})$$

with term weight w_i referring to the weight that a particular term $i \in [1, N]$ has in the query q or the document j. In other words, the term weight typically corresponds to the frequency, or a function of frequency, of a term in the query or the document. The similarity *sim* between the vectors of the query and the document are then calculated using the *cosine* of the angle between their vectors:

$$sim(\vec{q}, \vec{d_j}) = \cos \sphericalangle(\vec{q}, \vec{d_j}) = \frac{\vec{q} \cdot \vec{d_j}}{||\vec{q}||||\vec{d_j}||} \tag{3.2}$$

In space, the documents that are closest located to the query can than be considered the most relevant ones in the retrieval task. [JM08] More precisely, instead of calculating and comparing vector magnitudes, the angles between the vectors are employed calculating the *cosine similarity* (Eq. 3.2). Thus, relevant documents, i.e., closest vectors in high-dimensional space towards the query vector, are the ones with the smallest angle.

Word2vec. An efficient vector space model that maps words to vectors and is known as *word2vec* has been introduced by Mikolov et al. [MCCD13, MSC+13]. Word2vec is based on a continuous skip-gram model and provides a mapping of phrases into a 300 dimensional vector space. The mapping keeps and expresses a large number of precise syntactic and semantic relationships between words while compressing semantic similarity. It was developed on a part of the Google News dataset containing about 100 billion words and phrases.

In particular, they have developed an efficient approach to learn high-quality vector representations of words from a large corpus of unstructured text. Their continuous skip-gram model is trained to predict surrounding words of the current word. The surrounding words are sampled within a defined range before and after the particular word. [MCCD13] They have extended their model to enable training on phrases making their model even more expressive. The phrases are formed from unigram and bigram counts. [MSC$^+$13]

Further, they have demonstrated that performing basic mathematical operations on word vectors can actually support language understanding. In particular, meaningful results can be obtained through simple vector addition, e.g.,

$$\text{vec}(\text{“}\textit{Germany}\text{”}) + \text{vec}(\text{“}\textit{capital}\text{”}) \sim \text{vec}(\text{“}\textit{Berlin}\text{”}),$$

with ~ meaning that the vectors are close in space [MSC$^+$13].

We will exploit this word2vec model mapping the lines of our creative texts as well as the image tags of our image corpus into the word vector space (Section 6.3.3) to judge similarity between text lines and images on a textual semantic level (Section 6.3.5).

3.4 Information Handling

In order to handle textual information, two important technologies need to be distinguished. *Information Extraction (IE)* (Section 3.4.1) analyzes text to extract specific information directly from the natural language, whereas *Information Retrieval (IR)* (Section 3.4.2) deals with finding a relevant subset of texts from a larger set of documents [Cun05]. Thus, IE directly deals with the textual structure, whereas IR considers texts more as "bags of unordered words" [Wil97]. These directions are illustrated in Figure 3.7.

Both technologies can be applied on top of each other, i.e., as a preceding step for the other one. IR can be used to retrieve relevant documents for later analysis with IE techniques [Wil97]. Or, the other way around, the output generated by an IE system can be directly used for indexing in IR [Cun05].

For the task of text illustration, which is tackled within this thesis (Chapter 6), IE is first applied to extract relevant information from the textual input. Afterwards, in order to find suitable pictures to a given text (Sections 6.2, 6.3), the extracted information is employed to retrieve relevant images based on their associated textual meta information from a large collection. This is closely related to IR but with images being the requested information.

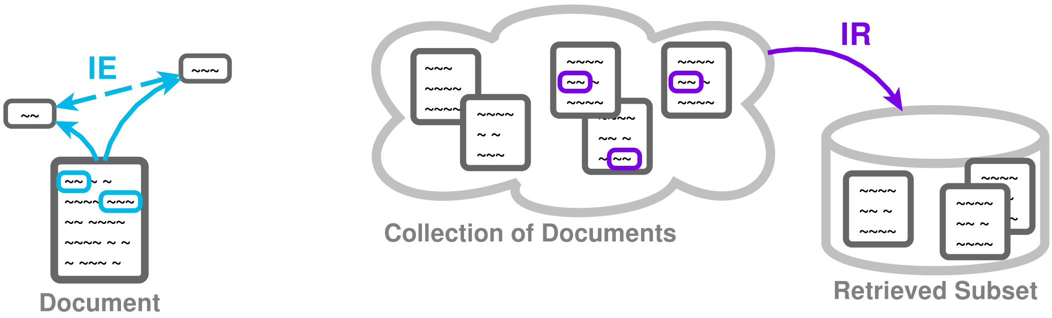


Figure 3.7: Comparison of the techniques Information Extraction (IE) and Information Retrieval (IR) in handling textual information. IR (right) retrieves documents from collections whereas IE (left) analyzes text to extract relevant information.

3.4.1 Information Extraction (IE)

Information Extraction (IE) forms an important part of NLP. It refers to the process that takes unseen documents comprising natural language as input and produces structured output from the explicitly stated or implied data which is found in the text [CL96]. Thereby, identifying specific information like names, dates, or certain events embedded in natural language, is a challenging task. Overall, the extraction problem is often considered as a combination of detection and classification [JM08]. Therefore, IE systems typically combine several techniques previously mentioned in Section 3.2 to solve problems like *recognizing named entities* or the *relationships* between them, just to mention a few.

Named Entity Recognition (NER). One of the most important tasks of IE is to identify the proper names, i.e., named entities like the names of people or places within a text. Thereby, the task of *named entity recognition (NER)* is a combination of detecting the ranges within the text that belong to such an entity and, then, classifying the found text segment with its according type, e.g., a person or a place. Thereby, mainly the following two types of ambiguity need to be resolved:

- *Same name → Different entities of same type.* This means that the same name can refer to different persons, e.g., "Mozart" can mean the father "Leopold" or the son "Wolfgang Amadeus".
- *Same name → Entities of different types.* This is the case if the same name can be associated with a person, a building, or a location and occurs often if, e.g., a building is named after a specific person. For example, "Guggenheim" can refer to the person "Solomon R." or "The Guggenheim" museum.

In general, NER problems are considered as the task of labeling sequences of words and are typically approached with techniques similar to POS tagging (Sec. 3.2.2), chunking (Sec. 3.2.2), or lookup-lists. In addition, these techniques can be employed by NER systems to extract suitable features from a training set of a representative collection of documents. Then, based on these extracted features, a sequential classifier is trained using machine learning and new sentences can be labeled by the trained classifier. [JM08]

An example of an output by an NER system with the bracketing notation of chunking is given in the following:

$$[_{PERS}\ \texttt{Manuel}\ ]\ \texttt{ate the candy in}\ [_{LOC}\ \texttt{Munich}].$$

Thereby, a *person (PERS)* entity named "Manuel" and *location (LOC)* entity called "Munich" are correctly identified and classified within the sentence.

To evaluate the performance of such systems, the named entities that are detected by the NER system are compared against human annotations using the standard IR metrics described in Section 3.4.3. Further, whenever all named entities are found, they can be clustered into sets that belong to the same entities under different names which is also known as *reference resolution*. [JM08]

Relations between Entities. As soon as entities are discovered within a text, a subsequent task is to identify the relationships between them, whereby especially binary relations are in focus. Some common relations are affiliations, part-of kinds, or geospatial ones. For example, assuming multiple person entities are detected within a text snippet, then a personal affiliation relation of the type

$$PER \rightarrow PER$$

that is supposed to be classified could be of the type *mother of* or *married to*. This problem can be solved using a supervised learning approach. Performance of such relation detection systems, i.e., if they have correctly detected all relations between the entities within a text, can again be measured using the IR metrics (Section 3.4.3) to evaluate the results of the system against manually created labels. [JM08]

Overall, extracting information like named entities or the relations between them are important problems within the field of Information Extraction. Huge attempts have been made to develop IE systems to solve such tasks when analyzing text as they allow to gain important facts out of documents. This thesis exploits the success of such IE systems and makes use of the generated output for further processing.

3.4.2 Information Retrieval (IR)

In general, the task of retrieving information simply comprises a request and a corresponding answer. Manning et al. [MRS08] motivate that, in principle, already getting the credit card from the purse to retrieve the number could be termed as such a task. However, as a field of study, they provide the following definition:

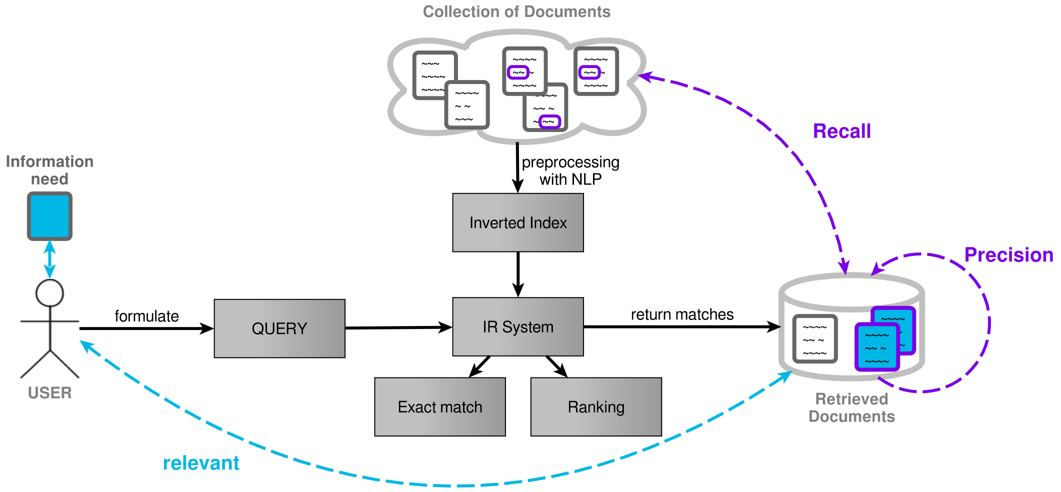


Figure 3.8: Components in IR. A user-formulated query as well as preprocessed and indexed documents from a large collection are compared in an IR system to identify matches and return relevant documents that meet the user's information needs. Approaches like exact matching or ranking systems can be evaluated by precision and recall measures.

> "Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." [MRS08]

In other words, the field of *Information Retrieval (IR)* typically deals with returning relevant textual documents from huge collections for textual search queries. A similar task is given in image search, with images being the retrieved documents. Related work on image retrieval will be outlined in Section 4.1.1.

An overview of the main components in the information retrieval process is illustrated in Figure 3.8. Typically, a user submits his information need in the form of a query. Then, an IR system matches the query with pre-processed and indexed documents from a large collection to return the relevant ones that satisfy the user's information need. In particular, a retrieved document is termed as *relevant* if the user considers the returned information valuable respectively his initial request [MRS08]. Overall, there are two main approaches, namely exact matching and ranking systems. The quality of the retrieval system can be evaluated by means of precision and recall measures which will be described in Section 3.4.3.

Document Content

In information retrieval, the content of documents consists of many rather unconnected words. Thus, word counts can be used to gain knowledge about the meaning of words in language, their rankings and their distribution within language.

Meaning of Documents. Overall, information retrieval systems often assume that the *meaning of documents* originates from the words rather than the structure between them. Thus, considering texts as sets of unordered words, i.e., ignoring the syntactic information between the individual words, retrieval approaches are typically seen as *bag-of-words* models (Sec. 3.3.2). [JM08]

Frequency Lists. As IR simply considers text as a set of individual words, *frequency lists* can be used to classify documents and, e.g., to identify the most common words within a text. Frequency lists are built by listing words together with their *word counts* and are typically sorted by the obtained frequencies.

Such word counts, i.e., counting the occurrences of the different words themselves or the different types of words, can be used to compare collections or to identify a certain text style. Generally, a quite uneven distribution of word types is typical. Often, a large amount only occurs once and makes it challenging for statistical NLP to gain knowledge about words that are barely observed. Such words that only occur once within a collection are also termed *hapax legomena* (Greek for "read only once"). Unfortunately, this problem still holds for larger corpora as some words are simply very rare. [MS01]

In this thesis, we will build a frequency list of relevant words from our creative text corpus to download a suitable subset of pictures from the YFCC100M data set for our image corpus as described in Section 6.3.4. The complete frequency list of over 400 relevant words is given in Appendix A.

Zipf's Law. The mentioned problem of a large existing amount of rare words is the basis for the early success of *Zipf's law* introduced by Zipf [Zip49]. This model can be employed to estimate the value of certain ranked items based on their rank. Therefore, word occurrences have to be counted within a large natural language corpus and a frequency list needs to be built in which the words are sorted due to their frequency of occurrence. Then, the *rank* of a word is simply the position of the word within the list. Therefrom, the relationship between the frequency f of a word and its rank r within the list can be explored. Zipf's law states that:

$$f \propto \frac{1}{r}$$

This means that, e.g., the most common word will appear two times the 2^{nd} most common word which is supposed to occur about two times as frequent as the 4^{th} most common word, and so on. Overall, following Zipf's law, the distribution of word frequencies within human language can be roughly described as consisting of a small number of quite common words in contrast to a great amount of words with rather low frequency, and inbetween, a middle amount of words with a medium frequency. [MS01]

Index Construction

In order to increase performance and avoid linear scanning of all texts every time a query is submitted, documents are *indexed* before being searched. To enable the generation of a suitable index, the documents need to be pre-processed using techniques from NLP.

Pre-processing with NLP. To create a suitable index for all documents in a collection, the text first needs to be analyzed linguistically. Thus, techniques from NLP (Section 3.2.1) are applied as a preceding step. Commonly used methods are:

- Tokenization
- Stop word removal
- Stemming

The textual data is *tokenized* to identify words that serve as *vocabulary* for building an inverted index [MRS08]. Besides, *stop words* (Figure 3.2) are typically identified and removed from the vocabulary as they are not as useful for search but might lead to a huge number of useless links in the index due to their frequent occurrence in text. Thus, removing the stop words also reduces the size of the inverted index [MS01]. Another commonly used NLP technique is *stemming* as it allows to match a certain query with documents that contain any of its morphological variants [JM08].

Inverted Index. Straightforward, linearly scanning all documents to find the ones which contain the query words would be a simple retrieval approach. However, to tackle web-scale collections of documents, support flexible query operations like match ranges around a specific word, or enable ranked retrieval, linear scans are too expensive or even not sufficient [MRS08].

Typically, an *inverted index* or *inverted file* is utilized as the central data structure in IR systems [MS01]. For all words occurring within a collection, an inverted index lists all documents containing the according word and, thus, eases searching for all documents that contain a query word [MRS08]. The structure of an inverted index

$$\begin{array}{ccl} \text{vocabulary} & & \text{documents} \\ w_1 & \rightarrow & d_3 \mid d_6 \mid d_{13} \\ w_2 & \rightarrow & d_1 \mid d_2 \mid d_{13} \mid d_{28} \\ \ldots & & \\ w_\Omega & \rightarrow & d_{113} \end{array}$$

Figure 3.9: *Structure of an inverted index. After extracting a vocabulary $V = \{w_1, w_2, ..., w_\Omega\}$ from the documents, an inverted index can be constructed in which each word w_i points to all documents d_j that contain the corresponding word.*

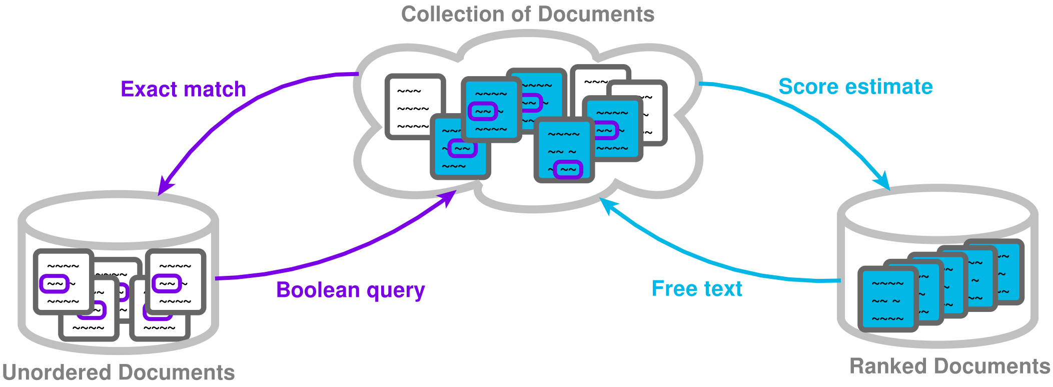


Figure 3.10: *Main approaches for ad hoc retrieval systems. Exact match systems like Boolean retrieval (left) return an unordered set of documents which exactly match the query, whereas ranking systems (right) return an ordered set of documents based on an estimated score.*

is indicated in Figure 3.9. In detail, after all documents are tokenized, a vocabulary $V = \{w_1, w_2, ..., w_\Omega\}$ is created consisting of all words w extracted from the documents. Then, for each word w_i in the vocabulary, links to all documents d_j containing the particular word are created.

Obviously, the construction of such an index is limited by hardware, especially for large collections as, for example, the complete vocabulary does not fit into memory. Thus, approaches dealing with web-scale solutions often have to distribute the index construction across computer clusters or support dynamic indexing to immediately reflect changes in the collection in the index [MRS08].

In this thesis, we will also create such an inverted file table for all image tags of our local YFCC14M Flickr set to allow efficient matching between query words and the tags of the 14M images (Section 6.3.3).

Ad hoc Retrieval Systems

The classical IR *ad hoc retrieval* task comprises a query entered by an unassisted user and a suitable list of documents returned by an IR system [JM08]. Thereby, two main models are commonly used and are illustrated in Figure 3.10. In *exact match systems* like *Boolean retrieval systems*, documents are returned if they precisely meet structured queries, for example, in the form of a Boolean expression. As such systems often lead to empty or unwieldy result sets, *ranking systems* became the preferred ones returning ranked results with estimated relevance towards the query [MS01]. Further, *web search* has become an essential retrieval task involving particular requirements arising from the explosion of data [MRS08].

Boolean Retrieval. *Boolean retrieval* is a model in which a query can be posed as Boolean expression, i.e., as combination of words connected with Boolean operators like AND, OR, or NOT. The retrieval system utilizes a precise language to exactly match the query expressions and, thus, "a document either matches or does not match a query" [MRS08]. As documents are only returned if they precisely meet the input query, empty result sets might be retrieved. On the other hand, if a large amount of documents match the input, the result set can get extremely unmanageable as there is no ordering provided. [MRS08]

Ranked Retrieval. Whereas Boolean retrieval systems require a precise language built upon Boolean operators, *ranked retrieval systems* allow for queries in the form of free text simply consisting of one or more input words [MRS08]. In contrast to exact matches, these systems then rank documents by estimating their relevance with respect to the query employing probabilistic methods [MS01].

A widely used representation is the *vector space model* described in Section 3.3.3. Especially its simplicity but also the strong correlation between spatial and semantic proximity make it a suitable model for ranking systems in ad hoc retrieval. As mentioned, the documents whose vector representations are closest to the vector representation of the query, i.e., the documents containing similar words as the query, are considered as relevant.

Web Search. With the growing availability of textual data in the web, an essential usage of NLP has become to query those large online repositories of textual information and extract meaning therefrom [JM08]. In *web search*, IR deals with returning relevant documents from these huge online collections for specific search queries whereby the documents are often web pages or parts of it. Although web search engines typically rely on ranking retrieval approaches, they meanwhile often include partial implementations of Boolean operators, e.g., for phrase search. However, whereas expert searchers exploit such options, most people rarely make use of them [MRS08].

Especially in the context of web search, IR systems need to efficiently handle tasks like indexing of documents on an enormous scale whereby a huge number of documents is widely spread over a huge number of computer storage. Additionally, such IR systems also have to deal with the hypertext explosion and content manipulations to boost web site rankings in search engines. [MRS08]

3.4.3 Evaluation of IR Systems

In order to evaluate the quality of information retrieval systems, more precisely, the relevance of the returned subset of documents, the two measures *precision* and *recall*

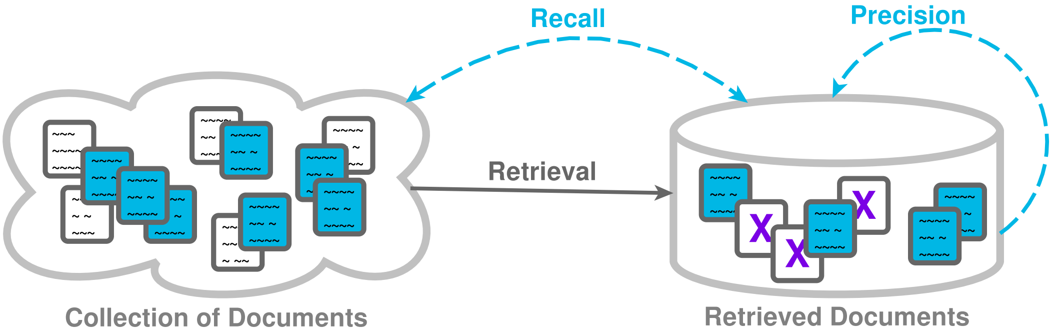


Figure 3.11: *Precision and recall measures to evaluate IR systems. Precision relates the relevant and non-relevant retrieved documents within the returned subset whereas recall correlates the relevant retrieved portion to all possible relevant documents available in the collection. Relevant documents are marked in cyan.*

are commonly used. For the evaluation, the following distinctions are made:

- relevant – non-relevant documents
- retrieved – not-retrieved documents

In particular, documents are *relevant*, if they satisfy the user's information need, and documents are *retrieved*, if they are returned by the IR system. The two measures are designed to provide answers to the following questions:

- "*Precision*: What fraction of the returned results are relevant to the information need?" and
- "*Recall*: What fraction of the relevant documents in the collection were returned by the system?" [MRS08].

Overall, those measures evaluate the relevance of the documents in the retrieved set. More precisely, the precision measure focuses on the retrieved documents relating the relevant and non-relevant ones within the returned set, whereas the recall measure correlates the relevant retrieved portion to all possible relevant documents available in the collection. The idea is illustrated in Figure 3.11.

In detail, assuming that the documents d_C in a collection C consist of relevant ones d_C^r and non-relevant ones d_C^n and the retrieved subset S of documents d_S consists of relevant ones d_S^r and non-relevant ones d_S^n, then the measures precision Pr and recall Rc are defined as:

$$Pr = \frac{|d_S^r|}{|d_S|} = \frac{\#(\text{relevant retrieved documents})}{\#(\text{retrieved documents})} \tag{3.3}$$

$$Rc = \frac{|d_S^r|}{|d_C^r|} = \frac{\#(\text{relevant retrieved documents})}{\#(\text{relevant documents})} \tag{3.4}$$

The precision *Pr* (Eq. 3.3) and recall *Rc* (Eq. 3.4) are the basic measures to evaluate IR systems [MRS08]. Note, that recall can only be evaluated if the amount of relevant documents available in the collection can be counted. Overall, although those measures have arisen from the task of evaluating IR systems, they are also used in IE systems to measure the recognition performance throughout a text.

This thesis will also make use of the precision measure to evaluate our approach for image retrieval by querying the web (Section 6.2.2). Thereby, images are considered as the documents.

4 Related Work on Text and Images

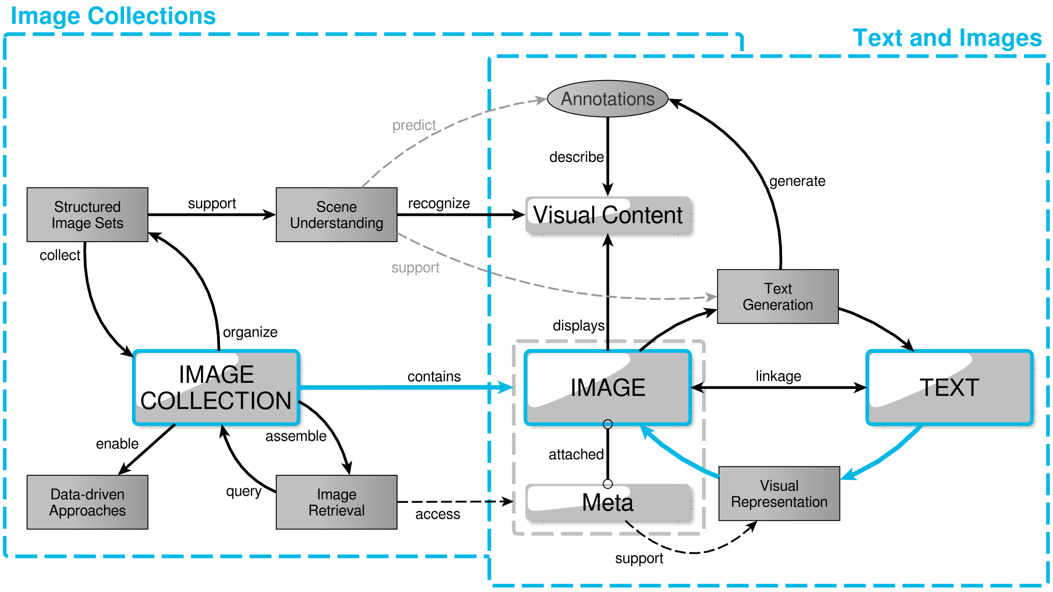


Figure 4.1: *Overview of previous work relating to images in connection with text. Online photo collections are enormous knowledge sources providing links between imagery and natural language and have provoked diverse research directions directly or indirectly exploiting the text–image links existing in such collections (Section 4.1). Related work has also investigated into the direct connection between text and images (Section 4.2).*

As indicated in Section 1.1, *text* and *images* present an intriguing natural connection that claims to be explored. Thereby, the previously mentioned enormous amount of online available data connecting natural language and images on a large scale has paved the way for diverse research in the areas of *computer graphics* and *computer vision*. Figure 4.1 shows a coarse overview of relevant research directions that have arisen from such image collections or benefit from the huge and still growing online available visual data and will be outlined in this chapter.

Several work directly makes use of the meta-data associated with the images, e.g., to retrieve relevant pictures, or to collect and structure sets from the large-scale visual data. Other methods rather indirectly exploit the text–image linkage as they focus on the huge visual knowledge provided by this source, e.g., data-driven approaches which depend on a large enough amount of images, or deep learning methods approaching scene understanding problems like visual recognition tasks which benefit from structured sets comprising diverse visual data. Corresponding research investigating in *image collections* will be discussed in Section 4.1.

Furthermore, techniques have arisen that directly explore the connection between *text and images*. Especially advances in scene understanding have supported research about translating pictured visual content into corresponding textual descriptions. Besides, the automatic generation of reliable image annotations like simple labels or detailed captions that precisely describe the depicted visual content can, for example, improve the retrieval task. At the same time, efforts towards the opposite direction, namely starting from text to create graphical representations like 3D scenes, or linking natural language with visual content like objects within pictures or whole images from online photo collections have also been made. Research exploring those directions will be surveyed in Section 4.2.

The present thesis relates to both, large-scale image collections as well as the direct connection between text and images (blue arrows in Figure 4.1). Throughout this thesis, we rely on visual data sets of varying sizes and diverse requirements for the specific tasks. In most cases, we exploit the enormous amount of images that is uploaded daily into online photo collections like Flickr with associated meta information, e.g., to derive aesthetically pleasing images with according user ratings (Section 5.2), or emotionally relevant images (Section 5.3). A huge amount of visual data providing initial links between images and natural language is also necessary to satisfy the needs of our illustration tasks (Sections 6.2, 6.3).

4.1 Image Collections

Meanwhile, online photo collections provide a tremendous amount of images associated with meta information and providing a wide range of visual quality and a huge diversity of displayed content (Section 1.1). Overall, in previous work, *image collections* have already been of interest for some time. Vast amounts of techniques have arisen to tackle the retrieval task (Section 4.1.1) which is fundamental to collect suitable subsets for further processing. Methods based on textual image annotations as well as pictured visual content have been developed whereby the aspect of efficiency has also been in focus to deal with large-scale data. Likewise, the recent explosion of visual data has enabled completely new approaches. Data-driven methods depend on the availability of a large enough amount of images and have arisen from those large community photo collections, e.g., to summarize large image sets, or composite new images from the data (Section 4.1.2). Further, exploiting the visual knowledge existent in the diverse online visual data, structured image sets have been created to support visual recognition techniques and, in particular, deep learning of visual content (Section 4.1.3).

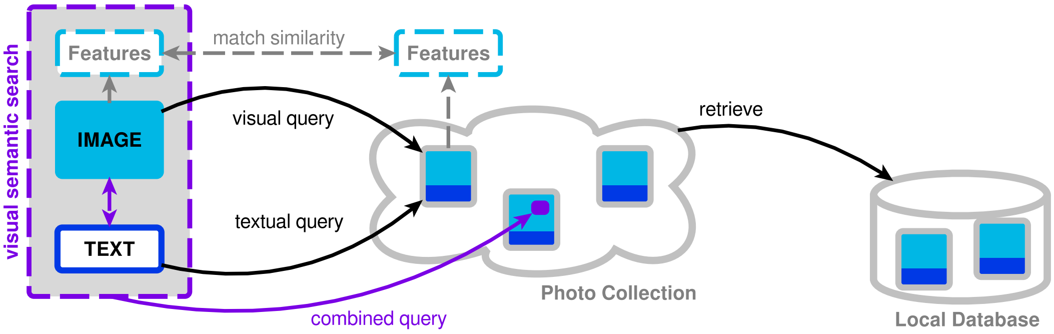


Figure 4.2: *Image retrieval and search. Queries are submitted to photo collections to retrieve images. Whereas text-based retrieval systems compare a textual query with image annotations, CBIR systems match visual features to find similar pictures based on a visual query. Connecting textual and visual knowledge allows for combined semantic search.*

4.1.1 Image Retrieval and Search

Retrieving images is the initial step to access online photo collections and assemble a visual set for further processing. As indicated in Figure 4.2, a query is typically submitted to an online image search engine, e.g., as provided by Google [1] or Yahoo [2], or a photo sharing site like Flickr to retrieve image results from the web that match the initial query. Text-based image retrieval systems typically compare a textual query with the annotations of the photos in the collection to retrieve matching images, whereas content-based image retrieval (CBIR) systems allow for visual queries and match features between the input sample and the data base to find similar ones. Recent success in connecting text and images allows combining knowledge from both modalities to search on a higher semantic level.

Text-based Image Retrieval. As mentioned in Section 1.1, textual queries are commonly used to search for images in online available visual data. Photo sharing sites like Flickr allow users to annotate their uploaded photographs with textual descriptions or a simple set of words, so-called “tags”, to reflect the pictured content or associate other information with a picture. This allows to find images based on textual input. Text-based image retrieval systems work similar to text-based information retrieval systems (Section 3.4.2) differing only in such that they consider the annotations of an image as a document. In order to return images based on their associated textual annotations, retrieval methods match the similarity between the input query and the image annotation and, thus, depend completely on the quality of the image description, i.e., how well the annotation describes the visual content of an image. Thus, as mentioned in Section 2.2.1, the semantic gap between

[1] https://images.google.com/
[2] https://images.search.yahoo.com/

the content an image displays and its associated meta-data attached by users or machines remains the biggest challenge for text-based image retrieval systems.

To support this retrieval task, research has investigated into several directions, for example, to assist users in the tagging process [SvZ08], refine given image tag lists [QHTM14], or enhance the quality of tag-based image search [YWHZ11]. Sigurbjörnsson and van Zwol [SvZ08] have developed tag recommendation strategies to directly support the user in the tagging phase based on observations they have derived from comparing image tags with WordNet categories. Their key observations comprise that user tags are not restricted to the pictured content, but also provide a broad context of information like the time or the location of the taken photograph. Further, an important retrieval task consists in ranking images due to their corresponding tags towards an input query. For example, Sun et al. [SBNNB11] have focused on relevance rankings and have proposed a system that allows empirical evaluation of such methods. Yang et al. [YWHZ11] have introduced a tag-based ranking scheme which combines relevance and diversity to avoid the appearance of irrelevant images in top search results. As users tend to submit imprecise input queries, their approach supports diversity in the image search results, in particular, they aim to cover several topics related to a simple input query. The aspect of diversity has also been considered by Qian et al. [QHTM14], but applied to refine image tags with the focus on providing greater diversity through the annotations. However, a broad survey of previous work in the field of tag-based image retrieval can be found in [SBNNB11]. In order to tackle the problem of noisy image responses in text-based search, this thesis presents a hierarchical querying algorithm to retrieve images of high precision for text snippets (Section 6.2).

Content-based Image Retrieval (CBIR). To overcome the semantic gap between textual annotations and pictured content, research has investigated in content-based image retrieval (CBIR) to search for similar images based on their visual content. As indicated in Figure 4.2, a query image is typically compared to a set of images by matching low-level image features. Therefore, most CBIR systems extract visual features like global color histograms or local shape descriptors in a preprocessing step [DLW05]. Then, feature vectors are calculated as image representations and compared to find visually similar matches. The distance between those vectors returns a value that indicates the similarity between the images. As outlined in Section 2.2.1, a challenge thereby is to identify meaningful visual features that a user would extract from an image. Overall, CBIR comprises techniques of representation, organization, and searching [FK15]. As great interest in research has occurred in CBIR, detailed reviews of previous work can be found in Smeulders et al. [SWS+00] and Datta et al. [DLW05, DJLW08]. Also focusing on visual information, a recent survey has been presented by Li et al. [LUB+16] which intensively studies the tagging process in terms of assignment, refinement, and retrieval to observe what in images is of relevance for people.

Initial attempts to extend traditional image retrieval, i.e., performing linear search over a data base to find closest matches to a query image, have been made by incorporating previously defined categories. Vasconcelos [Vas01, Vas02] have exploited hierarchical data representations for the task of image retrieval. Based on a data set of images clustered into several categories, Vasconcelos [Vas01] have introduced a hierarchical image indexing method which employs the density of those single image classes and have shown that incorporating multiple levels allows for a higher efficiency than traditional linear search of single matching images. This retrieval on the class-level has been extended by Vasconcelos [Vas02] to a search mechanism that additionally includes the image-level to improve precision.

Further, a key challenge of image retrieval lies in handling the huge amount of available image data [DLW05]. Thus, not only suitable features but also efficient methods to match similarity are required to successfully perform image retrieval on big data. Different descriptors and matching strategies have been used for large-scale image retrieval and similarity matching. Local descriptors such as *SIFT (Scale Invariant Feature Transform)* [Low99] have been employed in the *bag-of-visual-words (BOV)* approach [SZ03] which has been inspired by the *bag-of-words* model in NLP to represent unordered text (Section 3.3.2). More precisely, local image features are considered as individual visual "words" to build a visual vocabulary and an image is then represented as a vector containing the occurrences of those visual words. Sivic and Zisserman [SZ03, SZ08] exploit the efficiency of text retrieval and combine techniques like inverted file systems and document rankings (see Section 3.4.2) with methods from computer vision. In detail, a set of local descriptors is extracted at several salient image points, quantized and stored in an inverted file structure allowing for efficiently matching, e.g., an object with all frames of a video. Similar to bag-of-words models in NLP which ignore the syntactic information between textual words, the approach presented by Sivic and Zisserman ignores the global shape and spatial arrangement of an image. Zhang et al. [ZJC11] group multiple visual words to encode spatial arrangement in the inverted file structure. Another approach is based on global descriptors such as GIST which has been presented by Oliva and Torralba [OT01]. GIST forms a global description of the image with a low dimensional vector while preserving the spatial structure. Because of its low memory requirements it scales up to very large databases [DJS+09]. Johnson et al. [JGRF10] used GIST to organize large photo collections on the GPU with a SIFT-based geometric verification to further refine the ranking generated by the global descriptor. Whereas images of similar scenes do not necessarily show the same objects with similar geometric layout, a certain combination of features is typical. So, learning and classification methods have been used in combination with local and global descriptors. Xiao et al. [XHE+10] evaluated such descriptors on a large data base, the *SUN database*, which will be described in Section 4.1.3. Shrivastava et al. [SMGE11] have proposed a computationally intensive method to find visually similar images over different domains, learning features that are most important for a particular image. Their approach matches input images,

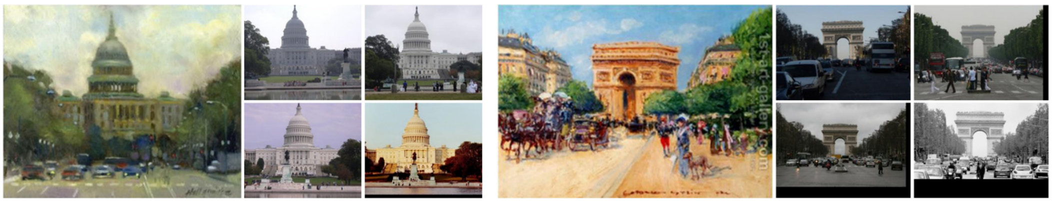


Figure 4.3: Approach for matching images over different domains by Shrivastava et al. The examples show paintings (large images) that are matched with photos taken, e.g., under different lighting conditions (Reprint from [SMGE11], Fig. 8, bottom row).

sketches, or paintings with photos that are taken, for example, under different lighting conditions. An example of their results is shown in Figure 4.3. Further, a promising descriptor that is invariant across different domains has been introduced by Shechtman and Irani [SI07] and has been employed by Chatfield et al. [CPZ09] to retrieve deformable shapes. More details on matching visual self-similarities and, in particular, the self-similarity descriptor [SI07] will be presented in Section 5.1.2.

Inspired by this line of research, this thesis also explores visual similarity in retrieved image data and will present an approach that finds similarities over different modalities on a large-scale (Section 5.1). In contrast to the mentioned learning and classification methods [XHE+10, SMGE11], we aim at efficiently finding similar images across a variety of domains without any prior learning steps. Therefore, we developed an efficient GPU-based version of the mentioned promising but computationally expensive self-similarity descriptor [SI07].

Visual Semantic Search. More recently, rather than retrieving images based on a simple textual query or performing content-based feature matching between images, incorporating semantic knowledge has become of high interest to search for semantically related visual data. As indicated in Figure 4.2, semantic knowledge is typically gained from exploring the connection between textual and visual information. Especially in the field of visual recognition, large attempts have been made during the last years to recognize visual content (Section 4.1.3) or to reason about objects and relationships which successfully enabled tasks like automatically generating complete sentences describing an image (Section 4.2.1). Overall, visual semantic search combines previously described retrieval mechanisms with such semantic relatedness.

Approaches have focused on combining textual and visual information to search objects in images [AZ12, HXR+16] or videos [LFKU14] as well as linking identified visual elements along video frames [LFKU14] or images in collections [HGO+10] to support browsing. Starting from a textual input query, work has focused on retrieving images depicting a particular object [AZ12] or retrieving locations of a particular object within an image [HXR+16]. Arandjelović and Zisserman [AZ12] have presented a 2-step procedure to retrieve all images that show a specific object based on a textual query. Therefore, they first search samples utilizing the Google image search engine and, then, use those samples to visually query the target database. Hu et al. [HXR+16] present an approach to localize an object within a given image based on a natural language query. Given a set of candidate locations in the input image, their recurrent network based approach combines local descriptors, spatial information about objects within the scene, and global context to calculate predictions for the initial candidate locations in relation to the textual query. As proposed by Lin et al. [LFKU14], textual queries have also been employed to assist in visual semantic search in videos. In their approach, they parse the video to detect visual elements like objects and transfer given textual video descriptions into a semantic graph to finally link, e.g., nouns with objects, or actions with verbs. Similarly, the enormous number of images includes many pictures that are somehow linked with each other, e.g., if they show similar objects under different perspectives or lightening conditions. Identifying such correlations is capable of presenting the underlying structure of the collection and eases the process of browsing such image data. Therefore, Heath et al. [HGO+10] exploit those implicit links and transfer them into a graph representation they call "Image Webs".

Further, semantic reasoning between objects has been studied by Zitnick et al. [ZP13, ZPV13] in the context of composing abstract clip-art scenes from textual descriptions (Section 4.2.2). They have demonstrated that image understanding benefits from fine-grained recognition of visual semantics and, thus, also improves image retrieval. In order to decode semantic information between objects, scene graphs have been utilized to improve semantic image retrieval [JKS+15, SKC+15]. Johnson et al. [JKS+15] have presented an approach that models objects together with attributes as well as relationships between those objects in such scene graphs which are then directly used as input to query for semantically related images. However, requiring scene graphs as input is rather inconvenient for users. Thus, Schuster et al. [SKC+15] have proposed an approach to parse image descriptions into such a scene graph which can than again be utilized as input for the retrieval task.

Overall, the focus of the described methods lies on retrieving suitable images from the web or searching and linking visual data matching similarity. However, other data-driven approaches have arisen exploiting the huge visual knowledge present in online photo collections and will be outlined in the following.

4.1.2 Data-driven Approaches on Web-scale Photo Collections

As previously mentioned, online photo collections like Flickr provide a tremendous amount of images with associated meta information allowing to assemble sets under varying aspects or for different topics such as, e.g., specific events or locations. Thus, this highly diverse imagery paved the way for many new data-driven approaches as well as enhancements or complete redesigns of existing techniques. To give a broad overview, we outline several techniques that make use of such data, e.g., for image compositing or scene reconstruction, collection-based summarization, or even extraction of underlying storylines. Some examples for data-driven approaches are shown in the Figures 4.4, 4.5, 4.6.

Image Compositing and Scene Reconstruction. In previous work, large community photo collections have been successfully exploited for a variety of different tasks towards the generation of new visual data such as image compositing [ADA+04, GSN07, CCT+09, EHBA09] or scene reconstruction [Sna09, FFGG+10], or enhancement of existing images by adding new data into existing imagery [LHE+07, HE07].

Typically, sketch-based compositing methods query photo collections based on simple freehand drawings created by users to composite a seamless image [GSN07, CCT+09, EHBA09]. Initial attempts to create a composite image with the aid of stroke drawings have been made by Agarwala et al. [ADA+04]. Starting from a small set of photos that show a similar scene, e.g., taken from different viewpoints, they combine parts of those images to enhance a single image in a user-guided process. Larger photo collections have been exploited by Gavilan et al. [GSN07]. Based on a rough sketch, their approach matches color composition as well as individual objects from the drawing with the photos in a collection to compose a collage. Chen et al. [CCT+09] utilize online photo collections for automatically creating photorealistic images from rough freehand sketches annotated with simple textual labels. To filter out undesirable search results, they propose an appearance similarity based filtering scheme arguing that images with similar content usually share similar appearance. Eitz et al. [EHBA09] propose a user-in-the-loop system to support the creation of a composite image. Based on user sketches without textual annotations, they query a large database of over 1.5 million images. Similar to Agarwala et al. [ADA+04] they provide a stroke based interface to select final regions which are then blended into a seamless composition.

Other approaches utilize community photos to insert new objects [LHE+07] or regions [HE07] into existing photographs. The approach presented by Lalonde et al. [LHE+07] uses a publicly available internet object database for the object insertion task to support the user in image compositing. Instead of investigating in manipulations such that the object fits the image, they focus on retrieving objects that already comprise the needed properties. Also starting from an existing image but with a missing region, a data-driven approach for automatic scene completion is proposed by Hays and Efros [HE07] (Figure 4.4). Instead of directly matching regions, their method starts by searching for similar scenes based on structure and color information in an image data base retrieved from the web. This approach allows them to patch holes in images with semantically valid regions without the need for user labels. Finally, they output multiple seamless composite images.

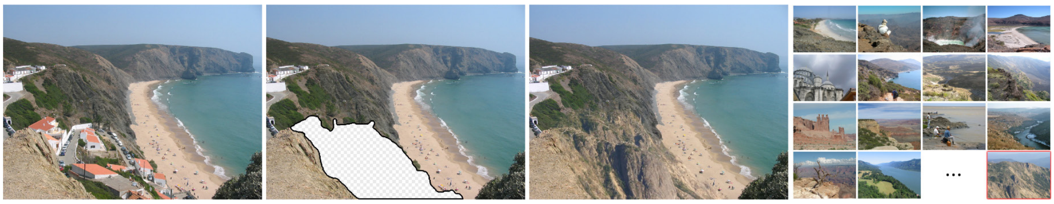


Figure 4.4: Data-driven approach for scene completion by Hays and Efros. To fill a missing region in an image, similar scenes are matched from a large photo collection. Left to right: Original image, input, output, matching scenes. (Reprint from [HE07], Fig. 6, first row).

Generally, image compositing often requires to search for particular regions or objects to synthesize new images or insert visual data into existing pictures. Especially searching and integrating sky regions is quite common. But also existing pictures sometimes require a replacement of the sky region, e.g., if pictures of events are taken under a quite gray sky and lead to a rather boring appearance of the photograph. Thus, to ease searching for a particular type of sky, Tao et al. [TYS09] have developed an interactive search system, called SkyFinder. They have downloaded over half a million sky images from the internet, automatically extracted semantic sky attributes (e.g., layout, richness, horizon) in an offline process, and now provide an interactive online search engine. Further, utilizing search engines to query for content usually leads to noisy results due to the weak nature of associated text on the internet. A method to identify and distill relevant images of a requested object from these large unstructured data collections has been proposed by Averbuch-Elor et al. [AEWQ+15]. Employing an algorithm for shape clustering based on previously extracted outer segment contours of the objects, they filter outliers with shapes that do not appear in tight clusters. This work can be applied as a pre-processing step to methods that require clean object sets.

Many of the mentioned methods [LHE+07, HE07, CCT+09, EHBA09] demonstrate that realistic images can be easily synthesized by utilizing the visual knowledge existent in large photo collections instead of relying on a complex underlying model. Inspired by that work, Johnson et al. [JDA+11] also make use of a large collection of photographs for matching color, tone, and texture regions to a computer-generated image in order to give them a more realistic appearance.

However, beside of the successful utilization of a large number of pictures to synthesize 2D images, research has also exploited web-scale photo collections for 3D scene reconstruction and visualization, for example, based on many images showing a particular landmark from multiple viewpoints [SSS06, FFGG+10] or to estimate the reflectance of a scene from Flickr images taken under distant lighting [HFB+09]. Initial success in reconstructing 3D points and viewpoints from unstructured web photographs using structure from motion (SfM) has been achieved by Snavely et al. [SSS06]. Their presented photo explorer application allows 3D navigation to browse the pictures. Focusing on efficiency to approach the large and growing online available amount of visual data, Frahm et al. [FFGG+10] have demonstrated large success in dense 3D reconstruction with high computational performance.

Image-based Summarization. Usually, photo collections cover a variety of different topics. Thus, methods for summarization aim at identifying those main themes within a collection [RKKB05, RBHB06], determining typical elements for a certain area [DSG+12], or creating summaries of images depicting similar objects to perform set-based manipulation [NNRS15].

To represent the highlights of a personal photo collection, several research has chosen a similar output as previously described compositing systems, namely combining the resulting images into a seamless collage [RKKB05, RBHB06]. The finally synthesized image claims to be an automatic visual summary. The approach proposed by Rother et al. [RKKB05] synthesizes one output image from a set of different consumer photos and has been extended in [RBHB06] to tackle computational complexity with a multistage optimization and scale to larger image sets. Further, a method summarizing visual elements that are typical for a specific geo-spatial area has been proposed by Doersch et al. [DSG+12]. Their data-driven approach relies on geo-tagged images and employs discriminative clustering to automatically discover specific elements, for example, windows or balconies that distinguish a city like Paris from other cities. The approach presented by Nguyen et al. [NNRS15] aligns images that show a specific objects towards a reference image with the aim of manipulating the set. They learn the limits of valid image manipulations based on such a collection of exemplar images while focusing on edits related to an object's shape as well as its appearance.

This thesis also presents work on manipulating the appearance of images to influence how an image is perceived by an observer (Section 5.3). However, the color manipulations performed to develop our EmoTune image filter are carried out independent of other images.

Story-based Visual Summarization. More recently, research has explored the generation of semantically meaningful summaries, i.e., extracting the underlying storyline of image sets [KX13, KX14, KMS15a] or videos [LGG12, LG13, XKS15, KHLS13, KSX14]. In this line of research, storylines typically describe the meaningful structure and consist of key events or common activities that repeatedly recur in the image data, or relevant moments in video streams.

Motivated by the huge number of online existing photo streams, Kim and Xing [KX13] propose preliminary research towards storyline reconstruction. They detect recurring moments in outdoor activities by aligning and segmenting different photo streams to find matching images. An example of their approach detecting collective storylines in photo streams for "scuba+diving" is shown in Figure 4.5. Building

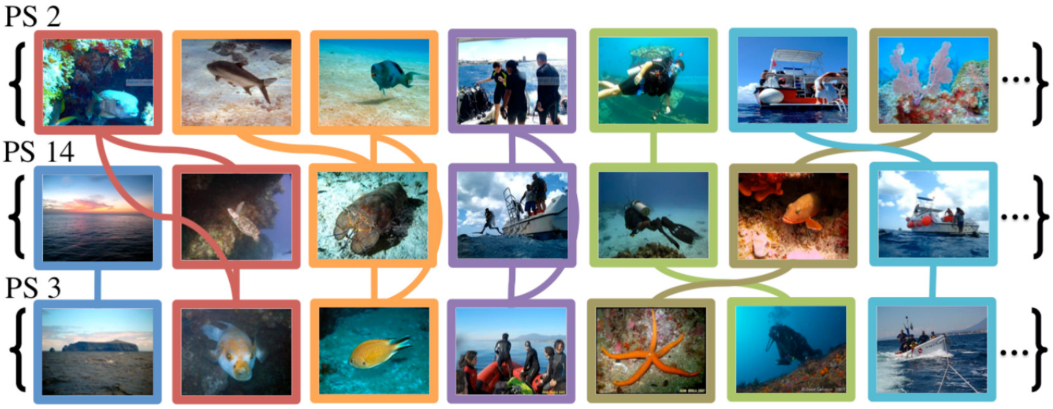


Figure 4.5: Visual storyline extraction from multiple online photo streams for "scuba+diving" through joint alignment by Kim and Xing. (Reprint from [KX13], Fig. 1(a)).

on this work, Kim and Xing [KX14] have investigated in summarizing sets and extracting storylines depicted in Flickr images of a particular event like the fourth of July. They output their structural summarization as a storyline graph. In addition to large image collections, Kim et al. [KMS15a] utilize the sequential structure of blog posts to aid in the story-based summarization task by jointly aligning blogs with photo streams, e.g., employing photo streams to interpolate between single blog images. As a key requirement of their approach lies in the availability of a large enough number of blogs and photo streams, they demonstrate their results on the popular event of visiting Disneyland.

Further, research has also investigated in creating story-based summaries of video streams. Generally, the main objective of summarizing videos is to enhance navigation in terms of simplification and efficiency. For example, egocentric videos tend to consist of long rather boring sequences between the meaningful personal moments. Thus, research has investigated in identifying the relevant scenes to ease navigation leading to summaries of the underlying story [LGG12, LG13, XKS15]. Those methods mainly differ in their representation. Whereas Lee et al. [LGG12] produce a compact storyboard depicting the key people and objects on a timeline, Lu and Grauman [LG13] present a story-driven summary of the identified video subshots of the main events. Xiong et al. [XKS15] extract multiple story elements, namely actors, location, objects, and events, to represent the storyline of egocentric videos allowing for higher flexibility as users can search based on their preferred elements. Other research has utilized large image databases to aid in the video summarization task [KHLS13, KSX14]. Khosla et al. [KHLS13] utilize web images as a prior for the creation of semantically meaningful summaries of user-generated videos. Similarly, Kim et al. [KSX14] use the images of a photo stream to identify keyframes in a user video, and, the other way around, utilize YouTube videos to enhance finding an ordering of a photo stream.

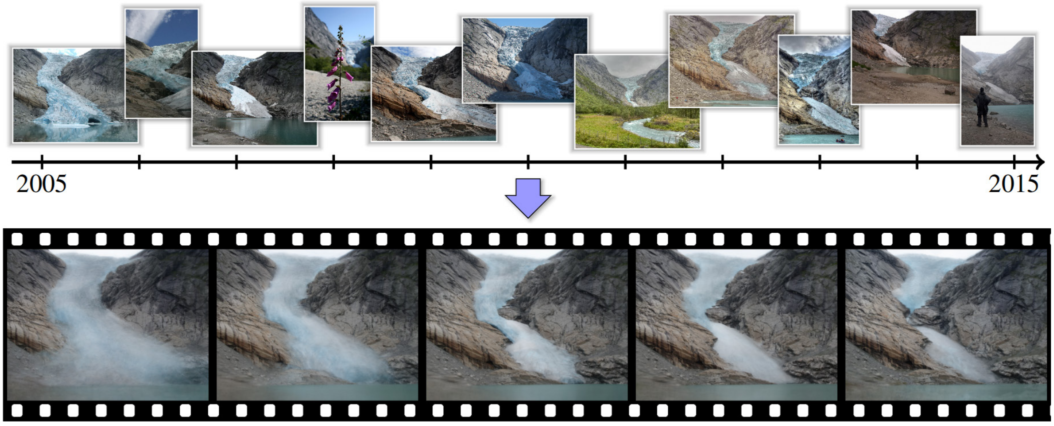


Figure 4.6: Data-driven approach employing online imagery to synthesize time-lapse videos of popular locations over years by Martin-Brualla et al. (Reprint from [MBGS15], Fig. 1).

Similar to this line of research, this thesis also presents work towards visual storytelling (Sections 6.2, 6.3). However, instead of summarizing an existing image set to extract the underlying story within the visual set, our work aims to build a completely new visual story from a textual basis. Besides, as we do not rely on images from a pre-existing set like a blog or a video, or a particular event, but deal with diverse topics, locations, photographers, day-times, and visual styles, the images differ largely in their visual appearance. Thus, to tackle visual consistency in a picture story, we present an approach that optimizes over different appearance constraints like visual style as well as image content features (Section 6.3).

Simulating Movements from Images. Large photo collections contain many pictures of similar objects or popular landmarks taken by different users at different times under varying perspectives, lighting situations, or even differing in their general appearance, e.g., similar objects in different colors. Work exists that explores groups of images with such slight appearance variations and aligns them to simulate dynamic movements. Initial attempts towards movement simulation from still images have been presented by Xu et al. [XWL+08]. They start from a still image showing a group of animals, extract the single motion states that are captured within the flock and combine them in a new sequence showing the motion cycle. Rather than starting from a single image, Averbuch-Elor et al. [AECOK16] extract “as-smooth-as-possible” sequences from image collections between a given image pair. Their method relies on a clean collection of foreground images of a specific category depicting different object instances. To identify smooth sequences they build upon shape distance as well as correspondences between contour points. Once a sequence is found, they provide data-driven morphs between the start and the end image. Large community photo collections have further been leveraged by Martin-Brualla et al. [MBGS15] to synthesize time-lapse videos of popular locations world-wide (Figure 4.6). Their data-driven approach exploits knowledge hidden in the data about changes in the world over the years. To tackle the vast variations of viewpoints and appearance, they obtain stabilization through a combination of structure from motion and stereo algorithms as well as a temporal filter.

In those approaches, not only obtaining a stable alignment but also a consistent appearance along the sequence plays an important role to avoid flickering, especially, if the sequence arises from different images [AECOK16, MBGS15]. In Section 6.3 we also aim for visual consistency along a sequence of pictures. However, in contrast to this line of research, we present an optimization over different visual features to obtain a consistent visual appearance in various visual styles.

To further exploit the diverse visual content existent in the enormous online available data, research in computer vision has collected and refined several sets from the web to support visual recognition tasks. An overview is given in the following.

4.1.3 Deep Visual Recognition on Structured Image Sets

Image data bases are fundamental to establish techniques that solve computer vision problems like scene classification or object detection. Besides, the enormous online available visual data carries a vast amount of knowledge reflecting the real world. Thus, visual recognition tasks benefit from this data as it comes with a huge diversity of different scenes, objects and relationships. However, the data is unstructured and rather noisy and needs to be prepared to support the development of suitable models. Thus, several benchmark sets of varying sizes have been created by collecting images from the web and provide structured and well labeled data. Further, within the last years, such large and structured data has largely supported the development of deep learning networks which further pushed the success in visual recognition. Deep learning methods have demonstrated their success on large-scale recognition challenges.

Structured Image Data Bases. The generation of image sets with well labeled visual data is of high importance as such structured data bases allow for fair comparison of existing techniques and support the development of new methods. Such benchmark sets consist of carefully selected images and thoroughly associated information and are typically built by collecting images from the web followed by a refining step, e.g., to clean the image set from duplicates, associate manual annotations like bounding boxes, or employ crowdsourcing with AMT (Section 2.3) to associate human labels.

A variety of smaller [FFFP04, GHP07, EGW+10, EEVG+15] and larger [TFF08, DDS+09, XHE+10, LMB+14] image sets have been created to serve as benchmarks for training and evaluation. Well labeled but smaller image sets like the *Caltech-101* [FFFP04], the *Caltech-256* [GHP07], and the *Pascal VOC (Visual Object*

Classes) [EGW+10, EEVG+15] have supported the development of many computer vision algorithms. The Pascal VOC created by Everingham et al. [EGW+10, EEVG+15] comprises a publicly available annotated data set and an annual competition which has been initiated in 2005 and presents several challenges, e.g., for scene classification or object detection. Although those data sets have contributed largely to establishing benchmarks, technical advances claim for larger data bases.

To evaluate the significance of image classification on a large scale, Deng et al. [DBLFF10] performed a study on a around 10K categories and about 9M images. Among others, their key findings comprise the challenge of large scale in terms of performance but, also, the importance of performing large scale categorization as knowledge gained from smaller data sets can not generally be transferred to large data. Further, their experiments indicate that a semantic hierarchy, like the structure existent in WordNet (Section 2.1.2), is capable to improve visual classification. Making use of the enormous amount of online data to support understanding scene content at a much larger scale, several data bases have been constructed employing WordNet into the generation process [TFF08, DDS+09, XHE+10] or focusing on objects within their natural context providing a large number of instance labels [LMB+14]. A set of around 80 million images has been presented by Torralba et al. [TFF08]. This so-called *TinyImages* data base consists of low resolution images which have been associated with loose labels out of around 75K nouns from WordNet and especially serves object recognition. To further support object recognition, Lin et al. [LMB+14] have shifted the task towards objects in their natural context within complex everyday scenes. They have proposed the *MS COCO (Common Objects in Context)* data set which presents a large number of labeled instances, in particular, 2.5M instances in 328K images. The focus of this data set lies on providing a large amount of instances per category to support the development of detailed object models. In order to support research in scene understanding, Xiao et al. [XHE+10] have introduced the *SUN (Scene Understanding) database* which was previously mentioned in Section 4.1.1. Focusing on the task of scene categorization, the SUN data base provides a big number of different scene categories, in particular, 899 categories and about 130K images. Starting from WordNet terms, they have selected the ones related to places and environments and have performed a refinement step resulting in 899 nouns of environmental scenes. Although they queried various search engines, several categories have resulted in rather few images. Thus, they end up with almost 400 well-sampled scene categories on which they have evaluated several state-of-the art scene recognition algorithms. Patterson and Hays [PH12] have extended the categorical SUN data base to support more fine-grained scene recognition through attributes providing 700 categories and around 14K images. However, a highly relevant large-scale data base has been presented to the community by Deng et al. [DDS+09] with the *ImageNet* described in Section 2.2.2. As mentioned, the ImageNet maintains the hierarchical structure of WordNet and meanwhile comprises over 5K synsets and around 3.2 million images. Thus, the ImageNet is not only large in scale, but also provides high diversity over various categories and enables a variety of tasks, e.g., image classification, or object recognition. Further exploring the ImageNet, Russakovsky and Fei-Fei [RFF12] have presented an approach to detect visual connections and discover new relationships between the given categories by learning semantic attributes.

Those mentioned benchmark sets have largely contributed to the generation and evaluation of many computer vision algorithms for various tasks. However, especially the ImageNet has significantly supported algorithmic advances in deep visual recognition.

Deep Visual Content Recognition. Fully understanding the visual content depicted by images comprises various visual recognition tasks ranging from detection and localization of single objects to studying the relationships between the objects, or categorizing the complete displayed scene. Therefore, huge efforts have been made in computer vision research to develop models that reliably detect objects or classify scenes with high accuracy and associate according textual labels. Besides, successfully recognizing visual elements, relationships, or the scene as a whole is a key requirement for text generation methods which, e.g., create captions describing the pictured content and will be outlined in Section 4.2.1. As previously described, the enormous amount of online available data has enabled the generation of large structured and thoroughly labeled image data sets which paved the way to successfully train suitable classifiers with fine-grained knowledge leading to high accuracy predictions in visual recognition tasks, e.g., the reliable detection of all objects in an image. Within the last years, such large-scale and well structured data has supported the development of deep learning networks which have demonstrated large success in visual recognition tasks.

To support the community with a large-scale benchmark for visual recognition algorithms, Russakovsky et al. [RDS+15] initiated the *ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)*, which they built upon the structured ImageNet data base described in Section 2.2.2. Following the previously mentioned earlier initiated PASCAL VOC challenge [EGW+10, EEVG+15], the ILSVRC competition similarly consists of two components, namely, carefully labeled visual data as well as standardized tests for tasks like image classification, object detection, object recognition, or object localization. Such competitions support the improvement of existing as well as the development of new algorithms. However, compared to PASCAL VOC, the ILSVRC provides a bigger amount of images and object classes, e.g, in its first year the ILSVRC2010 presented 1,461,406 images and 1000 object classes whereas PASCAL VOC 2010 consisted of 19,737 images and 20 object classes. This much larger scale of the ILSVRC enabled significant success in object recognition performance as well as relevant algorithmic advances in image recognition and retrieval [RDS+15]. In particular, deep learning methods such as convolutional neural networks (CNNs) have shown very successful performance in the ILSVRC competition and have emerged as powerful image representations for various tasks in object classification [KSH12, SZ14, OBLS14, SLJ+15, HZRS15, LRM15] or object detection [GDDM14]. In 2012, the *AlexNet* architecture presented by Krizhevsky et al. [KSH12] has been the first CNN that won the ILSVRC competition and, thus, laid the foundation for the usage of CNNs to solve computer vision tasks. With a relatively simple layout, their network demonstrated convincing performance as they outperformed other participants with a fair distance on the task of image classification with 1000 possible categories. Significant improvement has been achieved by increasing depths to 16 – 19 weight layers in a still relatively simple but deep network which has been presented by Simonyan and Zisserman [SZ14]. This powerful network is often referred to as *VGG* and performed well on image classification and localization tasks. Later networks have moved away from simplicity and sequential structure to rather complex frameworks like the *GoogleLeNet* presented by Szegedy et al. [SLJ+15] which participated, for example, with a 22 layers network at the ILSVRC competition. He et al. [HZRS15] proposed an ultra deep network, the *ResNet*, which they designed as a residual learning framework allowing them to train networks of substantial depth. For example, they evaluated residual nets with up to 152 layers on the ImageNet. To tackle the object detection task, Girshick et al. [GDDM14] successfully increased performance by combining region detection with CNNs in their *R-CNN* architecture instead of focusing on depth. Aiming at fine-grained visual recognition, Lin et al. [LRM15] have proposed a bilinear CNN architecture for image classification that comprises two feature extractors and supports pairwise interactions between them. Their architecture is based on networks trained on the ImageNet and combined with domain-specific fine-tuning to gain accuracy and speed.

In several parts, this thesis also builds upon the success of deep learning networks. We make use of this line of research in our visual storytelling task (Section 6.3) to match visual content similarity along image sequences. Therefore, we utilize the response vector of a deep neural network pre-trained for image classification, in particular, the VGG 16-layer model [SZ14], to judge content similarity between two images. Further, in order to approach the highly complex nature of visual aesthetics (Section 5.2), we propose a deep metric learning approach to position aesthetic relations in a high-dimensional space.

While this section outlined related work that deals with or benefits from image collections, the following section will survey previous approaches that investigate into the direct linkage between text and images.

4.2 Connecting Text and Images

Exploring the intriguing connection between *text* and *images* and, especially, identifying direct links between both types of information has become of growing interest in previous work. Thereby, the huge online knowledge source connecting pictures

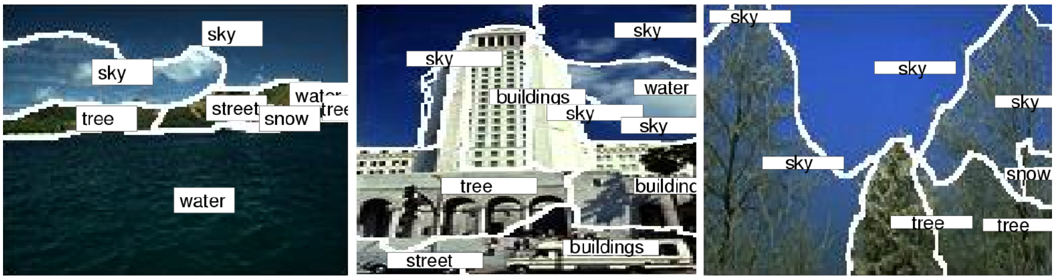


Figure 4.7: Approach annotating image segments with highest predicted words for corresponding regions by Duygulu et al. (Reprint from [DBFF02], Fig. 8).

and textual information along with technical advances to handle such big data succeed in supporting research to tackle the challenging semantic gap between visual representations and natural language (Section 2.2.1). Both directions, i.e., starting from images to produce text and vice versa are of substantial importance allowing for a variety of applications, e.g., providing descriptions of images or the surrounding environment to help visually impaired people, or supporting layman as well as artists in generating virtual environments or imagery. Several approaches have already investigated in translating visual content into natural language describing a picture (Section 4.2.1). But also employing natural language to generate 3D scenes or compose new images have been made (Section 4.2.2). Besides, initial attempts towards illustrating natural language by linking text to existing images from online photo collections has also been explored (Section 4.2.3).

4.2.1 Images to Text

Advances in scene understanding and, in particular, the development of deep learning methods (Section 4.1.3), which successfully contribute to the recognition of pictured content, have enabled research to extensively explore the relationship between images and natural language describing this imagery. Automatically attaching labels to image segments or detected objects (Figures 4.7, 4.8), or classifying complete scenes are fundamental steps to link visual data with textual information. However, automatically generating sentential image descriptions goes beyond identifying the components, e.g., objects or specific regions, of the displayed scene and understanding their relationships. The recognized information needs to be expressed in natural language. Advances in machine learning algorithms encourage this part of research allowing machines to learn from combined visual and textual sources. Examples of related work dealing with sentential image descriptions are shown in the Figures 4.9, 4.10, 4.11.

Region-based Labeling. A fundamental step towards connecting images with text has been explored on the pixel level in the area of semantic segmentation. Thereby, an image is segmented in semantically connected regions and textual labels are attached to those specific parts of a scene.

In the past, various approaches have investigated in linking image regions and individual words [BF01, DBFF02, BDF$^+$03, BJ05]. Barnard and Forsyth [BF01] combine words and image segments into a hierarchical statistical model to automatically annotate complete images with words. In order to further specify which image structure provokes which particular word, Duygulu et al. [DBFF02] have built upon the approach but focus on predicting keywords for specific image regions (Figure 4.7). They segment images and learn correspondences between the obtained regions and words whereby they group words with close meanings into clusters. Barnard et al. [BDF$^+$03] present and compare various models for the task of linking words with image regions. Inspired by achievements to predict words for images, Barnard and Johnson [BJ05] present preliminary work on how images can aid in the disambiguation of associated words. Based on images associated with 3 – 5 keywords, they learn a statistical model by jointly analyzing words and image regions.

More recently, to approach scene understanding in a more holistic way, research builds on semantic segmentation combined with knowledge from other computer vision tasks [LSFF09, YFU12, FSU13]. Li et al. [LSFF09] present a coherent framework that combines the tasks of classification, segmentation and annotation leading to a hierarchical output of semantic information for an image, i.e., the category of the scene, object regions with labels, and a list of keywords derived from the previously gained knowledge. Yao et al. [YFU12] investigate in a holistic scene model allowing for inference between the single tasks, for example, they incorporate object detection to aid in the segmentation problem. Following the holistic model of Yao et al., Fidler et al. [FSU13] additionally incorporate rich textual descriptions as input to improve performance. Therefore, short single sentences are parsed into nouns and prepositions, which are used to generate potentials in their holistic scene model.

Image Description Generation. In contrast to semantic segmentation, which aims to divide a complete scene into its regions and attach single keywords to those image segments, approaches for image description or caption generation exploit the success of visual recognition algorithms like object, person, or action detection and incorporate language models to provide semantically richer information.

Within the last years, the interest of research in visual recognition has shifted from solely detecting single objects or other specific classes towards exploring relationships between such visual entities and creating annotations for meaningful visual concepts. Sadeghi and Farhadi [SF11] have introduced such relationships as "visual phrases" and show that recognition of visual phrases works better than only

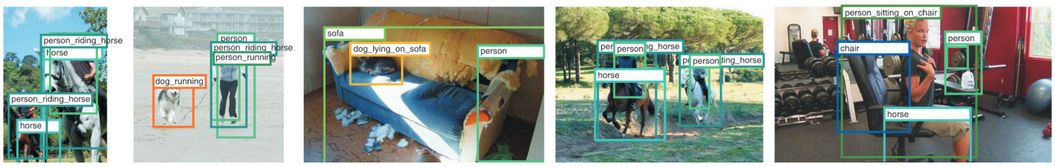
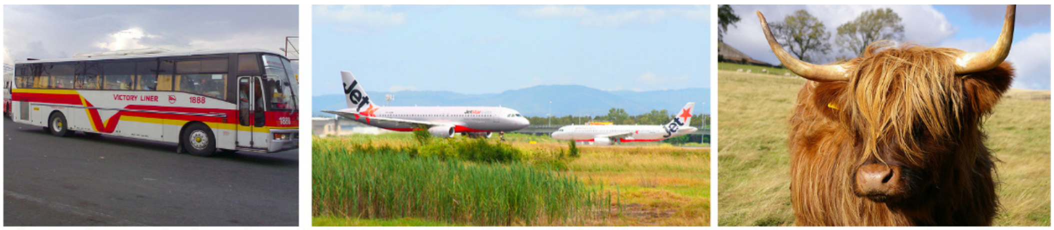


Figure 4.8: Detecting visual phrases in an image by Sadeghi and Farhadi. Decoding such relationships is often more accurate than solely object detection. The images show results after their decoding approach. (Reprint from [SF11], Fig. 6, images of last row).



This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

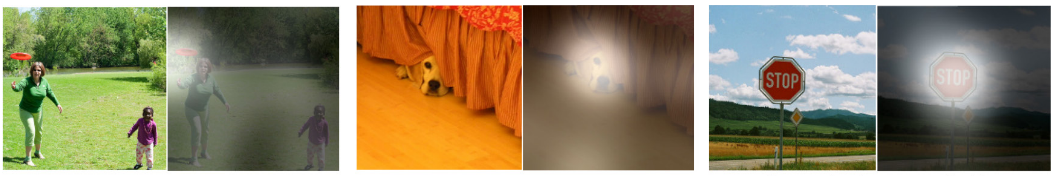
There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

Figure 4.9: Approach to automatically generate simple sentence descriptions for images by Kulkarni et al. (Reprint from [KPD+11], Fig. 4, first 3 examples in first row).

recognizing single objects and, thus, that recognition benefits from the knowledge about the relation between objects (Figure 4.8). Such concepts of relationships present a fundamental step towards complete sentence generation. For example, as indicated by Fang et al. [FGI+15], some abstract words like “beautiful” are hard to detect in an image and can not be clearly denoted by bounding boxes but have a strong correlation to visual patterns like “mountains at sunset”. As image captions typically provide rather salient information and allow learning a variety of concepts, Fang et al. directly train on image captions, learn to extract attached words from image regions, and generate novel captions from detected concepts.

Overall, automatically creating natural language descriptions from images has been presented in several lines of research [FHS+10, KPD+11, OKB11, KFF15, VTBE15, DCF+15, FGI+15, XBK+15, YJW+16]. Farhadi et al. [FHS+10] link images with simple sentences. They derive basic triple representations from an image consisting of an object, an action, and a scene which they then use to retrieve adequate sentences from a set. In contrast to retrieving sentences, other work has investigated in literally generating sentences. Kulkarni et al. [KPD+11] exploit statistics from parsing large text corpora as well as visual recognition algorithms and output relevant sentences for images (Figure 4.9). Ordonez et al. [OKB11] assemble a large filtered set of Flickr images associated with relevant captions to approach the challenge of generating simple image descriptions. Most recent approaches take advantage



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

Figure 4.10: *Caption generation approach combining deep learning with spatial model of attention (white regions) by Xu et al. Words that correspond to attention regions are underlined. (Reprint from [XBK+15], Fig. 3, first row).*

of deep learning based models and features [KFF15, VTBE15, DCF+15, FGI+15]. Thereby, two main directions have been established that both start from a CNN which is either followed by a maximum entropy (ME) language model [FGI+15] or a RNN [KFF15, VTBE15] to generate a sentential description. The output ranges from natural language annotations of image regions [KFF15] over image captions [FGI+15] to complete sentences describing a picture [VTBE15]. Devlin et al. [DCF+15] provide a comparison of those two directions. Other deep learning approaches incorporate models of attention mechanisms [XBK+15, YJW+16]. Whereas Xu et al. [XBK+15] propose a spatial attention model (Figure 4.10), You et al. [YJW+16] combine top-down and bottom-up strategies and focus on a semantic attention mechanism.

Rather than concentrating on the generation of a single description to a single image, research further investigated in considering images within a sequence to create captions that relate to each other [PK15, HFM+16]. Motivated by people taking series of pictures from specific moments, Park and Kim [PK15] present a multimodal architecture integrating convolutional as well as recurrent neural networks and learn from blog data to generate sequences of sentences along image streams. Further, aiming to shift basic understanding of visual content and the describing language itself more into the direction of human-like understanding and storytelling, Huang et al. [HFM+16] introduce a dataset that provides image descriptions on different levels. Their annotations range from captions for isolated images to story text for images considered within a sequence.

Visual Question Answering (VQA). Rather than providing a solid description for an image, research has also investigated in using the knowledge gained from the visual source to output a textual answer for a specific question [MF14, AAL+15, FPY+16]. In other words, given an image and a question, the task of visual question answering (VQA) is to automatically answer a question in relation to the image. Especially visually impaired users can benefit from this research direction [BJJ+10]. For the particular purpose of aiding blind people, Bigham et al. [BJJ+10] introduce a talking application that runs on mobile phones and links human workers to answer

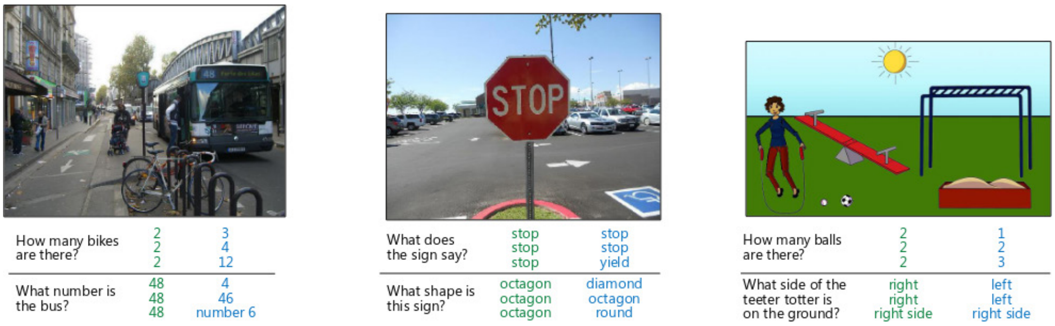


Figure 4.11: *Dataset providing answers to open-ended questions in VQA task by Antol et al. The examples show answers from AMT workers with (green) and without (blue) looking at the image. (Reprint from [AAL+15], Fig. 2, examples of last column).*

questions about visual content. However, recent approaches focus on automatic analysis and output generation. For example, Malinowski and Fritz [MF14] build on automatic visual scene analysis from semantic segmentation to automatically answer questions about realistic scenes. However, they rely on a predefined restricted world.

Antol et al. [AAL+15] have shifted the task to a version that claims to be "free-form and open-ended" and allows for higher diversity. They argue that, compared to solely producing image captions, the VQA tasks requires a more complex reasoning about the real-world. Thus, they provide a dataset consisting of real-world images as well as abstract scenes, open-ended questions, and natural language answers to support research towards this direction. Examples are shown in Figure 4.11. Another dataset, called "Visual Genome", has been collected by Krishna et al. [KZG+17] to further support such cognitive tasks like question answering which go beyond recognition and require to reason about objects, their interactions, and their relationships. As the task of VQA requires jointly modeling textual and visual information, Fukui et al. [FPY+16] have proposed a bilinear pooling approach to better capture associations between those modalities. They evaluate their approach on the two mentioned VQA datasets [AAL+15, KZG+17]. Focusing on the more specific direction of action recognition, Vondrick et al. [VOPT16] have explored the actual motivation for a depicted action to provide a textual answer for the question of *why* a particular action is performed by a person.

Video Descriptions. Naturally, the successful results of automatically generating descriptions of individual static images have encouraged research to transfer the task to dynamic video streams.

An earlier approach that associates schematic annotations with a video stream has been introduced by Goldman et al. [GCSS06]. They present a method to take short video clips and visualize them in a single static image using the visual language of storyboards (e.g., outlines, arrows, text). Instead of attaching specific motion words that describe the dynamics of the video stream, the generation of words or even complete sentences that reflect the content of video scenes became of higher interest [VXD+15, YTC+15, PMY+16]. Attempts to tackle the generation of video descriptions have been evolved from the recent success of Recurrent Neural Networks (RNNs). Venugopalan et al. [VXD+15] process short video clips and have proposed an end-to-end Long-Short Term Memory (LSTM) based model to generate sentences describing the main activity within the clip without the need of fixed sentence templates. Yao et al. [YTC+15] focus on incorporating the local as well as the global temporal structure, e.g., to maintain the order in which objects appear within the video. They model the local structure utilizing a 3-D CNN and make use of an attention mechanism to link global temporal information. Additionally to those approaches that rely on generating the next word locally based on previous words and visual content, Pan et al. [PMY+16] further investigate in relating the semantics between the visual content and an entire sentence within a visual-semantic embedding space. Overall, those attempts still provide rather short descriptions. Thus, Zhu et al. [ZKZ+15] tackle the problem by aligning books with their movie releases to obtain descriptions that are longer and richer and semantically on a higher level than previous ones.

In contrast, this thesis presents a converse approach that starts from a given textual description to assemble a set of relevant images and presents the resulting picture story as storyboard and, finally, renders them as video, more precisely as an animated slide show (Section 6.2). Further, we explore the generation of music videos in different styles based on our resulting picture stories (Section 6.3).

4.2.2 Text to Graphics

The previously surveyed methods (Section 4.2.1) have focused on understanding visual data to associate textual information like simple labels, captions, or complete sentences describing the visual content depicted by the imagery. However, the opposite direction, namely starting from a textual description to create a visual representation of language is even closer related to the aim of this thesis and the work we will present in Chapter 6. This direction has also been studied in a number of previous works that employ natural language to guide 3D scene generation or create 2D images.

Text to 3D Scenes. Several systems have been developed to render static 3D scenes from natural language descriptions [Win71, ADMF83, CW96, CS01]. Early systems like the SHRDLU program invented by Winograd [Win71] or the Put system from Clay and Wilhelms [CW96] have been rather limited in their linguistic input. Enhanced text processing allowing for more input flexibility has been incorporated by Coyne and Sproat [CS01]. They provide a graphics engine to render static 3D scenes from natural language descriptions and allow for user interaction to generate "natural" looking scenes including colors and textures. A more detailed overview of such systems that use language to guide the generation of virtual environments is given in Section 6.1.2. Similarly, this thesis also presents an approach for language based 3D scene generation (Section 6.1). Thereby, in contrast to previous work, we focus on automatically resolving spatial relations between 3D objects correctly while, at the same time, allowing for arbitrary textual input descriptions.

However, to generate arbitrary 3D scenes, a vast amount of suitable 3D models is needed. Online repositories like Google 3D Warehouse[3] are constantly growing which makes their accessibility more difficult. Thus, similar to the previously described task of retrieving images (Section 4.1.1), the retrieval of 3D shapes has also been of interest in research, e.g., based on informative shape descriptors [BBGO11] or supporting various input modalities including text [TD16]. A detailed overview surveying several approaches in the area of shape retrieval has been presented by Tangelder and Veltkamp [TV08]. Shape retrieval approaches often have to deal with the challenge of high variability in transformations, for example, as presented by Bronstein et al. [BBGO11] who have concentrated on the retrieval of nonrigid or deformable shapes. However, more recently a cross-modal retrieval approach has been presented by Tasse and Dodgson [TD16]. Their method combines labeled 3D shapes with semantic information existent within the labels in a vector space. They extend shape retrieval over several domains and allow for queries of multiple input modalities, e.g., words, sketches, or 3D shapes.

Further, instead of actually generating scenes, Kong et al. [KLB$^+$14] have presented an approach to tackle the Text-to-Image co-reference problem, namely identifying which visual objects within an image a text refers to. Natural sentential descriptions of RGB-D scenes are thereby exploited to improve 3D semantic parsing. As they deal with dependent sentences, they also address the problem of co-reference within the lingual descriptions.

Text to Images. Within the last years, work aiming at composing 2D images from textual descriptions has been evolved. An approach for 2D abstract scene generation has been presented in [ZP13, ZPV13], modeling scenes using clip-art images produced by workers on Amazon's Mechanical Turk (Figure 4.12). Whereas Zitnick and Parikh [ZP13] focus on the relation between single words and visual features, Zitnick et al. [ZPV13] extend their approach towards learning the correspondence between visual features and semantic phrases they derive from simple sentences. Therefore, they extract basic textual structures consisting of two nouns and a relation whereas subtle changes in the relation can lead to different visual interpretations. Kottur et al. [KVMP15] further exploit such abstract clip-art scenes to learn visual grounding as semantic context for words. They motivate that certain words might not share an obvious textual relation as integrated in text-only

[3] https://3dwarehouse.sketchup.com/index.html

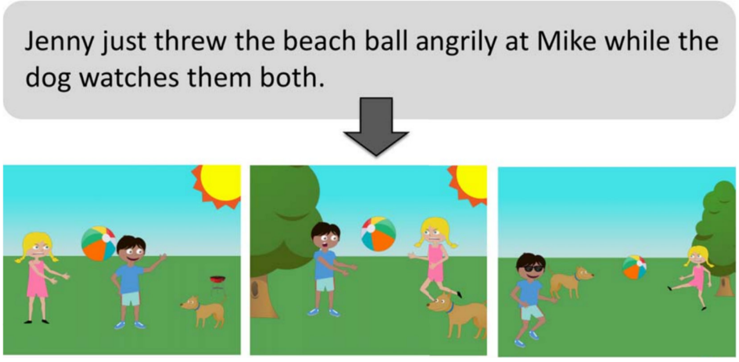

Jenny just threw the beach ball angrily at Mike while the dog watches them both.

Figure 4.12: Text-based generation of abstract scenes by Zitnick and Parikh. Images show similar scenes modeled by AMT workers. (Reprint from [ZP13], Fig. 1, first row).

word embeddings (word2vec, Section 3.3.3) but can still be visually connected and, thus, also semantically related. This means that, for example, although the words "eat" and "stare" do not seem to be closely related in a word embedding, they do share a visual connection as the action of eating often involves that a person stares at his food. Overall, although relying on abstract scenes, this line of research has presented a relevant step towards semantic reasoning in visual recognition and has supported several other tasks like visual semantic search (Section 4.1.1) or VQA (Section 4.2.1).

Recently, initial attempts towards the automatic generation of photorealistic images conditioned on textual descriptions has been proposed [RAY$^+$16, ZXL$^+$16]. They have utilized Generative Adversarial Networks (GAN) [GPAM$^+$14, DCsF15] which are able to successfully model complex multimodal data. Thereby, the biggest problem is to produce images of high resolution that show photorealistic details to a textual description. Whereas Reed et al. [RAY$^+$16] have been able to synthesize 64×64 images, Zhang et al. [ZXL$^+$16] have demonstrated a stacked version of GANs resulting in higher resolution images (256×256).

All these approaches aim at generating or composing completely new images from scratch. In contrast, this thesis largely investigates in dealing with existing imagery with primary focus on the huge online available amount.

4.2.3 Auto-Illustration

Most relevant to a primary aim of this thesis, namely automatically illustrating text using existing imagery (Sections 6.2, 6.3), only few approaches have yet been proposed to explore this direction.

Initial attempts have been made to connect text with existing imagery, e.g., to illustrate a certain event [JWL06, KMS15b]. Examples are shown in Figure 4.13.

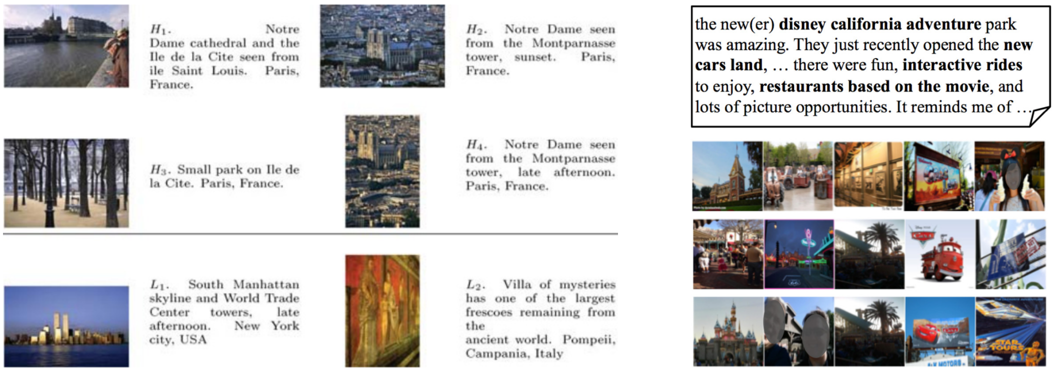

H_1. Notre Dame cathedral and the Ile de la Cite seen from ile Saint Louis. Paris, France.
H_2. Notre Dame seen from the Montparnasse tower, sunset. Paris, France.
H_3. Small park on Ile de la Cite. Paris, France.
H_4. Notre Dame seen from the Montparnasse tower, late afternoon. Paris, France.
L_1. South Manhattan skyline and World Trade Center towers, late afternoon. New York city, USA
L_2. Villa of mysteries has one of the largest frescoes remaining from the ancient world. Pompeii, Campania, Italy
the new(er) **disney california adventure** park was amazing. They just recently opened the **new cars land**, … there were fun, **interactive rides** to enjoy, **restaurants based on the movie**, and lots of picture opportunities. It reminds me of …

Figure 4.13: *Related work on auto-illustration. Left: Highest (top) and lowest (bottom) ranked images by Joshi et al. (Reprint from [JWL06], Fig. 4) illustrating a story about Paris. Right: Top ranked sequences with images from Disneyland retrieved for relevant words (bold fonts) by Kim et al. (Reprint from [KMS15b], Fig. 4, first example).*

Those approaches typically retrieve a set of representative images that display the idea of a piece of text describing a specific topic, e.g., a location or an event. Therefore, Joshi et al. [JWL06] incorporate an unsupervised ranking scheme to select pictures to illustrate a given story. More precisely, from a given text, they extract keywords with the aid of WordNet and polysemy counts and use those keywords to query candidates. Then, they use pairwise similarities (visual and lexical) to perform mutual reinforcement during the ranking and select the few highest ranked images to display the idea of the textual description. The types of stories they consider are short descriptions of a specific city like Paris (Fig. 4.13, left), a cultural landscape like the Loire Valley, or a piece of art. They work on about 125K well annotated images consisting of a personal photo collection as well as a digital library providing pictures of artworks in museums, all associated with well defined manual annotations. The method proposed by Kim et al. [KMS15b] learns an association between image sequences and multiple sentences (Fig. 4.13, right). Within a group of sentences, they use LSA to identify the most representative ones. Besides, blog posts are used for the training as they consist of image sequences and associated text. Content-wise, they select Disneyland as domain to present their approach due to the high amount of available blog posts with images. They mention that other popular events could be considered as long as a large enough set of blogs with images exists which is a precondition of their work.

In contrast, this thesis deals with a considerably bigger scale in terms of data as well as topics. Whereas Joshi et al. [JWL06] make use of a set of reliably manually annotated data, we utilize the enormous online available image data which is challenging as it comes along with noisy annotations. Further, the mentioned approaches [JWL06, KMS15b] focus on pre-selected domains. Although this leads to less noise and fewer outliers, the available themes and data are rather restricted. However, we aim at illustrating natural language that describes arbitrary subjects with meaningful and touching picture stories. Thus, we analyze images under different semantic aspects (Chapter 5) and build upon an extensive linguistic analysis (Chapter 6). We focus on a semantically close connection between text parts and images through the development of a hierarchical querying algorithm (Section 6.2) as well as an optimization procedure (Section 6.3). Further, we aim at maintaining the line of a given story instead of assembling a more global representative image set as presented by Kim et al. [KMS15b]. Therefore, we treat each text part with similar importance while searching for semantically close visual representations (Sections 6.2, 6.3). To emphasize the meaning of a story, we provide an optimization based approach along the storyline over semantic relatedness as well as visual coherency to retrieve a smooth visual appearance along the resulting picture story in various visual styles (Section 6.3). As the semantics of images is highly important for meaningful visual storytelling, we present work on analyzing images under different visual aspects in the following chapter.

5 Image Analysis

A single image comprises a variety of aspects that together lead to a specific appearance and transport visual information to an observer. The low-level image structure and the high-level meaningful appearance compose such visual information. Thus, in this chapter, we focus on analyzing those aspects to approach the semantics of an image. We exploit the image structure in terms of visual similarity to identify meanings and explore the connection between the visual appearance of images and an observer in terms of aesthetic appeal and evoked emotion.

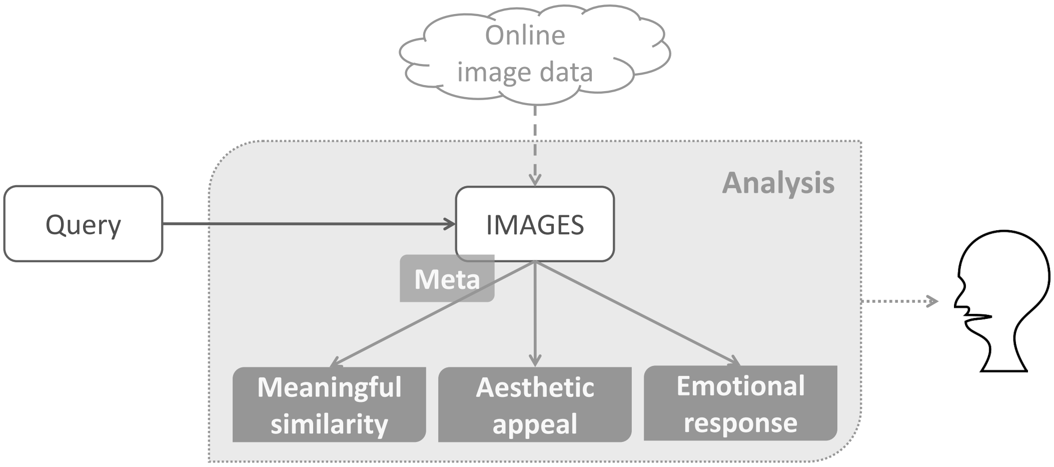


Figure 5.1: *Overview. In this chapter, we analyze the visual structure as well as the appearance of images in terms of meaningful similarities, aesthetic appeal, and emotional response of an observer. Images with meta data are retrieved from online collections through textual queries.*

Figure 5.1 gives an overview of the methods presented in this chapter to analyze the appearance of images under meaningful aspects. We explore the visual similarity of images and identify groups with similar meanings over different visual representations (Section 5.1). Further, we present an approach to learn rankings of the aesthetic appeal of images from online social behavior (Section 5.2) and investigate how simple global image modifications can actually change the emotional perception of an observer (Section 5.3). For those different meaningful aspects, we retrieve images from online photo collections with suitable textual queries according to the requirements.

5.1 Meaningful Visual Similarities

Throughout this thesis, the huge amount of online available data serves as a valuable knowledge source. However, as indicated in Section 1.1, the quality of the returned image results can differ and might lead to false positive ones, e.g., as image tags are not always reliable. Furthermore, in order to handle such a massive amount of visual information, algorithms often have to be redesigned.

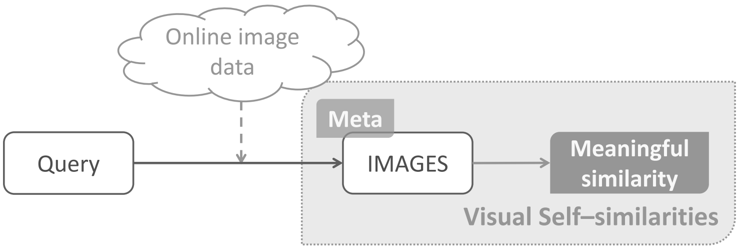


Figure 5.2: High-level concept. Image search results retrieved from online collections based on simple textual queries are explored to identify similar images due to their visual structure. Those meaningful visual similarities can resolve various meanings of ambiguous words.

The high-level concept of this section is indicated in Figure 5.2. Image search results are retrieved based on simple textual queries and explored to identify visual similarities in the data. We hypothesize that:

> $\mathcal{H}_{VisSim}$: *Different meanings can be identified by clustering visual similarities.*

As stated in $\mathcal{H}_{VisSim}$, we will show that our method is capable of identifying different meanings in search results from online collections by clustering the images due to their visual similarity patterns which leads to meaningful similarities that identify different senses of ambiguous textual queries (Section 5.1.5). To approach the visual structure of images and tackle the massive amount of visual data, we present an efficient approach to find visual similarities between images that runs completely on GPU and is applicable to large image databases. Based on local self-similarity descriptors [SI07], the approach finds similarities even across modalities. Given a set of images, a database is created by storing all descriptors in an arrangement suitable for parallel GPU-based comparison. A novel voting-scheme further considers the spatial layout of descriptors with hardly any overhead. Thousands of images are searched in only a few seconds. Additionally to applying our algorithm to cluster a set of image responses to identify various senses of ambiguous words, we will also present applications on re-tagging similar images with missing tags (Section 5.1.5) as well as locating an object within a real-world scene (Section 5.1.5). The work we present in this section is based on the following publication: [SHL12].

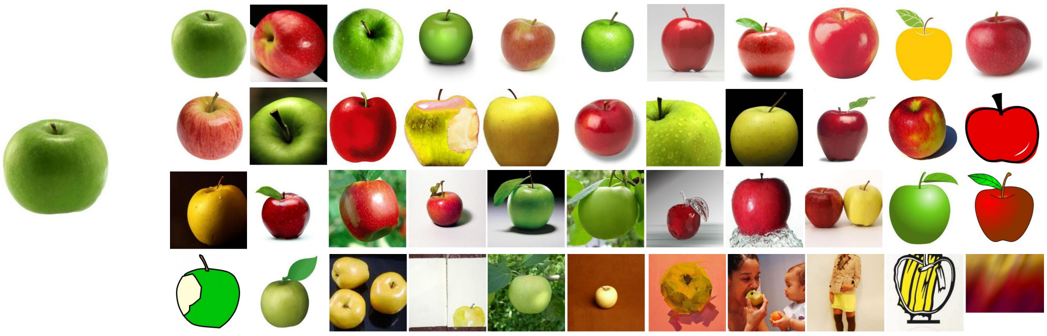


Figure 5.3: Some images (right) retrieved from Google Images by querying for "apple" and sorted by similarity to the template (left). Decreasing similarity from top left to bottom right.

5.1.1 Introduction

Finding similarities between images is a computational intensive task that is necessary in many computer vision applications, e.g., image retrieval and organization (Section 4.1.1), object detection or recognition. Typically, the comparison is based on extracted features representing important image properties. Mostly, it is assumed that multiple similar images share the same properties as well as the extracted features. A major challenge is the extraction of suitable features because they should be classified as looking similar if captured under varying lighting conditions, from slightly different viewpoints, or with partially occluded objects. The features have to account for changes in rotation, scale, illumination, color, texture, etc. Moreover, the set of common properties can vary drastically when taking images of various domains (photographs, drawings, sketches) into account.

In order to search for similar images, huge image databases such as Flickr typically focus on meta-data, tags, and textual search queries specified by users (Section 1.1) rather than the visual content. As previously mentioned, Flickr meanwhile contains billions of images and a simple search often yields thousands or even millions of results whereby the quality largely depends on the search term and the quality of the meta-data. Improvement on the quality of answers can only be achieved by taking, besides textual data, also the visual appearance of images into account. The local self-similarity descriptor introduced by Shechtman and Irani [SI07] encodes local similarities within an image region and successfully finds templates in other images. Unfortunately, since the computation of this descriptor is very expensive, applying it to large databases is a big challenge.

We present a variation of a self-similarity algorithm that makes it applicable to huge image databases. Descriptor generation and matching run completely on a modern GPU using CUDA. Due to our suitable representation of the descriptor database as well as a new voting-scheme considering the spatial arrangement with hardly any overhead, our implementation scales to databases with thousands of images that can

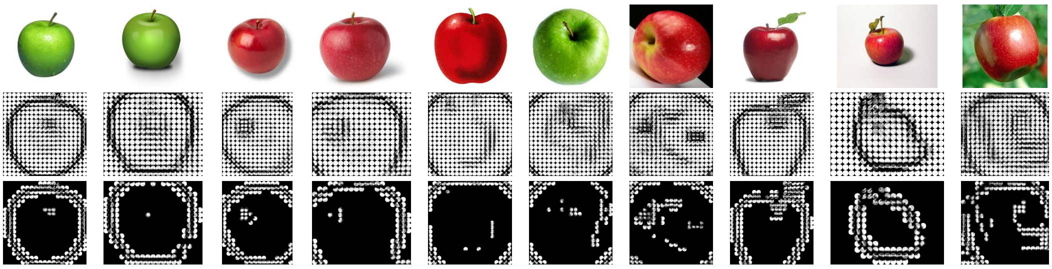


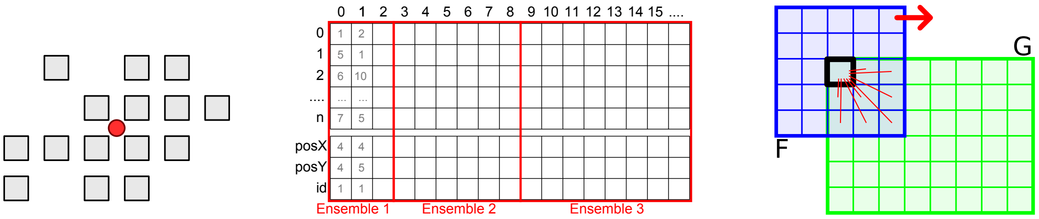
Figure 5.4: Some search results of Google images for query "apple" (top row) with our self-similarity descriptors (middle row) and only the informative ones (bottom row).

be searched in only a few seconds. Further, no additional pre-processing steps like learning or quantization are needed. Evaluation is performed with ETHZ, Caltech 101 and MIRFLICKR datasets as well as over one million images downloaded from Flickr. Even for matching a template with the large Flickr sets, we can compute over 1400 full image comparisons per second. Figure 5.3 shows a visual result of applying our approach to sort the search results we retrieved from Google Images for the query "apple" by similarity (right) towards a template image (left). In addition to our applications on meaningful clustering, image re-tagging, and object localization, our approach could also be used for real-time analysis of video streams.

5.1.2 Background on Matching Self-Similarities

As mentioned in Section 4.1.1, successful image retrieval largely depends on the selection of suitable features as well as an adequate matching strategy. However, the enormous growth of online photo collections requires efficient methods to match similarity between the visual features and sometimes claims to redesign existing algorithms to handle big data.

In particular, a promising but computational expensive descriptor that is invariant across different domains has been introduced by Shechtman and Irani [SI07]. Their local self-similarity descriptor, on which our implementation is based, has been developed on the observation that similar images do not necessarily share properties like colors, textures, or edges. Thus, measuring them is not always sufficient for comparison. These images are similar because their local intensity pattern is repeated in nearby image locations in a similar relative geometric layout which is captured in this descriptor. In detail, the self-similarity descriptor is generated by measuring the similarity of an image patch within its surrounding region. The sum of squared differences (SSD) between a 5×5 image patch centered at an image pixel and all 5×5 patches in the surrounding region is calculated and normalized, leading to a correlation surface that is subsequently transformed into a binned log-polar representation with 80 bins. Some of the descriptors are non-informative, because they do not capture any local self-similarity or they capture too much.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	....
0	1	2															
1	5	1															
2	6	10															
....	...	...															
n	7	5															
posX	4	4															
posY	4	5															
id	1	1															

(a) Ensemble (b) Database layout (c) Ensemble matching

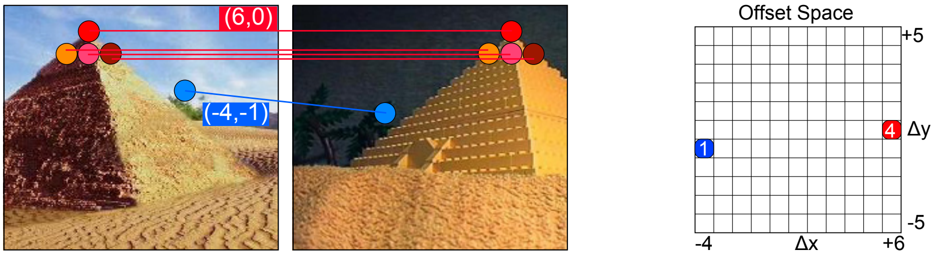
Figure 5.5: *Approach. (a) Ensemble is formed by informative descriptors (rectangles) capturing the spatial arrangement around the center (red dot). (b) Layout for storage of ensembles in device memory. (c) All descriptors in ensemble F are compared to all descriptors in the database. Matching descriptors cast a vote on the spatial arrangement.*

Non-informative descriptors are discarded and the remaining descriptors of an image form a global ensemble. Figure 5.4 shows the self-similarity descriptors generated by our approach (middle row) and the informative ones only (bottom row) for some search results from Google Images for querying “apple” (top row). Ensembles are similar if the distance between the descriptor values is small and the spatial arrangement of the descriptors is similar. Boiman and Irani [BI07] have used the self-similarity descriptor to detect objects in images based on both freehand sketches and real images with an optimized ensemble matching strategy. Their elimination of comparison calculations at locations where the similarity is probably very low leads to a scattered memory access pattern that does not fit well onto the GPU. Thus, we developed a different GPU-optimized strategy.

Further, as indicated in Section 4.1.1, Chatfield et al. [CPZ09] make use of the self-similarity descriptor to retrieve deformable shapes. They refine the sparsification of descriptors and study the influence of quantization on matching performance for large-scale retrieval using a BOV approach. In contrast, we do not need any quantization and, thus, avoid errors therefrom. Moreover, time-consuming generation of vocabulary is not necessary.

5.1.3 Approach and Implementation

Based on the self-similarity descriptor [SI07], we developed a simple approach that enables searching for similar images to a given template or calculate similarity values for an image set (e.g. Fig. 5.3) also in huge databases. Our system first creates the informative self-similarity descriptors of an image (Fig. 5.4) that form an ensemble (Fig. 5.5(a)) and stores the ensembles of all images in a database suitable for parallel GPU-based comparison (Fig. 5.5(b)). This database of ensembles is only created once and matching then operates directly on the database without any further pre-processing steps by comparing ensembles (Fig. 5.5(c)). Our efficient



(a) Similar descriptors in different images

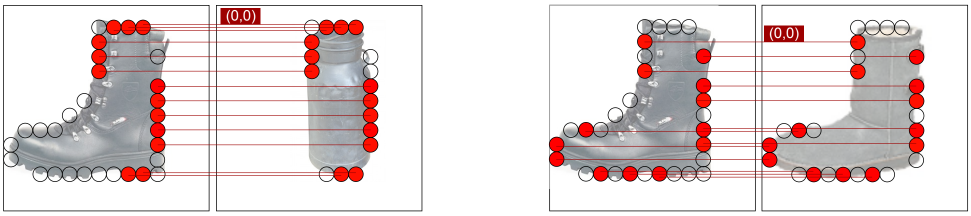
(b) Offset Space

Figure 5.6: *Voting in offset space. Lines connect similar descriptors (a). Red descriptors share same offsets and vote in offset space at corresponding positions. Blue descriptors vote for another offset. Maximum in offset space indicates best matching positions (b).*

implementation is able to compare more than 1400 images in only one second.

Database Handling. The first step is to compute the spatial ensembles as given in [SI07]. This step can be trivially parallelized on the GPU using CUDA. The ensemble of the descriptors of a 256 × 256 image occupy about 600 KB, typically containing 440 descriptors. In order to handle large databases a caching strategy is proposed. Ensembles residing in GPU device memory are swapped out to host memory if the device memory becomes exhausted. As the same strategy is used for host memory, ensembles are finally swapped out to disk which is determined by a simple LRU (least recently used) strategy. The ensembles are stored in a 2D array with one descriptor per column (Fig. 5.5(b)) ensuring fast access to same values in different descriptors in subsequent CUDA threads. As the ensemble a descriptor belongs to as well as its position within the ensemble have to be known in the matching stage, a second array stores additional information.

Matching Ensembles. In the matching stage, the ensemble of a template is compared to all ensembles in the database, yielding a similarity measure. The detailed comparison of ensembles is described in the following. We can either compare one query image to all images stored in the database or compute a weighted similarity graph between each pair of images in a cluster. In each case, multiple scales are supported. As already mentioned, the approach presented by Boiman and Irani [BI07] does not fit well onto the GPU. Thus, we decided to use a simple brute-force voting of first identifying the potential center and then calculating the score of an ensemble. Our approach is similar to a 2D cross-correlation (Fig. 5.5(c)). Every descriptor in the query ensemble is compared to the descriptors of all ensembles in the database. For small distances, the template descriptor casts a vote for a certain central position in a database ensemble. Votes indicate how many template descriptors are similar to database descriptors in the



(a) Comparing I with I'_1 (b) Comparing I with I'_2

Figure 5.7: *Regions in voting. Both times the same number of descriptors cast a vote (red). This would result in the same similarity, although shape of I is more similar to I'_2 than to I'_1. We increase the comparison score by taking the position into account (b).*

same spatial arrangement. Then, the votes are weighted by a number denoting how scattered the voting template descriptors are.

Details of the Matching Strategy. When querying for an image, the template ensemble I is compared to the whole database with ensembles I'_k ($k = 1..K$, with K images in database). Thus, at first the squared distance t between descriptor d_i of I and descriptor d'_{kj} of I'_k ($i, j = 1..80$ bins) is calculated by one CUDA thread per combination of template and database descriptor. If t is below some threshold T, the spatial offset between d_i and d'_{kj} is used to vote for an ensemble offset in offset space $S^s_k(\Delta x, \Delta y)$ (Fig. 5.6). This offset space exists for each ensemble in the database. In order to account for small deformations and variations in scale, each bin in offset space contains 3×3 offsets. As soft-weighting would require many memory accesses, we only increase $S^s_k(\Delta x, \Delta y)$ by 1 if distance t is below T. Because this is rarely the case, the number of memory accesses is very small. Thus, an atomic instruction is used that operates directly on global memory. The offset for which most descriptors voted indicates the displacement where the template image fits best to the database image.

However, the number of votes does not contain any information about the arrangement of the voting descriptors in the template ensemble. If only descriptors in a small region of I cast a vote, then the similarity is smaller than if descriptors were uniformly distributed over I (Fig. 5.7). In order to incorporate the spatial arrangement, we decided to include the position of the voting descriptors into the matching results. Therefore, we partition I into rectangular regions R_m (e.g. 5×5) and assign each descriptor to such a region by its position in the ensemble. In addition to $S^s_k(\Delta x, \Delta y)$, a second offset space $S^r_k(\Delta x, \Delta y)$ stores information about the regions where the voting descriptors are located. $S^r_k(\Delta x, \Delta y)$ is organized as a 2D 32-bit integer array. Bits b_0 to b_{24} are connected to a region. The bit b_v is set if a descriptor in region R_v casts a vote (Alg. 1). Then, offset $\Delta\mathbf{x} = \arg\max_{\Delta x \Delta y} m(\Delta x, \Delta y)$ and similarity $s = \frac{\max(m(\Delta x, \Delta y))}{r \cdot \max(c_I, c_{I'_k})}$ are calculated with $m(\Delta x, \Delta y) = S^s_k(\Delta x, \Delta y) \cdot \mathrm{popcnt}(S^r_k(\Delta x, \Delta y))$. While

Algorithm 1 Voting in offset space (pseudo-code)

```
for all $i, j, k$ in parallel do
    if $(t \leftarrow ||d_i - d'_{kj}||^2) < T$ then
        $\Delta\mathbf{x} \leftarrow \mathbf{x}_{d'_{kj}} - \mathbf{x}_{d_i}$
        $S^s_k(\Delta\mathbf{x}) \leftarrow S^s_k(\Delta\mathbf{x}) + 1$
        $S^r_k(\Delta\mathbf{x}) \leftarrow S^r_k(\Delta\mathbf{x}) \mid (1 << \nu)$
    end if
end for
```

r describes the number of informative descriptors in the template ensemble I, c_I and $c_{I'_k}$ is the number of descriptors in I and I'_k, respectively. The number of set bits in x is counted with popcnt(x).

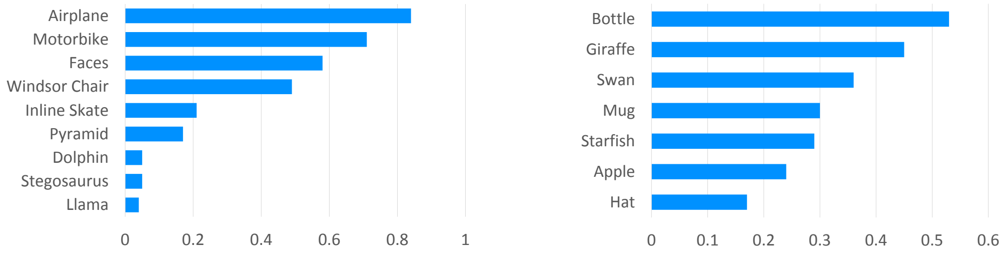
For comparison on various scales, the query is scaled to different sizes before comparing it with the database. Matching all ensembles in the database with each other needs K^2 comparisons. For L scales, the database must contain ensembles in L scales. Consequently, $(LK)^2$ comparisons have to be performed. Due to redundant comparisons, the complexity can be reduced to $(L+1)K^2$.

5.1.4 Evaluation

In order to analyze accuracy, speed, and memory requirements of our approach, we performed experiments on the datasets ETHZ Extended Shape Classes [SS08] (383 images in 7 categories) and Caltech 101 [FFFP04] (9145 images in 101 categories, we removed "BACKGROUND" and "Faces_easy"). For testing on larger image counts, we used MIRFLICKR [MJHL10] containing 1 M Flickr images and additionally downloaded over a million images from Flickr with some random categories.

Visual Results. The self-similarity descriptor already works well for finding similar forms over various domains. To validate our changes in the algorithm, we performed tests with the ETHZ and Caltech datasets. Thus, a database is searched for each image yielding a similarity value between the query and all images. The results are sorted by their similarity and the average precision (AP) is calculated from this list. The mean average precision (mAP) is calculated for each category.

The results for the Caltech (Fig. 5.8(a)) as well as for the ETHZ (Fig. 5.8(b)) dataset vary a lot depending on the different categories. Categories with images sharing a distinct shape work best (Airplanes, Motorbike, Faces). Other categories (Pyramid) contain images with very cluttered background that are not well suited as a template image. So, they have negative impact on the mAP of a category. Our results for ETHZ are similar to [CPZ09]. For measuring cross-domain matching, we applied several



(a) Caltech (b) ETHZ

Figure 5.8: *Mean average precision (mAP) for several categories in datasets.*

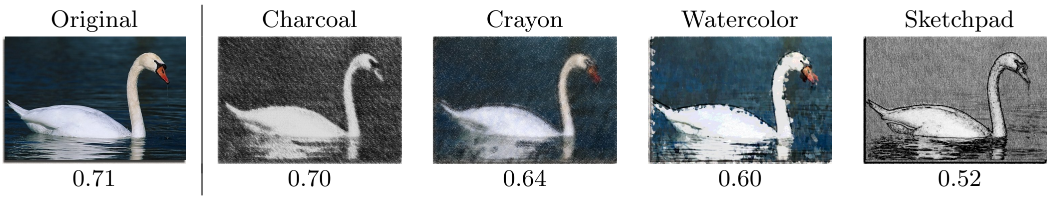


Figure 5.9: *Average precision (APs) for various domains. Effects applied to templates only. AP remains for every effect similar as for the original image.*

effects to the templates (Fig. 5.9) without changing the images in the database. As expected, the AP remains for every effect nearly the same as for the original image.

Performance. We measured performance on a NVIDIA GeForce GTX 580 with 1,5 GB VRAM. The host system uses an Intel Xeon X5660 CPU and 48 GB RAM. First, we only consider the generation of all descriptors of a single image, comparing our GPU version against the OpenCV implementation on CPU (Fig. 5.10). The CPU test was performed on an AMD Phenom II X4 965 CPU with an OpenMP-optimized version. For every image size, our GPU descriptor generation algorithm performs about ten times faster than the CPU implementation.

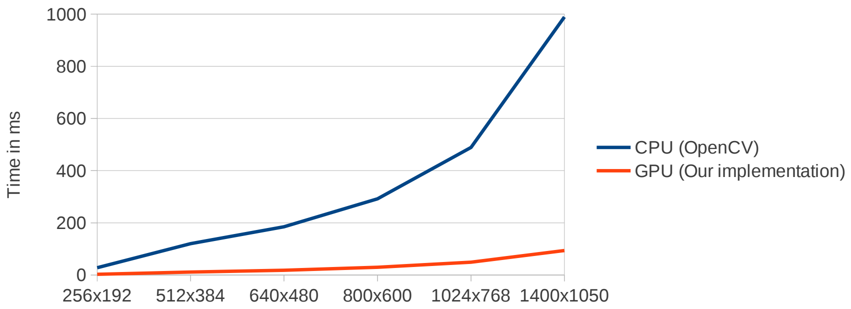


Figure 5.10: *Runtime comparison: OpenCV vs. our GPU version for descriptor generation from images of varying sizes.*

Table 5.1: Performance measurements for data sets of different sizes

	Images	Descriptors		Create DB	Matching one image	
		Total	Informative	Time	Time	Image comparisons per second
ETHZ	383	0.5 M	0.1 M	6s	0.23s	1644
Caltech	8,242	10 M	4 M	1min	3s	2849
ImageNet	31,183	42 M	16 M	8min	11s	2652
MIRFLICKR	999,997	1300 M	450 M	9h 6min	12min	1407
Flickr Images	1,087,007	1400 M	490 M	7h 38min	13min	1406

As already mentioned, we created the database such that our matching is performed very fast and can also be applied to very large image datasets without waiting for hours or even days. Results are shown in Table 5.1. Images were downscaled to 256 × 256 pixels. Times for creating the database as well as for matching vary depending on size of the image set and number of descriptors. For the Caltech set we even retrieved 2849 full image comparisons in only 1 sec. (load DB: 0.89 sec., match: 2 sec.). In order to compensate for variations, we also tested two larger datasets with images of varying sizes and content from Flickr: MIRFLICKR and images we downloaded for random categories. Both times, matching an image still resulted in over 1400 comparisons per sec.

This indicates a great speedup contrasted to [CPZ09], where comparing a single pair of VGA images with quantized descriptors took at least 20 sec. on a 2.4 GHz Pentium. Further, compared to a CUDA-based geometric verification step as employed by Johnson et al. [JGRF10] to organize large-scale photo collections on GPU (Section 4.1.1), our method promises performance benefits while performing not only geometric verification, but also advanced shape matching. In their work, verifying about 4,000 clusters of 30,000 images takes 40 minutes. Our implementation is also faster than various methods based on locality sensitive hashing functions, implemented for CPU [AMP11]. As, on average, only about 30% descriptors in an image are informative, more than half are discarded. The memory required to store the informative ones was 48MB for ETHZ, 1.1GB for Caltech 101 and even 139GB for our Flickr dataset. Thus, memory requirements are larger than with a quantized approach.

5.1.5 Applications

Large-scale image databases normally allow to search by textual queries which often are ambiguous and lead to images with different visual appearance. Due to our efficient implementation, the search results can be improved by taking the shape of the objects shown in images into account and clustering many similar images (around 800 images in about 10 min) or even re-tagging a large database.

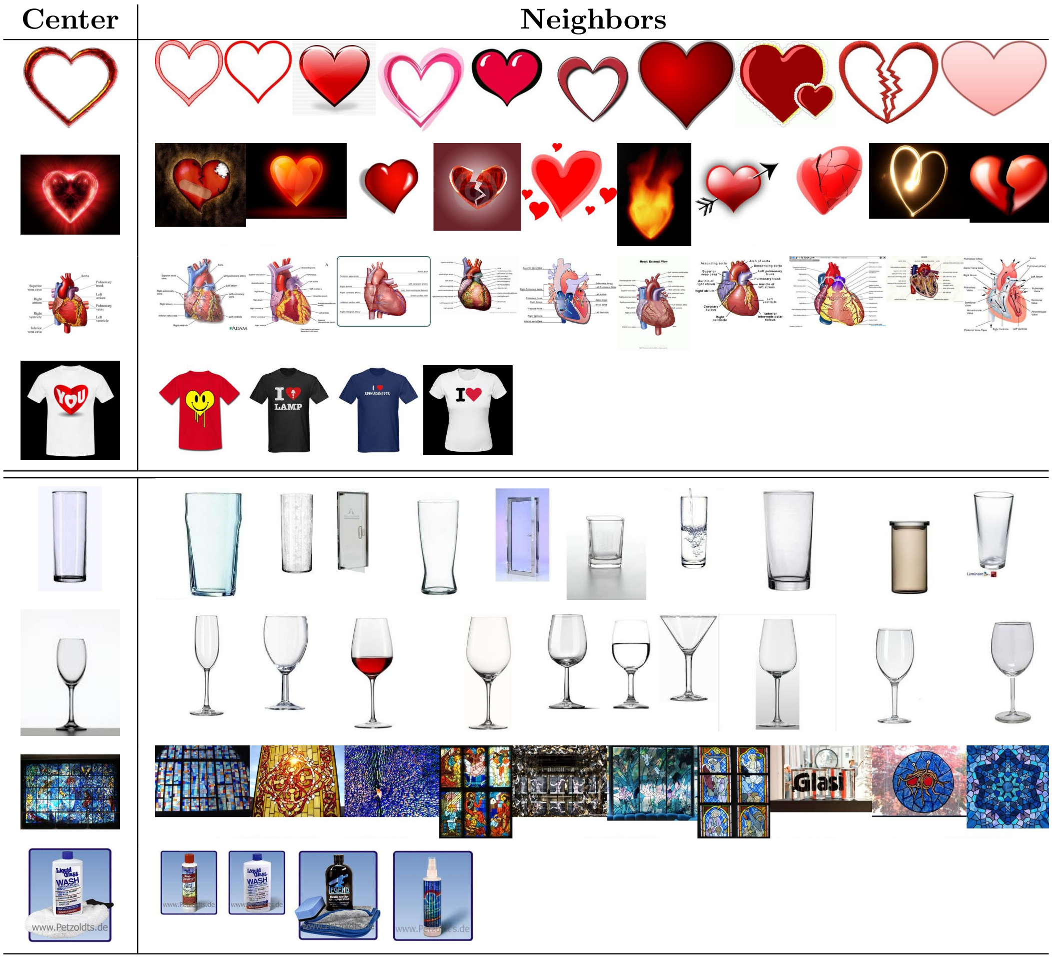


Figure 5.11: *Detected clusters in search results for "heart" and "glass" (Google Images) identify various meanings. Clusters consist of a center (left) and its neighbors in a certain range (right). Different forms show different meanings of ambiguous words.*

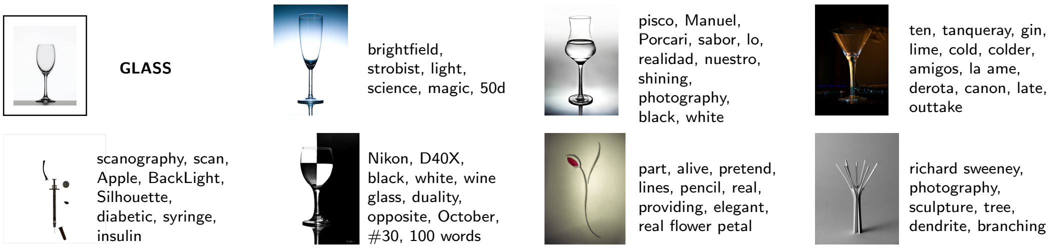

GLASS
brightfield, strobist, light, science, magic, 50d
pisco, Manuel, Porcari, sabor, lo, realidad, nuestro, shining, photography, black, white
ten, tanqueray, gin, lime, cold, colder, amigos, la ame, derota, canon, late, outtake
scanography, scan, Apple, BackLight, Silhouette, diabetic, syringe, insulin
Nikon, D40X, black, white, wine glass, duality, opposite, October, #30, 100 words
part, alive, pretend, lines, pencil, real, providing, elegant, real flower petal
richard sweeney, photography, sculpture, tree, dendrite, branching

Figure 5.12: Images (attached: tag-lists) retrieved in our Flickr set containing similar shape as template (framed) but not according tag. Images not necessarily show same object.

Clustering. Based on the calculated similarity values of our implementation, we generate clusters of similar looking images. The database is created with images retrieved from Google Images by searching for a single word. Then, comparing all images with each other results in a distance graph. The nodes represent images, the edge weight corresponds to the distance ($1 - \mathit{similarity}$). In this graph, cluster centers are found by searching for the node that has the most neighbors with a distance smaller than a threshold. This threshold is based on the average distance of all nodes in the graph. Finally, a cluster consists of the center and all its neighbors in a certain range. By repeating this process multiple times while removing the previously found centers and neighbors, several clusters are extracted. It is amazing what different clusters are found within some categories showing the ambiguity of the words. For example various meanings of "glass": different forms of glasses for drinking, windows or even glass wash liquids (Figure 5.11). These results correlate with our initial observation $\mathcal{H}_{VisSim}$ indicating that clustering visual similarities in retrieval results can identify different meanings of an ambiguous query.

The resulting clusters can be used for further processing, e.g., to clean up unlabeled categories. Besides, the retrieved cluster centers serve as basis for other possible applications like, e.g., re-tagging which is described in the following.

Re-tagging. As our implementation works very efficient, it can be used to quickly find multiple images containing a shape similar to a template image in a large photo database. The test photo collection with about one million images we downloaded from Flickr is enriched with a tag-list we also obtained from Flickr (textual information describing the image and added by Flickr users). Searching for a template image in the photo database may lead to a number of images that contain visual similarity to the template although not containing the according query in their tag-list (Fig. 5.12). In this case, an additional tag is added to the list which leads to a more complete tag-list. If the photo collection is then searched in a textual way, more true positive images are returned.

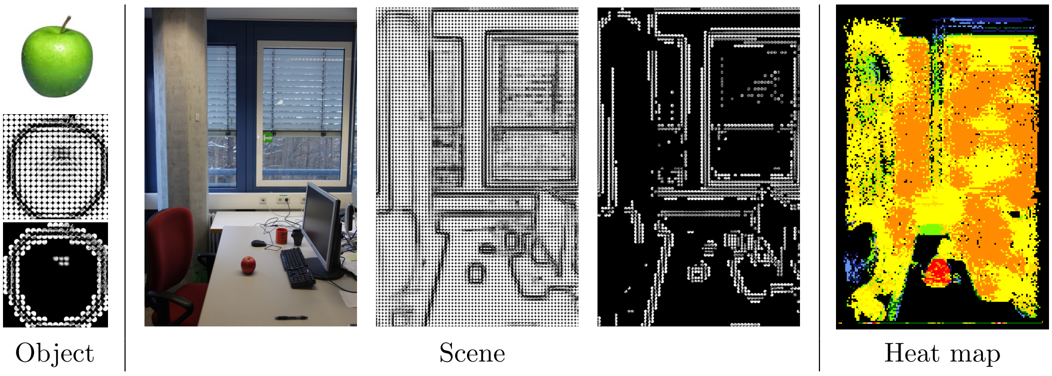


Figure 5.13: Self-similarity descriptors of a specific object (left) can be matched with the descriptors of a scene (middle) to identify the localization of the named object, e.g., in the form of a "heat map" (right).

Text-based Object Localization. Another simple application scenario is to locate an object in an environment by matching the self-similarities of an image from the named query object with the descriptors of a photograph of the surrounding scene.

In detail, based on a simple textual input request such as "Where is the apple?" or "Locate the apple.", the query object, in this case the noun "apple", is identified in a linguistic analysis. This information is then used to query an image set from the web and, as described before, cluster the results to potentially identify the most common meaning of the input word from the biggest cluster. Then, the template images within the cluster can be matched with a photo of the surrounding area in order to infer if the named object exists in the users environment and where it is located. An example is given in Figure 5.13. After the descriptors of a query image (left) and of a photo picturing the environment (middle) have been calculated, the result of matching the descriptors of both pictures can be visualized in a "heat map" (right). Thereby, the resulting similarity values are simply mapped to a color temperature scale that, for example, encodes the highest similarity in red. While the example presents a rather simple input text, more complex text phrases can be analyzed and queried using our hierarchical querying approach (Section 6.2) to retrieve a query image of the search object with high precision.

5.1.6 Conclusion

In this section we presented an efficient approach to find similar images in large datasets based on the local self-similarity descriptor and an ensemble matching strategy that runs completely on GPU using CUDA. We made some effort such that the descriptor database only has to be generated once for all images and, afterwards, enables efficient matching that works directly on it. New images can be directly added. Based on a novel voting-scheme to compare the spatial arrangement of descriptors, our GPU implementation searches the content of thousands of images in only a few seconds without any further pre-processing steps. Thus, images can be searched nearly instantly.

Evaluation is performed with several datasets. Depending on the image size and content, on average about 1800 image comparisons are carried out per second. Applying our implementation to clustering of a given set of images retrieved for a textual query leads to fascinating identifications of various meanings of ambiguous words as stated in $\mathcal{H}_{VisSim}$. Further, extending tag-lists of similar images can improve retrieval based on textual search and our implementation can also support locating an object in a real-world scene based on a textual input query.

Overall, in this section we focused on analyzing the low-level image structure based local self-similarities to efficiently find visual similarities over various domains. In the next section, we will concentrate on the high-level overall appearance of images and introduce an approach to analyze the similarity of aesthetic appeal between images on a high-dimensional level.

5.2 Aesthetic Rankings from Social Online Behavior

In the previous section, we explored similarity between images based on their low-level visual structure and identified resulting meaningful connections. In this section, we focus on the global appearance of images and analyze similarities and differences as relative comparisons between their aesthetic appeal. Therefore, as indicated in Figure 5.14, we exploit user ratings from social online behavior to learn rankings of how aesthetically pleasing an image appears compared to other images.

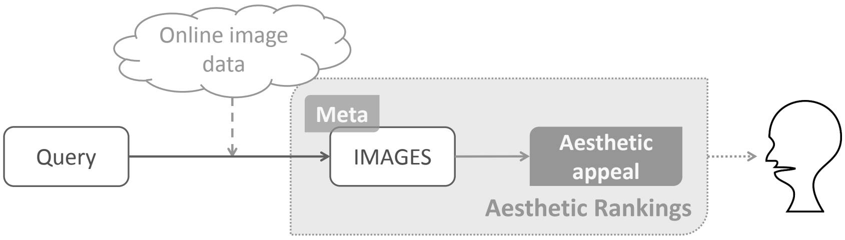


Figure 5.14: *High-level concept. Based on images from online collections, we explore their associated meta information in the form of user ratings to learn predictions of the aesthetic appeal of images from a high diversity of people.*

As mentioned in Section 2.2.4, aesthetics is a highly subjective term and personal aspects largely influence whether one likes a picture. Further, users can individually favor particular content on online platforms with a single click. We observe that:

$\mathcal{H}_{Aesth}$: *People favor aesthetically appealing images over less pleasing ones.*

To tackle the subjective nature of aesthetics, we incorporate a huge diversity of people by utilizing such online distributed multi-user agreements and assemble a large set of 380K images (AROD) with associated meta data. Based on our observation $\mathcal{H}_{Aesth}$, we derive a score to rate how visually pleasing a given photo is. We will evaluate our hypothesis which indicates a strong connection between favored images and aesthetically pleasing ones in a user study (Section 5.2.4) and use our derived model of aesthetics as input for our deep learning network (Section 5.2.5). Whereas previous work has tackled the aesthetic rating problem by ranking on a 1-dimensional rating scale, e.g., incorporating handcrafted attributes, we propose a rather general approach to automatically map aesthetic appeal with all its complexity into an "aesthetic space" allowing for a very fine-grained resolution. Without extra data labeling or handcrafted features, we achieve state-of-the art accuracy on the AVA benchmark set. As our approach is able to predict the aesthetic quality of arbitrary images or videos, we demonstrate our results on several applications (Section 5.2.7). The material of this section is based on the following publication: [SWL18].

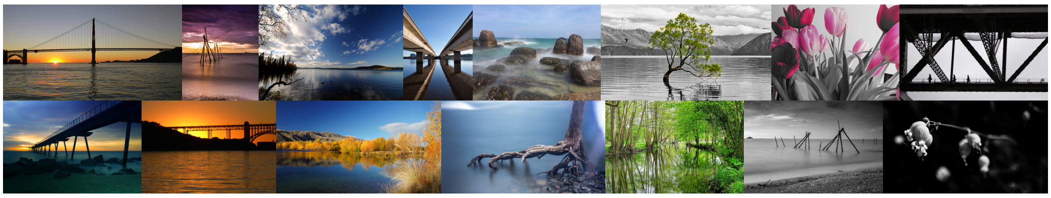


Figure 5.15: Aesthetically pleasing images. The complex matter of aesthetics depends on many factors like visual appearance, composition, content, or style, and makes it almost impossible to directly compare all images if they are similarly beautiful.

5.2.1 Introduction

The widespread use of digital devices allows us to take series of photos so as not to miss any big moment. Manually picking the best shots afterwards is not only time-consuming but also challenging. In general, approaches to automatically rank images towards their aesthetic appeal can be useful in many applications, e.g., to handle personal collections or for retrieval tasks. Overall, deciding how aesthetically pleasing an image appears is a highly complex matter and depends on a large number of factors like visual appearance, image composition, displayed content, or style. Figure 5.15 shows a set of beautiful images with different appearance. Assume one would score each of them separately, e.g., using grades $\{1,2,\ldots,10\}$ to obtain some granularity. This is not only a challenging task but, even more critical, the mapping of these scores to an *absolute* scale can lead to wrong relationships between them. Asking for *relative* comparisons is not only an easier task to accomplish but also results in a more reliable scaling. For images like in Figure 5.15 it is still almost impossible to directly compare all of them, e.g., the beautiful warmth of a sunset can hardly be generally related to the coolness of an image in style "noir". Overall, it is often unclear which particular attribute influences the aesthetic comparison of an image pair the most. Thus, we propose to arrange images in a high-dimensional space to gain a better understanding on a very fine-granular level of how the aesthetic appeal correlates between them without predefining specific factors. On saliency maps, we further demonstrate the necessity of considering global features in aesthetic tasks.

Additionally to the mentioned factors, differences in personal judgments have a large impact on the likeableness of a picture. As mentioned, today's online platforms allow users to easily favor or "like" certain content. Thereby, people usually "like" beautiful images or, in other words, aesthetically pleasing ones. Sometimes, people might also favor images for other reasons like the displayed content, e.g., picturing the newest mobile phone. Anyway, our user study shows that our derived model is still reliable. We consider both, the complexity of aesthetics in its high-dimensionality as well as a huge diversity of multi-user online ratings to obtain broad information about aesthetic relations without extra data labeling.

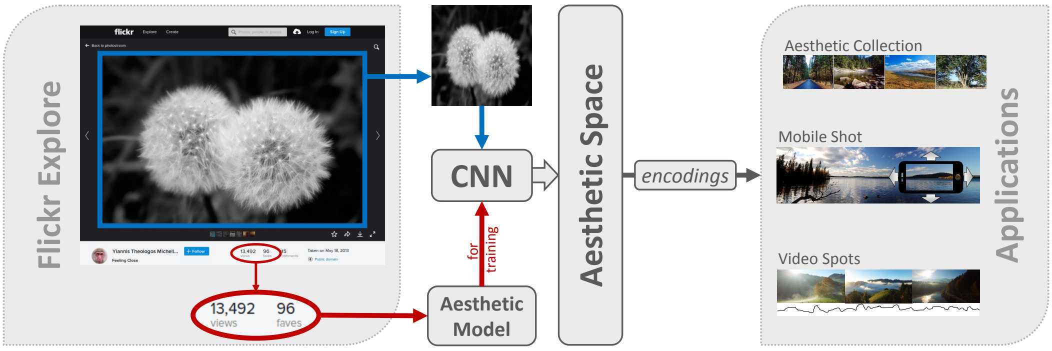


Figure 5.16: *Overview. Based on images assembled from Flickr, we derive a model that scores aesthetic appeal of a picture from its "views" and "faves". This model then guides the training process to learn fine-grained relations in the high-dim "aesthetic space". Our trained CNN can generate encodings for any arbitrary image leading to several applications.*

An overview of our method is illustrated in Figure 5.16. First, we assemble a large amount of images from Flickr and present a new database to exploit *Aesthetic Ratings from Online Data (AROD)*. Therefrom, we derive a model of aesthetics to score the quality of an image by making use of the huge amount of available online user behavior, the "views" and "faves". Then, we make use of deep learning and include the introduced measurements of aesthetic appeal indirectly as hints to guide the training process. Thereby, we only incorporate the information if two images are aesthetically similar or not instead of using the direct score. This allows us to consider every single image relatively to other images – even if they do not seem visually comparable, i.e., due to large differences in their visual factors like appearance, style, or displayed content (Figure 5.17). Our trained CNN is then able to directly learn an encoding of any given image in a high-dimensional feature space resembling visual aesthetics. Our "aesthetic space" encodes the complex matter of aesthetics, namely that not every pair can be compared directly, on a highly fine-grained resolution of relative distances. Finally, as those encodings can be obtained for any arbitrary image, we show how they can easily be transferred into several applications on images as well as videos. In summary, our main contributions are:

- A new large-scale data set containing dense and diverse meta information and statistics to reliably predict visual aesthetics and which is easily extendable.
- A model that approximates aesthetic ratings on a broad diversity without requesting expensive labels beforehand and validation in a user study.
- Formulating the complexity of aesthetic prediction as an encoding problem to directly learn the feature space allowing for fine-granularity of relative rankings on a high-dimensional level.
- Applications for resorting photo collections, capturing the best shot on mobile devices and aesthetic key-frame extraction from videos.

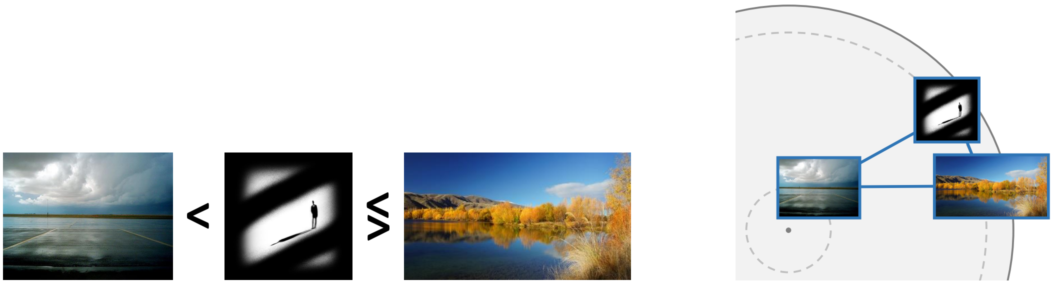


Figure 5.17: Some images might be similarly aesthetically pleasing and hardly comparable, e.g., due to different displayed content (left: middle and right image). Our aesthetic space allows to place them in relation to each other (right).

5.2.2 Related Work on Deep Aesthetics

Aesthetics in images plays an important role in research and has become of even bigger interest since the growing amount of online imagery. Due to the recent success of deep learning, some approaches transferred aesthetics into a deep learning problem. In the following, we review previous work that is closely related to this part of the thesis.

Aesthetics in Images. Previous research on visual aesthetics assessment focused on handcrafted visual cues such as color [OAH11, DJLW06, NOSS11], texture [DJLW06, KTJ06], or content [DOB11, LWT11]. Generally, no absolute rules exist to ensure high aesthetic quality of a photograph. Photo quality has been explored to distinguish between high quality photos and snapshots of low quality [KTJ06] or to classify between the aesthetic quality of a photograph taken from a professional vs a laymen [TLZ$^+$04]. Besides of quality, interest has arisen towards the importance of images. Thereby, previous work has exploited if and to which extent an image can be predicted as "popular" [KDSH14], "memorable" [IXTO11], or "interesting" [DOB11, GGR$^+$13, FHX$^+$14]. Thereby, aesthetics played roles like how it influences the memorability of an image [IXTO11]. Further, style attributes have been incorporated by Lu et al. [LLJ$^+$14] to improve aesthetic categorization. Related work on style recognition will be presented in Section 6.3.2. In addition to style, the composition of an image largely influences aesthetic pleasingness and has been explored in terms of rules or enhancement [Jac, LCWCO10, GLGW12]. Overall, many approaches have investigated a lot of work to find adequate attributes to approach aesthetics, e.g., generic image descriptors [MPLC11], or cues performing psychological experiments [GGR$^+$13]. Datta et al. [DJLW06] extract visual features based on artistic intuition to differentiate between aesthetically pleasing images and displeasing ones. Dhar et al. [DOB11] estimate aesthetic quality on attributes humans might use. They select the three broad types composition, content and sky-illumination to train a classifier. Luo et al. [LWT11] use regional and global features to obtain a content-based assessment whereas Lo et al. [LLC12] propose aesthetic features with high computational efficiency. Other methods have focused on classifying the aesthetic appeal restricting their content to consumer photos with faces [LGLC10, LLC10], consumer videos [MOO10, BNL$^+$13], or other visual domains. For example, Li and Chen [LC09] considered learning computational models of aesthetics on paintings whereas Campbell et al. [CCQ15] focused on evolved abstract images. In contrast to those previous methods, we aim for a general approach to explore the global overall aesthetic appeal without any necessity to restrict image content or define any specific attributes or properties.

Deep Metric Learning. As already mentioned in Section 4.1.3, deep learning networks have emerged as powerful tools to gain a better understanding of visual content, e.g., to detect or classify objects in an image. Overall, neural networks are capable of organizing arbitrary input in a latent space. Approaches directly manipulating this space have been successfully applied to signature verification [BGL$^+$94], face recognition [CHL05, VHW16] and comparing image patches [ZK15] for depth estimation. Hereby, feature representations of the inputs are optimized such that they describe similarity relations within the data. Therefore, metric learning methods such as Siamese networks [CHL05] and Triplet networks [HA15] are widely used. Inspired by those successful networks, we now approach the aesthetic learning problem by directly optimizing a metric to position aesthetic relations in a high-dimensional space.

Deep Learning Aesthetics. Transferring aesthetics into a deep learning approach without defining hand-crafted features has been formulated as a categorization problem based on extracting patches for training [LLJ$^+$14, LLS$^+$15]. However, reducing visual content to small patches can destroy the global appearance which is important for aesthetic tasks. In contrast, we incorporate the entire image and demonstrate the importance of global features on saliency maps (Figure 5.25).

Other methods have considered image quality rating as a traditional classification or regression problem predicting a single scalar information real or binary [MMP12, KSL$^+$16]. Thus, they do not meet the complex nature of aesthetics as they oversimplify the task. They focus on a single scale problem that even humans might not be able to solve as they probably disagree on the actual level of visual pleasingness. Further, these approaches either use hand-crafted features [KSL$^+$16] or examine a data set of small annotation density [MMP12, KSL$^+$16]. In contrast to those methods, we make use of deep metric learning to transfer the problem of aesthetic ranking into a high-dimensional feature space representation. We rely on the plain image without defining any kinds of attributes.

Table 5.2: *Comparison of different data sets containing images for judging visual aesthetics.*

properties	AVA [MMP12]	AADB [KSL+16]	**AROD (ours)**
max ratings (per image)	549	5	**2.8M**
mean ratings (per image)	210	5	**6868**
rating distribution	normal	normal	**uniform**
number of images	250K	10K	**380K**
avg. image size	602×689	773×955	**1926×2344**

Our Work. In summary, many approaches have focused on handcrafted visual cues or on finding adequate attributes to approach aesthetics. Several methods oversimplify the complex nature of aesthetics or examine a data set of small annotation density. However, as judging visual aesthetics is a highly individual process with different biases it is crucial to include the consensus of a large number of distinct individuals using dense information rather than from a small group of people. Thus, we aim for a general approach to explore the global overall aesthetic appeal with all its complexity in a high-dimensional space and without restricting visual content or defining any specific attributes or properties. We exploit the success of deep metric learning to learn a global ranking between the aesthetic appeal of images. In addition, to incorporate a huge diversity of people, we utilize online distributed multi-user agreements and assemble a large data set.

5.2.3 Data Sets

The training of deep networks requires large annotated data sets [RDS+15, LMB+14] to obtain reliable results. Further, as visual aesthetics of photos is highly subjective depending on the current mood as well as any emotion, training a data-driven model requires extensive, diverse annotations. To overcome flaws of previous benchmark sets, we introduce a new data set with a comparison given in Table 5.2.

Previous Data Sets

AVA. The AVA data set [MMP12] provides 250K images classified in visually well-crafted and mediocre ones on a fixed scale. These images are obtained from a professional community of photographic challenges. Through their annotation process only a very small amount of annotations are collected in comparison to the dimensions of social network members comprising also non-professional photographers. Note, to reliably judge image aesthetics it is inevitable to consider the consensus of highly diverse participants.

AADB. Recently, Kong et al. [KSL$^+$16] introduced a new aesthetics and attributes data set (AADB) comprising of 10K images. Each individual image score in AADB represents the averaged rating of five AMT (Amazon Mechanical Turk) workers, who are *asked* to give each image an overall aesthetic score. In addition, they provide attribute assignments from 11 predefined categories as judged by AMT workers. Their database maintains photos downloaded directly from Flickr which are likely to be not post-processed in contrast to professional results contained in AVA [MMP12].

Our Flickr Subset

Whereas AADB is quite small, the image data of AVA seems rather biased. Besides, both only provide a small amount of collected ratings (Table 5.2). Thus, we propose a new, much larger data set comprising aesthetic ratings from online data (AROD). This data can be downloaded immediately, including meta-data as well as extensive, diverse labels, without the need to collect extra ratings spending additional time, effort, and money.

AROD. A single click allows users to give feedback to media content. We propose to use this information. E.g., Flickr allows to add any photo to a personal list of favorites, which is counted as “faves”. Since this feature is optional, users are absolutely free to add a particular image to their favorite list. Their only motivation is to tag a photo which is worth to remember. In addition, these images are uploaded without a purpose to participate in a concrete challenge and are not limited to a specific topic.

To collect these images we crawl around 380K photos from Flickr including meta data such as their number of views, comments, favorite list containing this photo, title of the image and their description from the Flickr website. Our collection contains images which were published and uploaded between January 2004 and November 2016. As each photo is visited ~7K times on average, this allows for a much finer granularity and gives more hints about aesthetics of images compared to previous data sets. Based on this data, we derive a model to obtain information about aesthetic pleasingness of the underlying image.

5.2.4 Model of Aesthetics

In online platforms, people usually tend to “like” beautiful images or, in other words, aesthetically pleasing ones. Thus, we now aim to explore those multi-user agreements and turn them into a new useful measurement towards aesthetic appeal. We extract time-independent statistics, the “faves” and “views” (Fig. 5.16), which contain information traits about the underlying image quality.

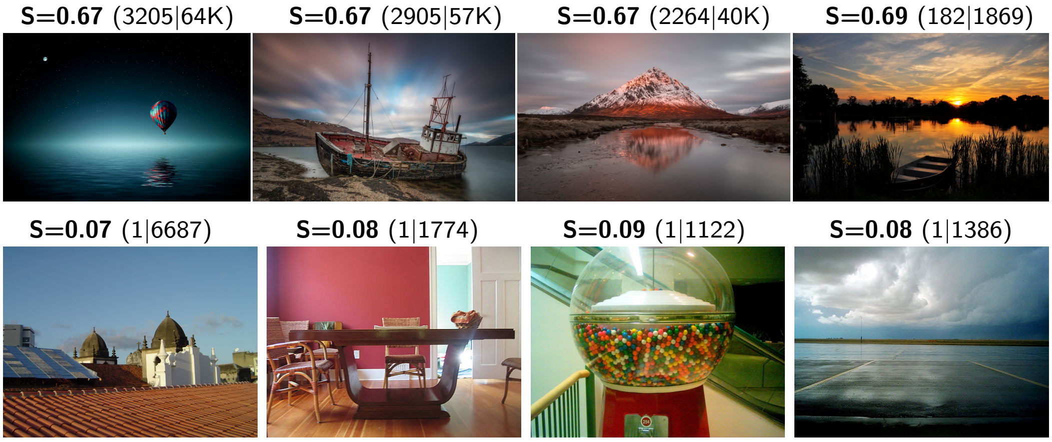


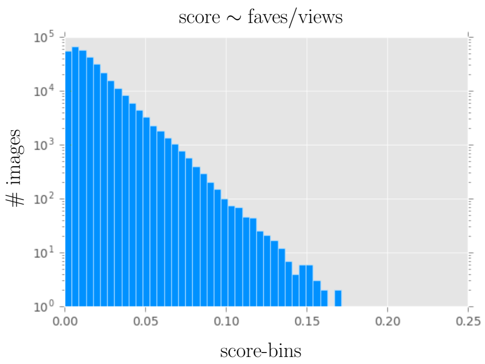
Figure 5.18: *Images from our data set with score S approximated from attached meta information (#faves|#views). The upper row shows images i with large values in $S(i)$ and the bottom row presents examples with relatively low scores $S(i)$. Photos (top row) by Lenny K Photography (left), Dave Holder (two middle ones), Tim Ross (right).*

Model Definition

Previous attempts tried to directly regress some score or trained a simple binary model [NOSS11, DJLW06] to decide whether an image is visually pleasing or ordinary. To overcome the classification approaches Kong et al. [KSL+16] employ a modification of the Siamese loss-function [BGL+94] to re-rank images according their predicted aesthetic score. In contrast to [KSL+16, NOSS11, DJLW06], we will leverage traits from freely available information in social networks to score the image quality. These statistics are only used as hints to guide the training process rather than as a direct label or score.

To judge the pleasingness of an image we examine the relation between the *"views"* (number of visits) and the *"faves"* (number of clicks that favor image) as a proxy for visual aesthetics. Both these landmarks are highly dependent of visual aesthetics and encode the visual quality in all its facets. In addition, the low hurdle of creating a feedback ("like" or "favor") allows to average information being orders of magnitude larger compared to data sets obtained via AMT. This is especially necessary, when treating images which are highly debatable. As common in population dynamics we assume exponential increase of the views $\frac{\mathrm{d}V(i)}{\mathrm{d}t} = r_{V(i)} \cdot V(i)$ and the faves $\frac{\mathrm{d}F(i)}{\mathrm{d}t} = r_{F(i)} \cdot F(i)$ over time $t \in \mathbb{N}$ for any arbitrary image $i \in \mathcal{I}$ with growth rate $r_{(\cdot)} > 0$. This allows us to approximate the score $S(i)$ of the image quality –independent of time t– by

$$S(i) \sim \frac{\log F(i)}{\log V(i)}. \tag{5.1}$$



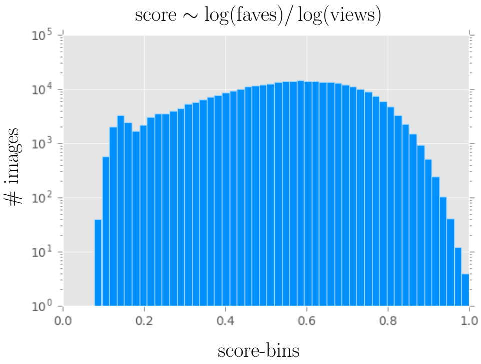


Figure 5.19: *Distribution over our data set. Distribution directly relating the plain faves and views (left) and our approximated model of the collected score $S(i)$ (right). The uniform distribution allows us to even judge images with borderline ratings.*

This time-independence of any image i is necessary when using images with different online life-spans. In addition, the model in Eq. (5.1) accounts for the effect of getting more faves per image being a popular user at Flickr due to the mechanism of followers. Note, the action not to add an image to ones "faves" contains valuable information, too! Considering the score $S(i)$ gives a criteria to rank images $i \in \mathcal{I}$, which values can be imitated by neural networks (see Figure 5.18). A histogram of the distribution of $S(i)$ (Eq. (5.1)) is illustrated in Figure 5.19. The uniform distribution of the data shows that the data has high entropy which allows us to even judge borderline images.

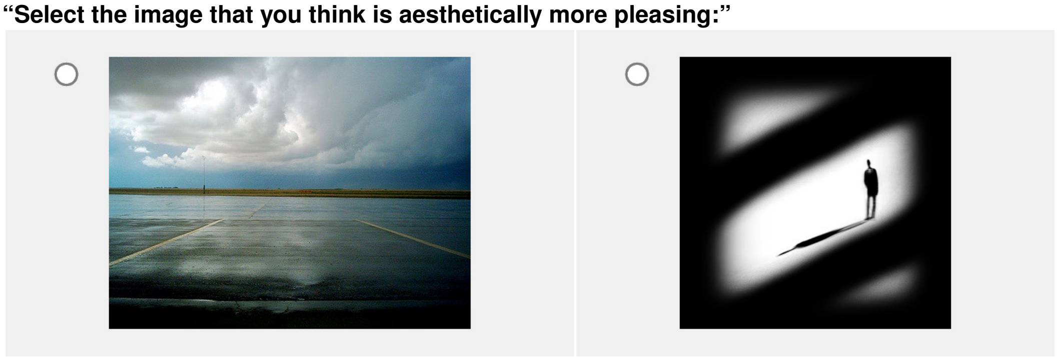
Human Evaluation

As we introduce our aesthetic model as a score based on online behavior from uncontrolled user clicks, we validate the usefulness of our derived metric in a controlled experiment. We formulate our hypotheses as follows:

$\mathcal{H}_1$: Our derived "aesthetic model" based on freely available ratings from an uncontrolled human online behavior is reasonable.

$\mathcal{H}_2$: Higher scored images are also rated better in a controlled user study and worse ones are also rated worse.

Rating the aesthetic quality of an image is highly subjective and differs from person to person. Thus, performing a user study over a diversified crowd is inevitable to validate trends. As mentioned in Section 2.3.1, Buhrmester et al. [BKG11] stated that Amazon Mechanical Turk (AMT) yields reliable data on a demographically diverse level. Thus, we make use of AMT to evaluate our aesthetic model.



"Select the image that you think is aesthetically more pleasing:"

○ ○

Figure 5.20: Example as seen by AMT workers. The task (top) is to select one image of the presented pair (bottom).

Experiment Setup. To overcome differences in internal ratings between persons, we aim for *relative ratings* instead of an absolute scale. Further, to ensure that images obtaining a higher score are really more pleasant than lower scored ones, we design the study as pairwise preference tests. Thereby, the AMT workers are presented two images with different scores. An example is given in Fig. 5.20. In each binary forced-choice task, the Turker is asked to select the image that is "aesthetically more pleasing". We directly ask for aesthetic selection to ensure that our score derived from online "faves" is a suitable measure to rate aesthetics. From our downloaded data set, we evaluate 700 randomly selected image pairs. Each pair is presented to 5 Turkers. To negate click biases ordering as well as positioning are randomized.

User Study Results. In our user study, we randomly test image pairs with varying distances between the scores derived by our model. Thereby, the lowest scored images obtained at least one fave. All evaluated distances are listed in Table 5.3 and according boxplots are displayed in Figure 5.21. A small distance means that our derived scores are very similar and that the images are almost identically pleasing towards aesthetics. However, setting the minimal distance between the scores of the 2 images in a pair to 0.1 is rated towards the similar

Table 5.3: User study results. More similar rating decisions of Turkers are obtained for larger distances $\Delta = \left|S(i) - S(j)\right|$ between our derived scores $S(\cdot)$ of the images within a pair. The p-value and α-level are calculated with the Mann-Whitney U-test [MW47] to evaluate statistical significance.

distance Δ	**> 0.1**	**> 0.2**	**> 0.3**	**> 0.4**	**> 0.5**	**> 0.6**
mean μ	0.78	0.85	0.88	0.89	0.89	0.89
variance σ^2	0.07	0.05	0.04	0.04	0.04	0.04
significance level α	10% ($p < 0.10$)	5% ($p < 0.05$)				

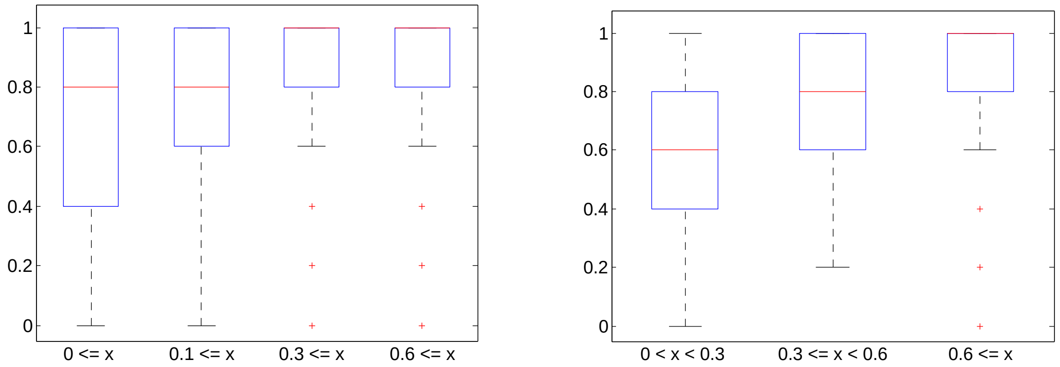


Figure 5.21: *Boxplots of user study results. Rating decisions for distances between image-pairs from several lower bounds (left) as well as grouped into three classes of smaller, middle, and larger distances (right).*

direction by already 78% of the Turkers. Further, for score distances bigger than 0.4, even 89% of the test persons agreed with the selection of the better image. Overall, we obtain ratings with surprisingly small variance. Besides, the already relatively small variance even further decreases with increasing distance. This indicates a high agreement between the different Turkers. The same trend is visualized by the boxplots displayed in Figure 5.21. The left plot directly correlates with the results denoted in Table 5.3 and displays boxes for all distances larger than several lower bounds. Agreements of aesthetic appeal between test persons and our model scores increase with growing distances between the underlying scores. However, outliers indicate the highly subjective task of rating aesthetics. Overall, even when considering only the very small distances ($0 < x < 0.3$) as shown by the left box in the right plot, the agreements of the test persons correlate with the underlying model directions towards the selection of a better image and, again, increase towards bigger distances.

As verified with a Kolmogorov-Smirnov test [Mas51], the underlying data does not come from a normal distribution. Thus, we verified statistical relevance performing the Mann-Whitney U-test [MW47] which rejected the null hypothesis for all distances at least at the 10% level ($p < 0.10$) and for $\Delta > 0.3$ at the 5% level ($p < 0.05$) revealing statistical significant dependency between the scores of our model and the user study ratings ($\mathcal{H}_2$). As we explicitly ask the Turkers to rate due to the term "aesthetically pleasing", our presented score $S(i)$ can really be seen as an aesthetic measure validating our first hypothesis $\mathcal{H}_1$. These findings also confirm our initial observation $\mathcal{H}_{Aesth}$ which states that people favor images that they consider aesthetically pleasing and typically rate them higher or rather distribute a "like" to such images than to less appealing ones.

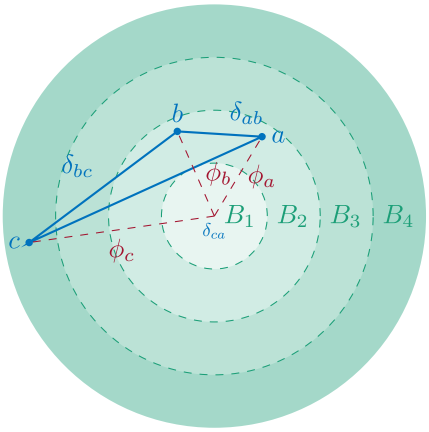


Figure 5.22: *Previous approaches treat aesthetic learning as a low-dimensional problem [KSL+16] which projects encodings onto one dimension or into discrete bins [MMP12] although images might not be directly comparable (Fig 5.17, left). Rather than learning a bin-mapping for each image $i \in \{a, b, c\}$ into bins B_i or directly ϕ_i, we propose to learn pairwise distances δ_{ij} to resolve the highly complex matter of aesthetics in a high-dimensional space considering the images in relation to each other (Fig. 5.17, right).*

5.2.5 Learning Aesthetics

As the visual quality of images is naturally hard to encode in a single scalar and it is hard to match images to discrete bins of aesthetic levels (Fig. 5.17, left), we aim for directly learning an encoding of a given image in a high-dimensional feature space resembling visual aesthetics in contrast to 1-dim ranking as in [KSL+16](Fig. 5.22). We will refer to the feature space as the *aesthetic space*. Ranking approaches like [KSL+16] predict scalars and inherently assume that image orders are possible on a 1-dim discrete or continuous rating scale. Hence, while a latent group of images might be globally misplaced in the aesthetic space, our formulation allows to still order the images within the specific group correctly.

Encoding Aesthetics

Inspired by metric learning [CHL05, HA15], our approach is to directly optimize *relative* distances

$$\delta \colon \mathcal{I} \times \mathcal{I} \to \mathbb{R}, \quad (i, j) \mapsto \|\Phi_i - \Phi_j\|_2$$

between encodings Φ_i, Φ_j from image pairs (i, j). We use a CNN to learn these encodings, which will be described later in detail. Importantly, this training procedure can be done without associating images to any specifically requested ratings or score from human annotators. Instead, it solely uses the information if two images are similarly aesthetic or not on an almost *arbitrary* scale. We minimize the triplet

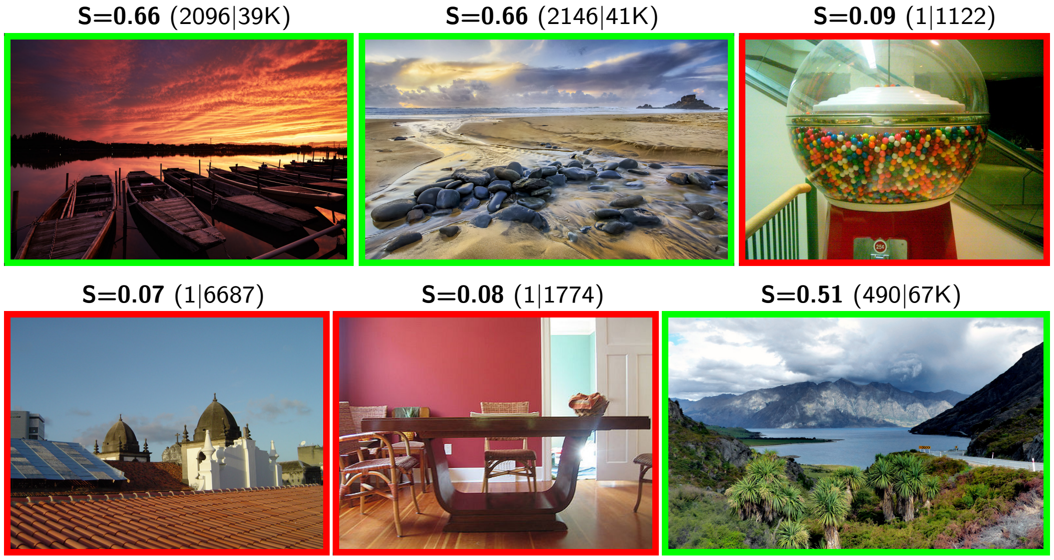


Figure 5.23: *Image triplets for training with scores $S(i)$. Each triplet consists of either 2 good and 1 bad image concerning its approximated quality (top row) or 1 good and 2 bad ones (bottom row). Good and bad images are framed green and red respectively. Photos by Mirai Takahashi (top left), Vitor Pina (top middle).*

loss function [HA15]

$$L_e(a,p,n) = \left[m + \left\| \Phi_a - \Phi_p \right\|_2^2 - \|\Phi_a - \Phi_n\|_2^2 \right]_+ \tag{5.2}$$

for images a,p,n and some margin m. Here, $[x]_+$ denotes the non-negative part of x like the ReLU activation function. This loss resembles a visual comparison, i.e., the distance between two mediocre images a,p should be smaller than the distance to a well-crafted image n and vice versa. Note, our objective function is not directly built on predicting $S(\cdot)$ for a particular image on a *specific* scale and range. To decide whether two images are aesthetically similar or not we use our score $S(i)$ to guide the sampling of the training data consisting of image triplets

$$D = \left\{ (a,p,n) \text{ with } \alpha < \frac{|S(a) - S(p)|}{|S(x) - S(n)|} < \beta,\ x \in \{a,b\} \right\} \tag{5.3}$$

with $\alpha, \beta \in \mathbb{R}$. Thus, any pair (a,p) with a rather small difference in the score allows for adaptively sampling of much harder negatives n by rejecting triplets with too large differences. An example of such image triplets is shown in Fig. 5.23. We allow (a,p) to contain images with higher or lower score than n for generating balance training data. This approach has the following advantages:

1. Every single image can be considered during the training *relatively* to other images, which also allows to train on highly debatable images.

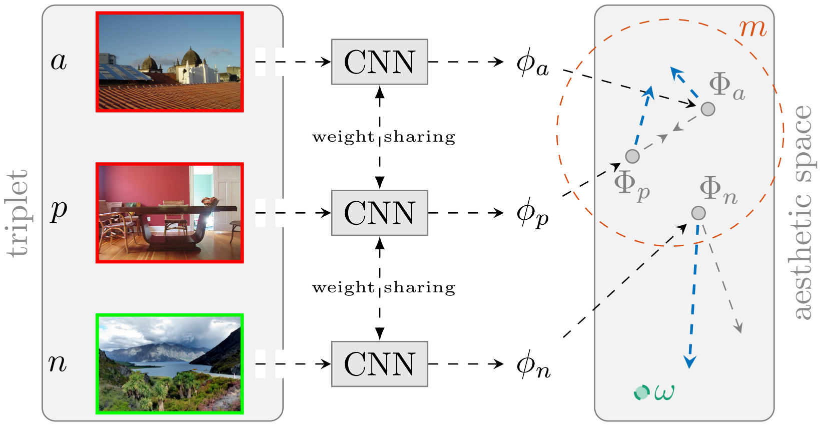


Figure 5.24: *Triplet loss. For each triplet (a,p,n) with anchor point, we aim at encoding aesthetically similar images a,p nearby and force a larger distance to aesthetic dissimilar images n. Adding L_d to L_e alters the update directions wrt. the aesthetic space origin ω.*

2. There is no need to either learn a scalar or solve a binary classification problem in the fashion of ranking [KSL+16] or aesthetic-label prediction [MMP12]. Instead, we learn the encoding itself.

Rating Aesthetics

As the encodings space $\subseteq \mathbb{R}^d$ is only a partially ordered set, for any two images x, y knowing the aesthetic distance $\left\|\Phi_x - \Phi_y\right\|$ has no information if x should be considered as more visually pleasing than y. Thus, ordering multiple images is not possible. If an "universally accepted" worst image ω would exist, then one might simply use the learned distance $\delta(x,\omega)$. But as we are allowed to rotate the entire space, a more practical solution is to force the encoding into a particular direction. We therefore add

$$L_d(a,n) = \operatorname{sign}(s(n) - s(a)) \cdot [||\Phi_a|| - ||\Phi_n|| + \tilde{m}]_+ \tag{5.4}$$

as a directional term to the loss function. This leads the triplet loss by reducing the norms of encodings belonging to less visual pleasing images and increases the norms of well crafted images. Note, that we again do not directly use any absolute score values from our data model. Altogether, we minimize the "directional triplet loss":

$$L(a,p,n) = L_e(a,p,n) + L_d(a,n)$$

to get a natural ordering by the Euclidean norm and relative distances. The effect of L_d is depictured in Fig. 5.24.

Learning the Aesthetic Space

Network Architecture. We use the standard ResNet-50 architecture [HZRS15] f_θ with trainable parameters θ to learn the encodings $\Phi_i = f_\theta(i)$. We add a projection from the *pool5* layer creating a 1000-dimensional descriptor Φ for each frame. Please refer to [HZRS15] for more details. Training was done on two Nvidia Titan X GPUs using stochastic gradient descent with initial learning rate 10^{-3} which is divided by 10 when the error plateaus.

Sampling Training Data. We randomly sample images from our entire collection on-the-fly according to D in Eq. (5.3). We estimate the cardinality of D as $|D| = 7 \cdot 10^{12}$ from tracking the reject-rate during training. Hence, no data-augmentation is required, which would further influence aesthetics. As ResNet expects the input to have the size $224 \times 224 \times 3$, we resize the original image to match the input dimensions. Although, this down-sampling might remove small details, it keeps the relations of the image content. Further, we are interested in the aesthetics quality, rather than the photo quality from a computational photography viewpoint.

From Space to Scale

To allow for multiple applications, e.g., ranking a set of images, it can be necessary to map our derived encodings within our high-dimensional space to a relative scale. As described earlier, while a latent group of images might be globally misplaced in the aesthetic space, our formulation allows to still order the images within the specific group correctly. Thus, we simply consider the norm of the encoding $||\Phi_i||_2$ as the projection score. Thereby, independent of the positions of the encodings in space, the relations between them stay maintained on the scale.

5.2.6 Experimental Results

We pursue two ways of evaluation in quantitative evaluation on the common benchmark set and qualitative evaluation to analyze the internal network mechanism. Further results in combination with applications are presented in Sec. 5.2.7.

Quantitative Evaluation. For a fair comparison to previous approaches, we fine-tune our model network to the distributions of the ratings in the AVA dataset [MMP12]. This is done using a subset of the AVA training data to predict discrete labels instead of relative embeddings. Table 5.4 shows such a quantitative comparison in accuracy to previous methods. Obviously, using an indirect approach such as ranking (Reg–Rank [KSL+16]), which resemble the nature of aesthetic judgments

Table 5.4: *Performance comparison on AVA data set. Different models with according accuracy. Our approach outperforms all models that do not use additional information (right) and even most methods that include additional information during training (left).*

Additional information during training	
method	accuracy
RDCNN [LLJ$^+$14]	74.46 %
Reg–Rank+Att [KSL$^+$16]	75.48 %
Reg–Rank+Att+Cont [KSL$^+$16]	77.33 %

No additional information during training	
method	accuracy
Alexnet–FTune [LLS$^+$15]	59.09 %
Murray [MMP12]	68.00 %
Reg–Rank [KSL$^+$16]	71.50 %
Reg [KSL$^+$16]	72.04 %
SPP [LLJ$^+$14]	72.85 %
DCNN [LLJ$^+$14]	73.25 %
DMA [LLS$^+$15]	74.46 %
Ours	**75.83 %**

much better than standard approaches like classification [LLJ$^+$14, LLS$^+$15, MMP12] yields also better performance on this benchmark set. Ours further boosts this accuracy significantly, which we attribute to the more natural choice of our loss formulation. In contrast to previous work [LLJ$^+$14, KSL$^+$16], we do not rely on a dedicated neural network architecture using a rather common model design. Results on the left use additional information such as attribute data or content-description. Hence, although we trained on a data set which was constructed with literally no extensive explicit labeling, we outperform *all* previous methods relying solely on ratings they obtained in an expensive process. Further, learning from the consensus of many Flickr users is sufficient to gain higher accuracy (our network) on the AVA benchmark set than recent approaches with additional attributes (Reg+Rank+Attr, RDCNN). Note, these attribute categories are acting essentially as a prior and were selected after consulting professional photographers [KSL$^+$16].

We expect to further improve our results when adding more explicit information about the content like in the construction of "Reg+Rank+Att+Cont". As our main focus is to exploit freely available information solely, this explicit meta-information can be image-related comments and tags.

What is the network looking for? Judging the visual quality of an image is totally different from plain object recognition tasks. When extracting relevant information, which is used by the neural network to perform aesthetics prediction, it is possible to visualize prominent traits in the input. To extract these saliency maps, we use guided-ReLU [SDBR15]. It is based on the idea, that large gradients of the output wrt the input have a high impact on the actual network prediction. Fig. 5.25 highlights those pixels in the input with large impact. Hence, this information is strongly coupled with the encoding in our aesthetic space. It clearly shows how our network considers larger regions in the image space compared to sparse saliency along gradients in the untrained network. More precisely, the network model reveals

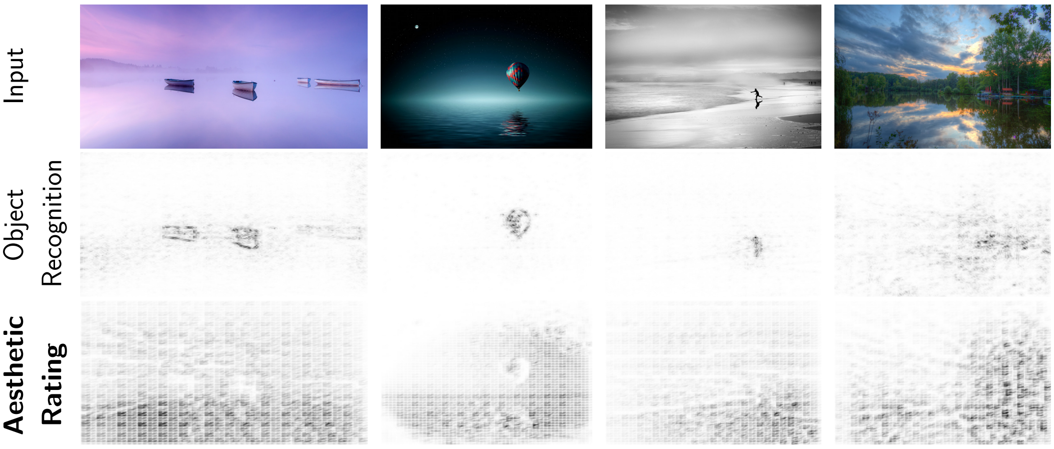


Figure 5.25: Different photos (top), related saliency maps for vanilla ResNet (middle), and our model (bottom) produced by guided-ReLU [SDBR15]. Darker region indicates higher influence on the actual network prediction. Photos (left to right) by john mcsporran, Lenny K Photography, Vitor Pina, Christian Kortum.

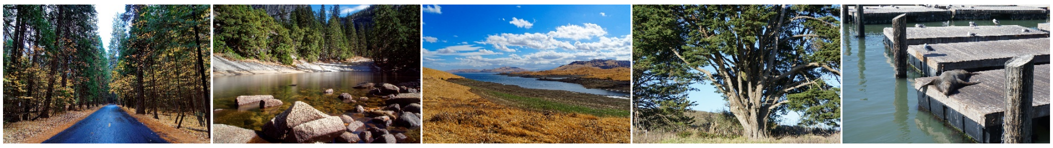


Figure 5.26: Aesthetically resorted set of photos with decreasing score from our provided tool starting with the visually most pleasing image (left).

high synergy effects between surrounding regions of objects whereas the vanilla ResNet (trained on object recognition) rather focuses on the objects (Figure 5.25).

5.2.7 Applications

In order to demonstrate the usability of our approach, we apply our derived aesthetics prediction score to images as well as videos allowing for several applications. Thereby, we map the encodings from space to a relative scale as described in Sec. 5.2.5 maintaining fine-granular relations.

Aesthetic Photo Collection. First of all, we support resorting an arbitrary photo collection due to our predicted relative aesthetic scores between the images. An example of a small set of aesthetically sorted images is shown in Fig. 5.26. This tool can facilitate to quickly resort one's holiday collection and directly share the best moments without time-consuming manual browsing of the usually rather large set of pictures.

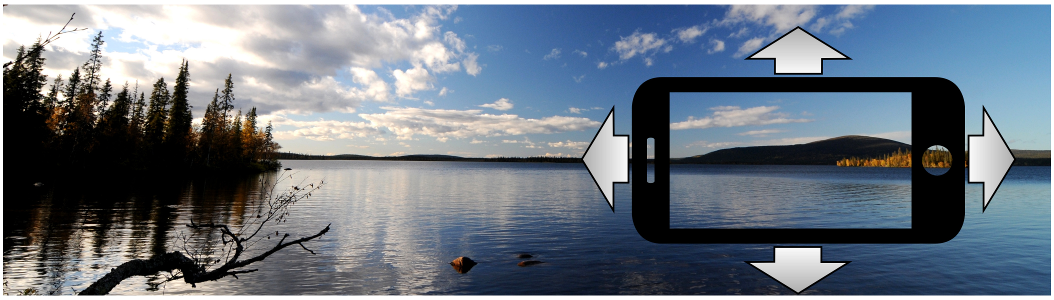


Figure 5.27: Best mobile shot. Based on slight movements in any direction, the application automatically captures the best shot.

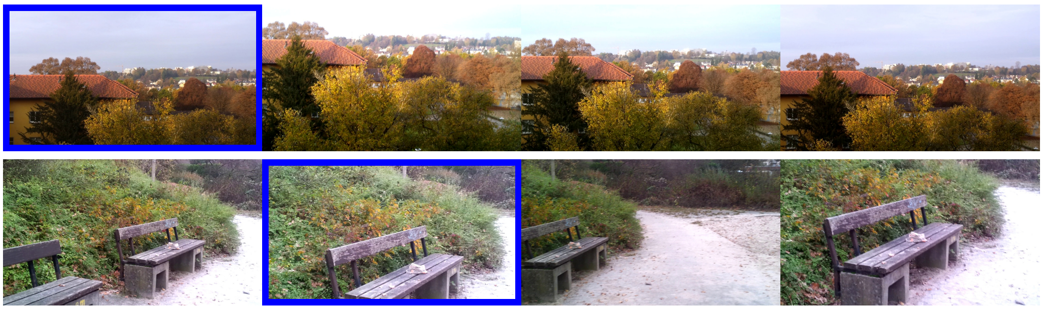


Figure 5.28: Best predicted image (blue frame) during capturing. The movements were recorded with a mobile device.

Best Mobile Shot. A commonly known situation is that people want to take a picture but are not completely sure what the best shot of the view could be. They tend to take mulitple pictures and just postpone the decision process. This can even lead to missing the one best shot completely. We provide a simple application that allows slightly moving the phone around and temporarily captures a video. The idea is illustrated in Fig. 5.27. All the single images are then analyzed and rated by our system and the image of the best view is saved. The application supports the user to directly obtain the best aesthetically pleasing image and prevents the time-consuming decision process afterwards. Fig. 5.28 shows several frames from movements we recorded with a Samsung Galaxy SII phone and the predicted best shots. Sky proportions, saturation and the tension of the overall image layout play an important role within the decision. Due to its small memory footprint of only 102MB containing the network weights, running this application directly on mobile devices is easily possible. This application could further be extended to lead the user to the best shot during the movement while indicating better directions.

Video Spots. Similarly, our system is able to find great shots in a video. Those shots can be selected as aesthetic key frames or, e.g., in documentary films, to identify

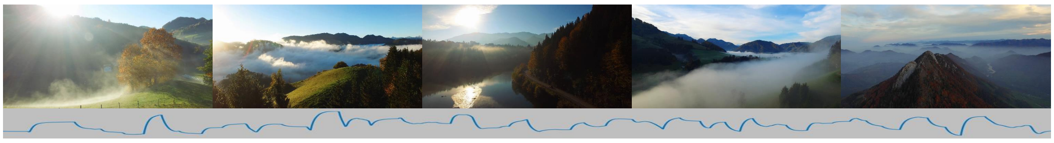


Figure 5.29: *Best video spots. Each frame is extracted at the peaks in the score signal.*

the most wonderful places or spots. Therefore, we calculate a complete prediction curve along the video displaying the aesthetic relation between the frames. Fig. 5.29 displays an example of a video and the according aesthetic prediction curve. Kalman filtering is applied to smooth the final predictions over time. Extracting the frame scores is done at a speed of 140fps on a NVidia GTX960. Embedding common videos requires only 25% of the actual playback time demonstrating high efficiency and enabling real-time applications.

5.2.8 Conclusion

In this section we introduced a new data-driven approach which learns to map aesthetics with all its complexity into a high-dimensional feature space. Additionally, we make use of online behavior to incorporate a broad diversity of user reactions as rating aesthetics is a highly subjective task.

In detail, we assemble a novel large-scale data set of images from social media content. Hereby, aesthetics ground-truth scores for training are obtained *without* explicitly requesting user ratings in a time-consuming and costly process. Hence, our dataset can be easily extended, as our approach requires effectively no labeling-efforts using freely available information from social media content. The assumption of our underlying model is validated in a user study which also confirms our initial observation $\mathcal{H}_{Aesth}$ namely that people typically prefer images that have an aesthetic appeal. To automatically judge aesthetics, we formulate the aesthetic prediction directly as an encoding problem. Consequently, we propose a more natural loss objective for dealing with the complex task of learning a feature representation of visual aesthetics. Our focus lies on the abstract representation of aesthetics using online media. Thus, we solely rely on a commonly used model architecture and use a much weaker training signal which leads to state-of-the-art results on previous benchmarks. Finally, we confirm the success of our model in several real-world applications, namely, resorting photo collections, capturing the best shot and a smooth aesthetics prediction along a video stream.

Overall, this section focused on exploring the aesthetic appeal of pictures. Such an appealing look of an image largely influences if the viewer likes what he sees. However, the emotional feeling evoked in the observer can still differ, i.e., an aesthetically pleasing image can provoke a large range of individual feelings. Thus, in the next section, we will investigate in the emotional effect of pictures.

5.3 Changing Emotional Response to Images

Additionally to the aesthetic appearance of images that we analyzed in the previous section, pictures are often touching whereby they trigger a specific emotional response or even change the mood of the observer. Whereas the last section focused on analyzing the personal judgments of a huge diversity of people, this section investigates in gaining deeper insights into the direct connection between the appearance of a picture and the feelings of a person. Therefore, as indicated in Figure 5.30, we retrieve images from online collections by querying for emotion labels occurring in their attached textual meta data.

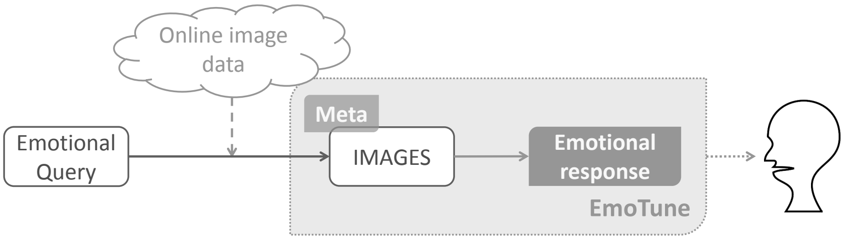


Figure 5.30: *High-level concept. Images are retrieved from online collections by querying for emotion labels. The emotional effect of the images towards an observer is directly explored.*

We directly study the emotional influence of pictures on an observer and, therefore, investigate in simple global modifications of the visual appearance, i.e., modifying the brightness, the saturation, or the color temperature. We hypothesize that:

> $\mathcal{H}_{EmoTune}$: *Global image modifications can change the emotional perception.*

To evaluate our statement $\mathcal{H}_{EmoTune}$, we collect empirical data on the retrieved images associated with emotion labels in a user study (Section 5.3.6). In particular, we analyze the valence, a simple positive–negative rating, evoked by the stimulus with respect to the different modifications (Section 5.3.4) and their strengths. We show that these relationships tend to be linear only in a limited range while sometimes stronger modifications lead to the opposite effect. Pushing the modifiers towards the boundaries we derive from those ranges and combining them successfully shifted the emotional response on 92% of around 80 samples. From these findings we derive our "EmoTune" filter which allows for almost linear control by combining specific modes (Section 5.3.7) and demonstrate successful application to both images and videos (Section 5.3.9). The material presented in this section is based on the following publication: [SFFL17].

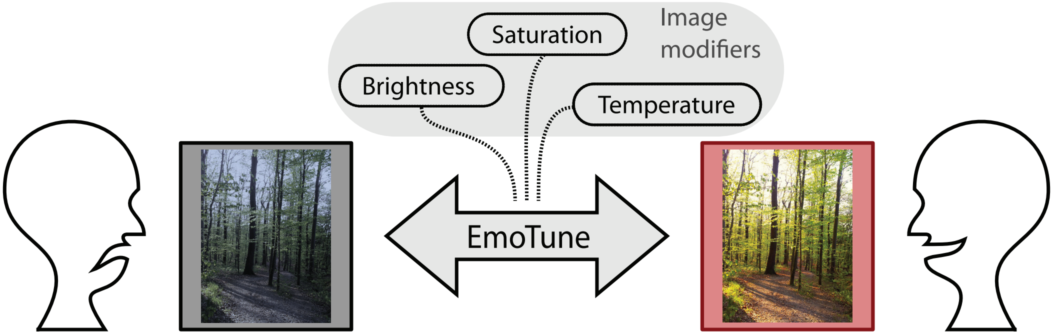


Figure 5.31: EmoTune. Brightness, saturation and temperature are altered to predictably change the evoked emotion of an observer.

5.3.1 Introduction

When viewing an image or watching a movie, a complex emotional response is evoked in the observer depending on multiple factors such as the displayed content, the observer's memories, or the appearance of the image. For sure, the evoked emotions are highly subjective. Thus, if artists provoke a certain mood they exaggerate their stylistic devices to carry out an impression. To complete the affective atmosphere of a film, the movie industry performs color grading not only to correct exposure errors or balance color but also to create a certain style influencing the audience's perception. Professional colorists make shots appear warmer or cooler and are able to set a different mood of a scene [vH10].

Generally, an appealing look of an image influences if the viewer likes it, but the evoked emotional feeling can still differ. In this work, we explore the emotional effect, more precisely, we aim at *tuning* an image to predictably alter the emotional perception ("EmoTune" – Figure 5.31). While the image content plays a major role in the viewer's emotional response, the semantics behind objects and their arrangements are particularly difficult to assess. Thus, we concentrate on content independent visual characteristics.

Overall, our work investigates how simple changes in basic global image modifications have a large impact on the emotional perception. We perform four types of experiments on images from two different sets where each image has been previously labeled with the associated emotion. One of them is a new broad image set we assembled by hierarchically querying images from Flickr for specific emotional categories [MFL+05]. It allows to manipulate arbitrary images independent of their content. Further, in our empirical study, we investigate how the manipulation of basic image characteristics, namely hue, brightness, saturation, color temperature and local contrast affects the emotion an image evokes. The emotional state is expressed by subjective self-assessment of the so–called *valence* [BL94], rating positive or negative emotions on a five point Likert scale. Our observations show

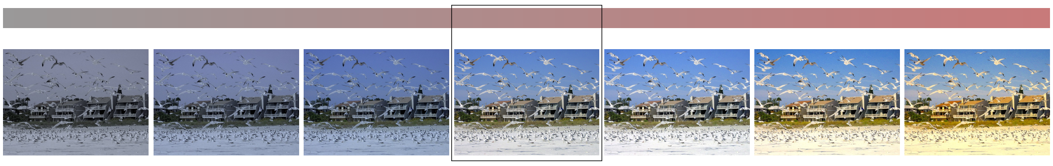


Figure 5.32: *Tuning an arbitrary input image (center) to evoke a rather negative (left) or positive (right) emotional response. Based on a single input parameter the EmoTune filter alters the brightness, saturation and color temperature to increase the valence of an observer. Original photo (center) by AntyDiluvian.*

that depending on the mode and the strength of the modification one can boost, attenuate or even flip the emotional response in a rather controlled way. The study further revealed limits within which the modifiers do operate as expected. Beyond those particular thresholds they quickly negate the intended change. As the derived modifications are additive to each other, they can be applied jointly to achieve an even stronger tuning. Based on these insights we develop a novel heuristic filter that changes a given image such that the triggered emotion can be tuned relatively to the input using a single parameter (Figure 5.32). We compare our tuned results to color gradings of an expert and establish the same trend. Thus, our developed *EmoTune* slider can support laymen as well as experts to influence the evoked emotion of films, games, advertisements, or website content to their liking.

5.3.2 Related Work on Color and Emotion

There exist numerous publications on which emotions are conveyed by an image and how, both on the psychological and physiological level. Most previous work deals with categorizing and understanding emotion in existing visual material. Besides, if setting the mood of an image, color plays an important role as already slight changes have a strong impact on the visual appeal. However, whereas image appearance and color manipulation with the help of a reference image became a popular research topic, only little research was done regarding the manipulation of images to affect the emotional response except in the field of non-photorealistic rendering (NPR).

Emotion Categorization. In order to distinguish between different emotions, two main directions have been established, namely, categorical approaches [Dar72, Ekm99] and dimensional ones [Wun96, OST57, Rus80, RWM89]. Categorical approaches assign discrete labels to specific emotions, e.g., Darwin [Dar72] has assumed that emotions are modular using terms like *anger* or *fear*. Ekman [Ekm99] has defined characteristics to distinguish such basic discrete emotions. In contrast, dimensional approaches map emotions into a space. For example, Wundt [Wun96] has introduced dimensions of pleasant–unpleasant and low–high intensity to differ between emotions whereas Russell et al. [RWM89] has used the two dimensions pleasure–displeasure and arousal–sleepiness for assessing the emotional response in an "Affect Grid". Due to the high rate of growth in this field, Ekman [Ekm16] has lately explored general agreements of scientists about emotion in a study revealing that both directions are similarly fundamental. Other outcomes coincide with specific emotion labels, or the agreement on an existing relation between specific moods and emotions. In order to provide a standardized collection of visual emotional stimuli Lang et al. [LBC08] have used a dimensional method to develop the IAPS (International Affective Picture System) dataset. The IAPS is a collection of images and their respective ratings on the three-mode valence-arousal-dominance scale. Additionally, Mikels et al. [MFL$^+$05] have defined discrete emotional categories on parts of this standardized image set. We make use of both, the IAPS as well as the categories to explore emotional effect of visual data.

Color Manipulation. Manipulating image color to change the visual appearance has been mainly addressed in the field of example-based color transfer, e.g., to map the color concept or look from a reference image to another individual photo [RAGS01, RP11, BPD06], a video sequence [CSN05], or from a reference video to another video clip [BSPP13]. Initially, Reinhard et al. [RAGS01] have proposed a simple algorithm based on statistical analysis to map color characteristics between images with the focus on finding a suitable color space in which simple operations can be applied. The work of Reinhard and Pouli [RP11] has further explored different color spaces in the context of color transfer. Several approaches exploit the idea of matching color statistics between images to perform color transfer. Chang et al. [CUS04] incorporate characteristics of human color perception and restrict the set of basic color categories based on psycho-physiological experiments, whereas Pitié et al. [PKD05] have proposed a non-parametric method to match arbitrary distributions. An enhancement technique based on a two-scale non-linear decomposition that allows the user to transfer the look as a combination of tonal balance and detail from one photo to another has been introduced by Bae et al. [BPD06]. Wang et al. [WYX11] approach image enhancement by learning rules for color and tone adjustments from example image pairs, more precisely, pairs with one picture before adjustments were made and a corresponding one from afterwards. An approach to unify color and texture transfer is given by Okura et al. [OVB$^+$15] to map a change in appearance and, at the same time, incorporate the generation of new image content without destroying structure which can arise from texture transfer. Further, an efficient image enhancement method mapping non-rigid dense correspondences (NRDC) between two images with similar content has been developed by HaCohen et al. [HSGL11]. This method is especially designed to work on image pairs taken under different acquisition conditions. Their work has been expanded to photo collections optimizing the color consistency along multiple images by propagating color from one photo to a set of compatible images [HSGL13].

To be applicable to larger image collections, Park et al. [PTSSK16] increase efficiency recovering sparse pixel correspondences. Transferring the style from one image to an entire video sequence has been approached by Chang et al. [CSN05]. Starting from video data, Bonneel et al. [BSPP13] have introduced a method to transfer the color concept of a reference video clip to another video sequence. To avoid flickering, they have proposed a differential-geometry-based scheme to interpolate between per-frame color transformations.

Further, multiple images have been considered to alter a single image based on attributes like season [LRT+14] or time [SPDF13]. Laffont et al. [LRT+14] study high-level scene attributes, e.g., "snow" or "autumn" and have proposed an appearance transfer method to adjust an outdoor scene based on 40 attributes. Altering an image to show a different day time has been approached by Shih et al. [SPDF13]. To identify the modification of the color appearance they make use of time-lapse videos locating a frame with similar time as the input and another frame with the requested output time.

In contrast to those example-based color transfer methods, we modify the color appearance of an image without any additional reference picture. Besides, as color has an enormous impact on the visual appearance of images, we will incorporate a color feature for color-based image selection of neighboring pictures in our text illustration methods (Sections 6.2, 6.3).

Color and Emotion. The effect of color on emotions has been reported as relationship between perceived valence and hue (wavelength) revealing that blue is the most pleasant color, yellow the least pleasant, and less bright as well as more saturated colors are more arousing [VM94, Sim06]. Further, emotional response to images has successfully been used to categorize images by machine learning approaches exploiting the IAPS [YvGR+08, LZT10, MH10]. Yanulevskaya et al. [YvGR+08] and Li et al. [LZT10] train SVMs based on a number of holistic image features such as GIST or Gabor features, whereas Machadjdik and Hanbury [MH10] focus on concepts they derive from psychology as well as art theory.

However, manipulating images in the context of emotional effect has mainly been explored in the field of NPR (non-photorealistic rendering) to study the evoked affect of renderings [DBHM03], interactively reflect the user's emotional state derived from the user's facial expression in painterly renderings [SBC06], or dampen the viewer's emotional response [MML11, ZZ10]. Mandryk et al. [MML11] have explored different NPR algorithms whereby a considerable shift towards the neutral range has been achieved using the painterly style of Zhao and Zhu [ZZ10] which unfortunately removes most of the image detail. Contrarily, we focus on provoking a stronger emotional response rather than dampening it.

Other approaches enhance image harmony by shifting specific colors [COSG+06] or investigate in color themes [OAH11, CFL+15, WJC12]. Cohen-Or et al. [COSG+06] have exploited harmonic color schemes to establish a harmony between foreground and background of an image. O'Donovan et al. [OAH11] have studied color themes, more precisely, sets of five colors, to derive preferred ones using online-image datasets and a machine learning approach. Chang et al. [CFL$^+$15] have proposed a simple interactive tool to recolor an image by only editing such a color palette. Further, in contrast to previous work that aims to create completely new imagery from text (Section 4.2.2), approaches exist that make use of single affective words to manipulate the image appearance based on such color palettes [WJC12] or precisely adapt specific parts of an image [LSBS12, LS15]. Wang et al. [WJC12] have exploited 5-color themes to adjust the appearance of a picture based on an affective input word. Learning semantic concepts between words and the image structure has been successfully explored by Lindner et al. [LSBS12, LS15] to enhance images and provide re-renderings with spatially varying color enhancements depending on concepts like "grass" or "autumn". Additionally, Lindner and Süsstrunk [LS15] have presented work on identifying color triplets for given color names as well as finding corresponding palettes of five harmonic colors to semantic expressions.

While research concentrates on affective understanding of images, movies are also of research interest, i.e., Wang and Cheong [WC06] present a systematic approach combining psychology and cinematography to address affective understanding.

We now aim to modify the visual appearance of an image independent of its content or representation and without deploying prior learning steps. Following previous insights, we will explore the response to color and intensity in our experiments.

5.3.3 Data Sets

In order to study emotional influence on a broad and diverse set of images, we will explore various modifications (Section 5.3.4) for their caused emotional response on two different image data bases. In particular, we make use of the IAPS but also assemble a new set from Flickr which is more independent of content and resolution. Both are described in the following. In addition, we will apply our resulting heuristic filter not only to images but also to a selection of video clips (Section 5.3.9).

IAPS. The mentioned IAPS [LBC08] data set consists of images evoking a particular emotion with measured certainty. Additional labels assign emotional categories defined by Mikels et al. [MFL$^+$05]: Negative norms (anger, disgust, fear, sadness), positive norms (amusement, awe, contentment, excitement) and undifferentiated ones. Unfortunately, the IAPS mainly consists of images selected based on strong content (blood, insects, faces, ...) rather than displaying some mood or a certain atmosphere. Besides, most of the images in the IAPS are of rather low image quality, e.g., noisy, blurred, or low resolution.

Flickr. Inspired by the IAPS, but aiming at a diversity of images in adequate quality which are independent of content and resolution, we download images from Flickr based on the same emotional categories [MFL+05]: We define the set of negative norms as $\kappa_{neg} = \{anger, disgust, fear, sadness\}$ and the positive ones as $\kappa_{pos} = \{amusement, awe, contentment, excitement\}$. To query for precisely matching images, we combine all negative norms into the powerset $\mathcal{P}(\kappa_{neg})$ and all positive terms into the powerset $\mathcal{P}(\kappa_{pos})$. As we will show later in Section 6.2.2, very specific textual queries result in higher semantic precision with regard to the retrieved images but might return only a small amount or even no image result. Thus, to assemble a large enough set with images of high precision we apply the hierarchical querying approach described in Section 6.2.4 to both powersets separately. In detail, for each subset $P = \{p_1, ..., p_N\} \in \mathcal{P}(\kappa)$ we combine the elements $p_n \in P$ by conjunction to query for specific images. The next hierarchy level is a disjunction of conjunctions of all subsets of the $N-1$ elements and so on. The most general level queries for the disjunction of all N elements p_n. This querying method leaves us with different levels of specialized images and enables further selection within a broad set.

Overall, we obtain around 25K images for the negative set and 17K images for the positive one. For further processing, we manually select a small controllable subset of about 450 images, 214 negative and 237 positive ones, ensuring their mapping into the correct positive – negative class.

Videos. In order to apply our derived image filter to videos, we also crawl single short videos that are presented under creative commons license (CC0) from the video sharing site vimeo [1].

5.3.4 Image Modifiers

When image content is presented to humans, they can react with a broad range of emotions from very negative to very positive ones. However, our main objective is to influence the viewer's emotional response regardless of the image content. As indicated in previous work, color is a very powerful device to change the appearance of an image without performing a semantic analysis. Thus, we aim to identify a set of modifiers that is capable of changing the observer's evoked emotional response relatively to a given input image I towards more negative I^- or more positive I^+. Therefore, we evaluate simple but effective image modifications operating in HSV color space on hue, saturation, brightness, as well as on the color temperature and the local contrast (Figure 5.33).

Hue (H). The dimension hue in HSV color space defines how similar colors appear. It has been explored to have noticeable impact [VM94]. Our modifier I_H regulates

[1] `https://vimeo.com`

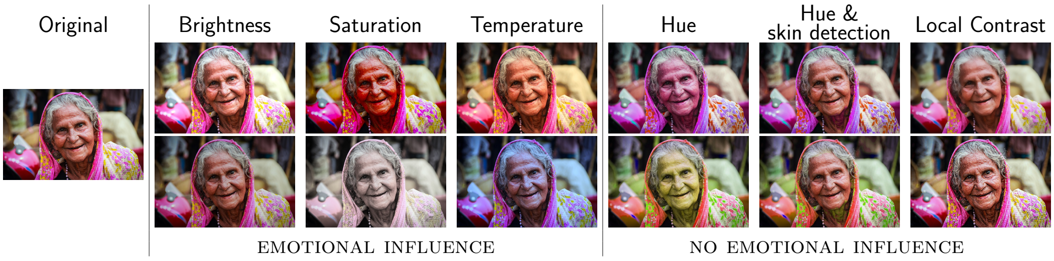


Figure 5.33: *All evaluated modifiers. First column original image, column 2-7 modified versions. Top row positive effect expected, second row negative effect expected. For the three rightmost columns no clear change in the emotional response was observed. Original photo by Anwar Shamim.*

this channel. Further, observers react highly sensitive if humans and, especially, faces are shown. As shifting the hue might result in a rather unnatural look, we apply face detection to handle them separately.

Saturation (S). The saturation in HSV determines the colorfulness C of a color relative to its own brightness B.

$$S = \begin{cases} 0 & B = 0 \\ C/B & otherwise \end{cases}$$

This means, that it is inherently limited by the minimum and maximum possible brightness. Saturation already played an important role in research on color and emotion [VM94, Sim06], and, thus, we evaluate an adequate modifier I_S.

Brightness (B). The third dimension in HSV represents the brightness. When viewed in the usual RGB color space, changing the brightness results in a change of the mean RGB value of the image. In digital images, this change is limited by the dynamic range of the imaging pipeline. As light, or the intuitive imagination of day against night largely influences triggered feelings, we apply a brightness modifier I_B on the luminance channel of an image.

Color Temperature (T). Color temperature denotes the color of light emitted by a black body heated to temperature T and is expressed in kelvin K. As certain color categories are closely related to the impression of an image giving it a rather cool or warm tone, we also incorporate a modifier I_T. We choose $6500K$ (color temperature of daylight) as the basis of I_T.

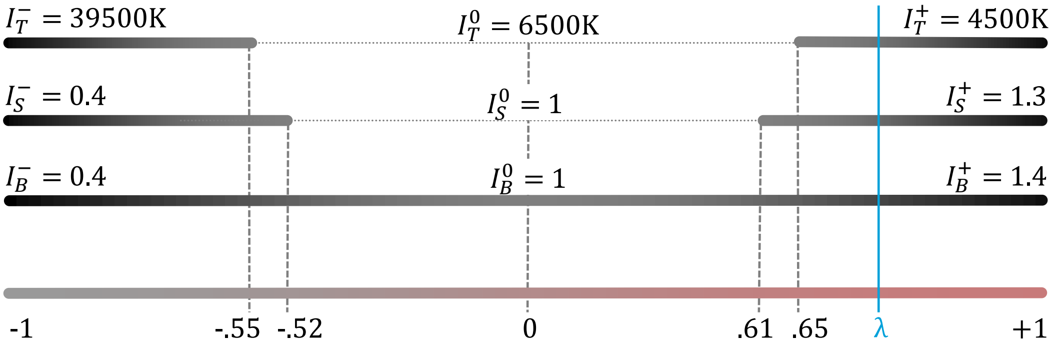


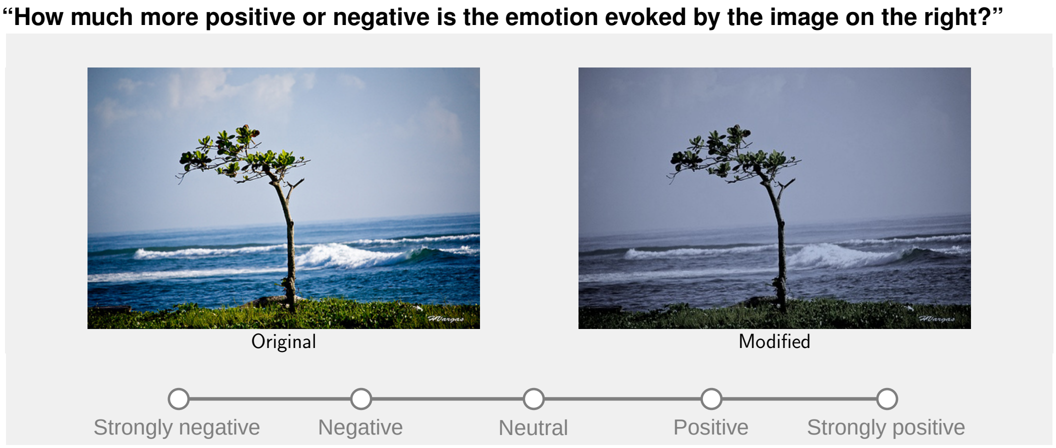
Figure 5.34: *Contribution of final modifiers. Estimated parameters, bounds and fadings are mapped into tuning space to derive the final filter. The emotional response is shifted towards negative or positve relatively to the input image.*

Local Contrast (*L*). In [DNSG07, CHLMT07] it has been reported that there is only a small relationship between emotional impression and spatial frequencies, especially in the higher frequency bands. However, we think that boosting detail or smoothing an image might still influence the overall impression. In order to prevent destroying the visual content by modifying the global contrast, we implemented the local contrast modifier I_L based on edge-avoiding wavelets (EAW) [FAR07, HDL11] to evaluate boosting and smoothing details on three scales of the EAW.

Combining Modifiers. Further, we aim at exploring the relation between the strength of the single modifiers and their resulting response, the limits within which the modifiers operate as expected, and suitable combinations to strengthen effects. Individual modifiers and combinations are evaluated in a user study.

5.3.5 EmoTune Model

Based on those mentioned modifiers, we develop a heuristic image filter that changes a given image such that the triggered emotion can be tuned relatively to the input using a single parameter. The final composition is visualized in Figure 5.34. The later described experiments reveal that only the modifiers B, S and T control the valence in a reliable way. Thus, we measure how much each single modification I_m with $m \in \{B, S, T\}$ changes emotion relatively to the input state I^0 of an image as well as compared to the combination of several modifications. We further exploit boundaries for each modifier towards the negative I^- as well as the positive I^+ direction. Those boundaries as well as empirically established relative strength allow us to project them into a *tuning-space* which maps the possible range of emotional changes onto a simple scale. The final image filter is derived as follows: Let $\lambda \in [-1;1]$ be the desired relative change in emotional response. Then, the final



"How much more positive or negative is the emotion evoked by the image on the right?"

Figure 5.35: *Example as seen by AMT workers. The task is to answer the question (top row) on a 5pt Likert scale (bottom row). Original photo by Hernan Vargas.*

image O is computed from the original image I as

$$O = (\mathrm{T}_{\gamma(\lambda)} \circ \mathrm{S}_{\beta(\lambda)} \circ \mathrm{B}_{\alpha(\lambda)})(I). \tag{5.5}$$

The final parametrization of the model is estimated in an extensive user study which is presented in the following.

5.3.6 Human Evaluation

We want to derive a suitable mapping from the presented modifiers to our final EmoTuner such that the viewer's evoked emotion is predictably changed. Thus, as emotions are highly subjective, we performed experiments on Amazon Mechanical Turk (AMT) to obtain general ratings from a wide variety of people on a high demographically diverse level [BKG11] (Section 2.3.1). Overall, 86 Turkers participated during our testphase and completed a total of around 57K HITs.

In our setups, each task is rated on a five point Likert scale in the range $[-2;2]$, noted as *{strongly negative, negative, neutral, positive, strongly positive}*, and is presented to 10 of the mentioned 86 Turkers. To explore the particular effects of the modifications, we designed experiments of four different types:

Generate Ground Truth. As the IAPS is a very well evaluated image database, we only verify ground truth for the subset of around 450 images we queried from Flickr. In this experiment, each image is categorized by Turkers based on the mentioned 5pt scale. For further processing, we only keep the images that were rated into the same category as they have been downloaded before and images of either category that have been consistently rated as neutral. Out of both, IAPS and Flickr, we randomly select 150 images: 50 negative, 50 neutral and 50 positive ones.

Evaluate Single Modifiers. This test identifies which modification influences emotion in terms of direction and amount. In a side by side setup, we display the original image and a modified version (Figure 5.35) and ask Turkers to rate "how much more positive or negative" the modified image seems. This time the scale item is to be selected relative. We generate modified versions by emphasizing or dampening the image property by a fixed value (e.g., brightened to 140% or darkened to 60% of the original intensity) and produce the final test set by random permutation. This experiment revealed that only *brightness*, *saturation*, and *color temperature* tend to control the valence in a reliable way whereas it was hardly altered by shifting *hue* or change in *local contrast*. Thus, we focus on the successful modifiers B, S and T.

Evaluate Modifier Limits. With the same setup as in the previous experiment, we evaluate the strength and consistency of the emotional change with respect to the applied strength of the modifier. This means that, for example, besides a 40% modification in the previous experiment we further test on 50% and 60% in both directions to see how the modification operates in extreme settings. For each modifier we determine the limit up to which the assumed effect is observable.

Evaluate Combined Modifiers. The successful modifiers are further tested on their joint effect when applied simultaneously. We generate images with all possible combinations, tuples, and triples for the previously determined modifier strength limits and perform a similar experiment as the evaluation of the single modifiers. The results show that the various modifiers are additive but influence the emotional response to different degrees. From these findings we derive relative weights for intuitive EmoTuning.

5.3.7 Results of Experiments

For all previously described experiments we analyze the mean μ and the variance σ^2 of the emotional changes induced by the modification over all images in each ground truth category. The observed trends were the same on the IAPS and the Flickr image set.

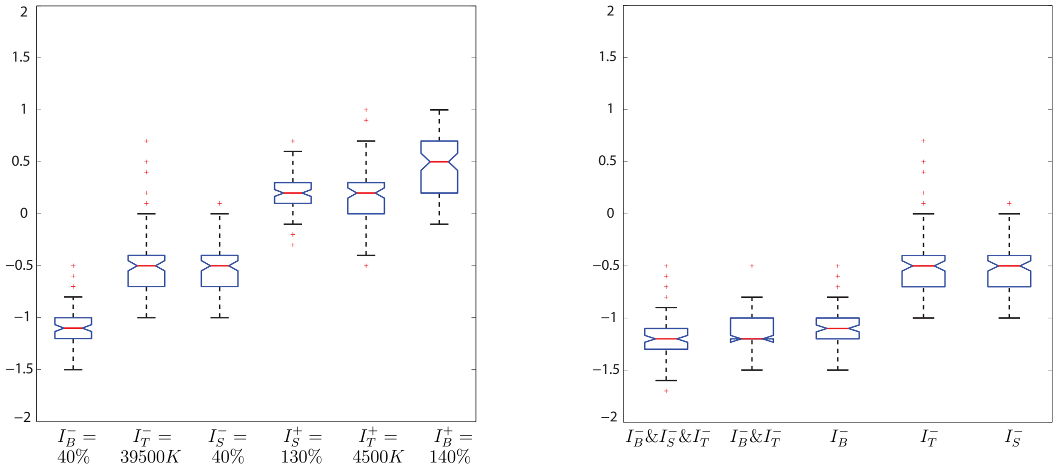


Figure 5.36: *Boxplots. Left: Our empirical boundaries. Right: Single modifiers and combined ones leading to negative direction of the EmoTune filter.*

Results for Individual Modifiers

First, we explore the effect of each individual modifier shown in Figure 5.33. Their produced shifts in emotional response for the most successful modifiers are summarized in Figure 5.36 (left).

Brightness. Changing the brightness of an image has the most profound effect. Darkening an image pushes the reaction towards negative, whereas brightening the image makes it more positive to look at, even if rather negative content is shown. In our tests, increasing the brightness to about 140% achieved the maximum shift in response while the darkening is stopped at a minimum of 40%. In the entire range, the response is monotonically increasing as indicated by the third experiment.

Saturation. Desaturating the colors of an image makes the emotional response more negative, while saturating the colors typically yields more positive responses albeit not as strongly. For saturation there is an upper limit to how much it can be increased before the trend gets actually less predictable and tends to be more negative (Figure 5.37). The predictable range is estimated to be between 40% and 130% of the original saturation.

Color Temperature. Changing the temperature towards cooler colors (i.e., more bluish) makes an image generally more negative. In the opposite direction, using warmer colors initially has the desired effect of provoking a more positive reaction. But, if the latter modification is applied too much, it can lead to a rather aggressive

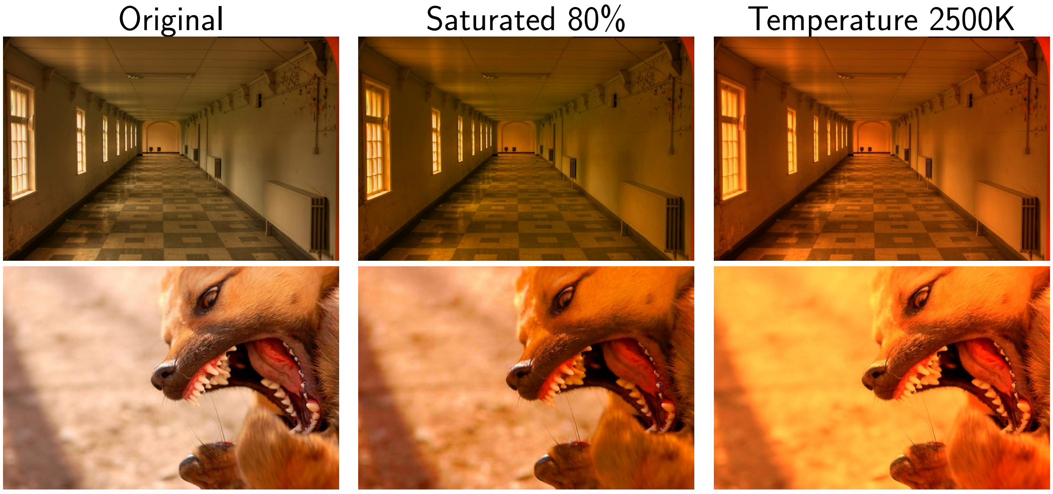


Figure 5.37: Example of extreme saturation and color temperature on neutral (top row) and negative rated input (bottom row). Both values lead to more negative impression (scary or aggressive). Original photos (left) by Leo Bissett (top), Laurentiu Nica (bottom).

appearance (especially red tones) and push the emotional response towards the negative (Figure 5.37, right). However, changing the temperature within the range of 4500K to 39500K only led to false labels in less than 5% of the input images.

Hue. Shifting the hue of an image in HSV color space quickly makes it look artificial, especially if humans are visible (Figure 5.33). Consequently, images with shifts in the hue were rated neither more positive nor more negative than the original. Even matting humans in the images to keep their hue information intact did not change the outcome. As the modification of hue was not successful we did not explore this attribute further.

Local Contrast. Similar findings were retrieved by changing the local contrast. Examples for the modifications are shown in Figure 5.39. As indicated in Figure 5.38, when increasing or decreasing the local contrast, the valence is hardly altered even if different frequency bands are involved. Most of the median values of the presented various scales are 0 indicating no predictable change. Only EAW $0-0-0$ and EAW $2-2-2$ resulted in a slight change to -0.1 and 0.1 respectively. But even for the slightly positive trend, the lower quartile has a stronger trend towards negative. Thus, our results confirm the previous findings of [CHLMT07, DNSG07].

Overall, several individual modifiers, namely the brightness, saturation, and color temperature, were able to reliably influence the emotional perception. Thus, we focus on these successful modifiers and evaluate combinations of them to derive our final filter.

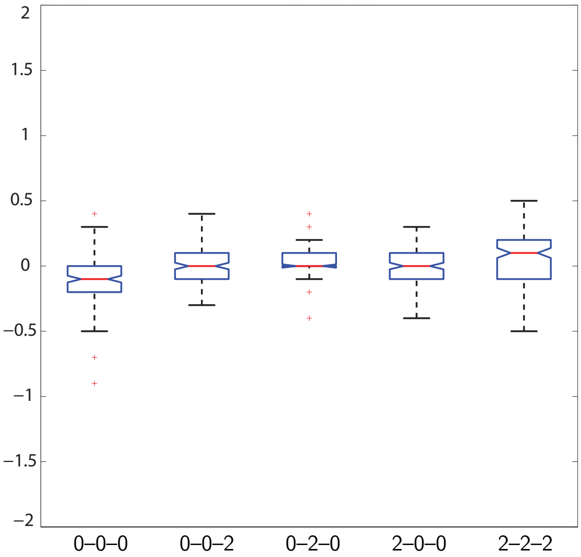


Figure 5.38: *Results for local contrast modifications. Altering the details on various scales in either direction of an EAW has hardly changed the perceived effect, even when including different bands (0 – removal of detail, 2 – details boosted).*

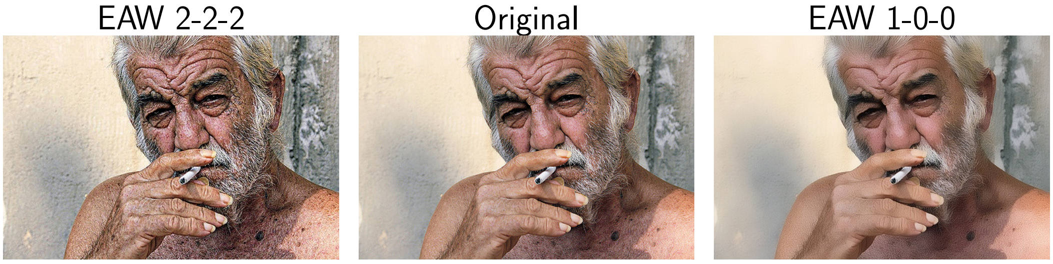


Figure 5.39: *Changes in the local contrast hardly effect the valence. Left: Boosting the contrast equally on all scales of an edge–avoiding wavelet. Middle: Original image. Right: Details on the first two scales removed. Original image by Flickr user: miss pupik.*

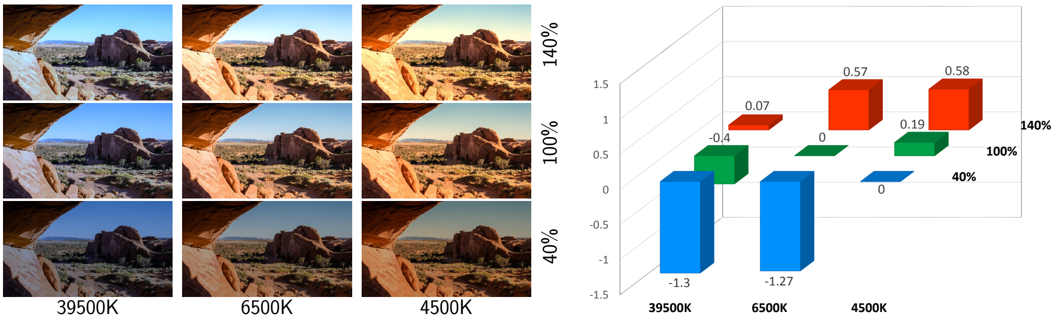


Figure 5.40: *Left: Example combining temperature T (cols) and brightness B (rows). Right: Average changes for combining T and B over all images arranged as on the left. The modifiers can amplify or dampen each other.*

Results for Combined Modifiers

From the individual modifiers, only the brightness, saturation, and color temperature tend to control the valence in a reliable way. Thus, in the last experiment, we analyze the effect of combining these successful filters. Thereby, each modifier is used up to its estimated reliable bounds.

Figure 5.36 left shows plots for the final modifiers and the determined bounds. The boxes indicate that the positive trend is detected for the positive modifiers and, similarly, the negative trend for the negative modifiers. Overall, the negative trend is detected stronger than the positive one. Further, our experiment showed that the effect on the emotional response is actually boosted if multiple modifiers are combined beyond what is possible with just a single modifier (Figure 5.40). Again, the example indicates a stronger effect towards negative direction, e.g., negative temperature (39500K) combined with darkening (40%) leads to a negative change of -1.3 on the rating scale $[-2;2]$. Those examples also visualize that applying one modifier in the opposite direction of another modifier partially neutralizes the effect. More specifically, combining negative temperature and positive brightness (140%) results in a minimal positive average change of 0.07 and, in the other direction, positive temperature combined with negative brightness even in no predictable change. In all tested combinations, the change in brightness had the most dominating effect followed by saturation and color temperature (Figure 5.36, right). Combining the modifiers strengthens the effect and also slightly increases the variance. Besides, pushing the modifiers to their evaluated boundaries and exploiting their derived combinations correctly shifted the emotional response in 92% on around 80 test cases. Overall, those results correlate with our initial observation $\mathcal{H}_{EmoTune}$ which indicates that global modifications can actually change the emotional response of an observer.

Table 5.5: *Estimated boundaries where modifiers still produce predictable monotonous changes. B, S are expressed with respect to 1 being the input image. For T the input image is assumed to be given at 6500K.*

Modifier	**negative boundary**	**mid-point**	**positive boundary**
Brightness (B)	0.4	1	1.4
Saturation (S)	0.4	1	1.3
Temperature (T)	39500K	6500K	4500K

Table 5.6: *Depending on the label associated with the input image the possible range of achievable valence tuning can be limited.*

		achieved change				
		- -	-	0	+	++
input label	- -	–	✓	✓	✓	–
	-	✓	✓	✓	✓	–
	0	✓	✓	✓	✓	–
	+	✓	✓	✓	✓	✓
	++	–	✓	✓	✓	–

Parameterization of EmoTuner

Based on the performed experiments, we derive a set of final modifiers with suitable parameters that are joined into the final EmoTuner (Figure 5.34). The provided empirical evidence indicates that the amount of change in emotion can be steadily increased by parallel application of the final modifiers within boundaries in which they produce predictable monotonous change (Table 5.5). The parameters for the model (Eq. 5.5) are computed as:

$$\alpha(\lambda) = \begin{cases} 1.4 \cdot \lambda & \text{if } \lambda \geq 0 \\ 0.4 \cdot \lambda & \text{if } \lambda < 0 \end{cases}$$

$$\beta(\lambda) = \begin{cases} 1.3 \cdot (\lambda - 0.61)/0.39 & \text{if } \lambda \geq 0.61 \\ 0.4 \cdot (\lambda + 0.52)/0.48 & \text{if } \lambda < -0.52 \end{cases}$$

$$\gamma(\lambda) = \begin{cases} 6500 - 2000 \cdot (\lambda - 0.65)/0.35 & \text{if } \lambda \geq 0.65 \\ 6500 + 33000 \cdot (\lambda + 0.55)/0.45 & \text{if } \lambda < -0.55. \end{cases}$$

The modification of brightness is dominant over the entire range while its gain reduces towards extreme values. The modifications in saturation and color temperature are faded in towards the boundaries to support a steady growth in response. The particular thresholds when the other modifiers fade in are derived by the relative strength of the modifiers. For example, normalized to the maximum extend of the negative brightness modification, the negative saturation only achieves 48% in the evoked valence scale and therefore is applied only in 48% of the negative range of λ.

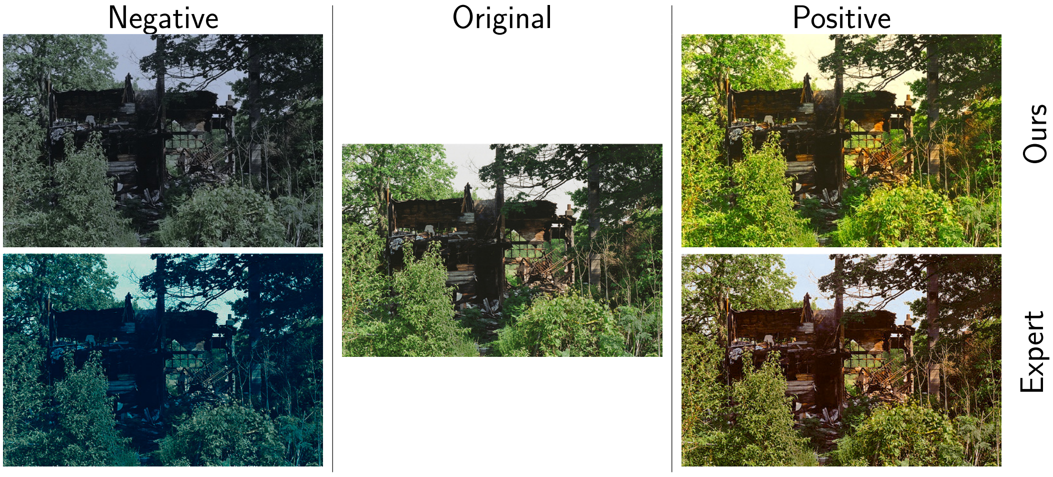


Figure 5.41: *Example comparison of our results (top row) to modifications of an expert (bottom row) to evoke a rather negative emotion (left) or to shift an image towards a more positive emotion (right). Although final results might still vary, a similar trend can be observed. Original Flickr image (middle): Bruce Nagel.*

5.3.8 Discussion

Overall, the modifications tend to influence the valence more strongly towards negative than towards positive emotions (Table 5.6). However, the absolute scale needs to be analyzed in a more detailed experiment. One aspect that we could not control in the AMT study is the display setting which can affect the strength of the individual responses. Anyway, by presenting the reference image on the same monitor the relative trend should still be kept.

One limitation of the proposed modifiers is that they cannot always overrule emotions triggered by semantic connotation of the image content. As shown in Table 5.6 original images that are labeled extremely positive or negative can only be emotionally tuned to a smaller extent than the average input image. Further, we demonstrated the importance to adhere to specific range limits for color saturation and temperature. Otherwise, one can for example easily produce the impression of being close to fire (Figure 5.37).

Comparison to Expert Grading. Additionally, in order to find out how an expert modifies images to evoke a rather positive or negative emotional response and if our tuned results indicate a similar direction, we requested a professional colorist to perform such modifications. On a subset of 30 images from the previously generated ground truth, we asked the expert to globally modify each of the images such that the evoked emotion is as positive or negative as possible. Figure 5.41 shows an example comparing our tuned results (top row) to the ones of the expert

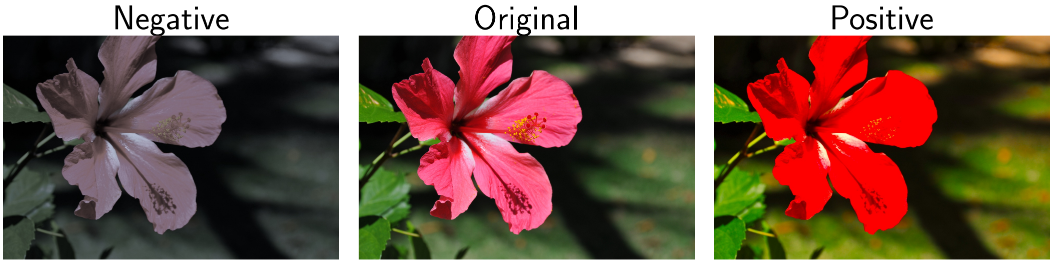


Figure 5.42: *Full range of the EmoTune slider. The estimated negative boundary (left) is relatively dark whereas the positive extreme (right) indicates an almost artificial look.*

(bottom row). The original image (middle) is altered to evoke a rather negative emotion (left) or shifted towards a more positive emotion (right). Overall, it seems that our proposed combination of modifications aims towards a similar trend. However, final results might still differ, e.g., compared to our negative version (top left), the negative suggestion of the expert (bottom left) seems extremely bluish. Nevertheless, an additional inking step could easily be applied on top of our version. Thus, although our filter can reduce workload for experts, they might need to add a finalizing step on top to fully cover their individual intentions.

5.3.9 Applications

Finally, our "EmoTuner" can be applied as a single slider to any input image and even works on videos.

EmoTune Slider Application. In a small application, we implemented the derived combination of modifiers as a single slider that allows the user to tune the evoked affect of an arbitrary image towards a stronger positive or a stronger negative response on the range $[-1;1]$ relative to the initial image (Figure 5.32). The example in Figure 5.42 demonstrates the full range of the EmoTune slider. As any image can be selected as input, tuning towards the negative effect might already appear relatively dark whereas the positive extreme can result in an almost artificial look. However, crossing those derived boundaries usually does not provide more benefit but can even turn the direction back towards the opposite effect as indicated in Figure 5.37.

Application to Videos. Further, we extended our emotional tuning application to work on videos. On individual scenes with little changes in lighting, the required emotional tuning state is simply applied to each frame. In videos with multiple scenes, we assume that the observer might request a different emotional appearance from scene to scene and support the user with an interactive tool (Figure 5.43). It

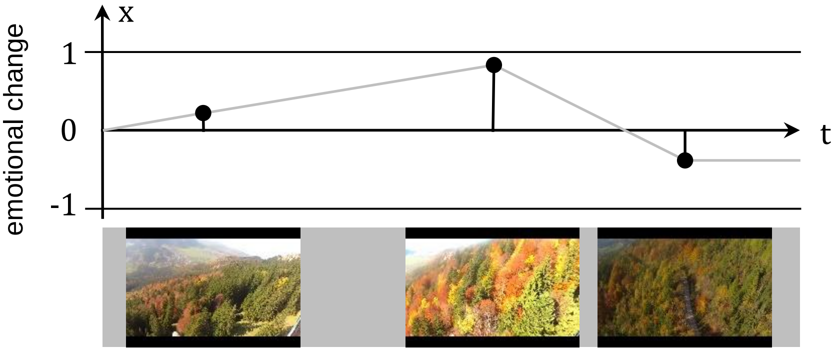


Figure 5.43: *User-supported tuning of video stream. The user can pick several frames (bottom row) and slide them to a required emotional state. Linear interpolation allows for a smooth tuning along the stream.*

allows the user to select single frames and to tune each of those selected frames separately to a required emotional state. To obtain a smooth tuning along the video, linear interpolation is applied between the tuned states of the selected frames. As long as the distance between the selected key frames is not too small, a temporally coherent video is generated.

5.3.10 Conclusion

In this section, we introduced a straightforward tool to modify the emotional response to images and videos. In a user study, we analyze the influence of simple image modifications on the evoked valence and reveal specific ranges where the tuning of brightness, saturation and color temperature work most efficiently or might produce unintended results. Based on these findings, we design our EmoTune filter to actively control the emotional response and combine the three successful manipulations to strengthen the effective change. Finally, we demonstrate applications to "tune" arbitrary images or smooth video sequences in real-time.

Of course, the filtered result cannot always overrule the emotion evoked by the depicted image content, e.g., a very negative image is rarely tuned towards positive. However, the filter works quite well in both directions for a broad range of input images. Besides, pushing the modifiers towards their bounds and combining them successfully shifted the emotional response on 92% of around 80 samples. This demonstrates the already large impact of those simple manipulations and correlates with our initial observation $\mathcal{H}_{EmoTune}$. In future, even stronger emotional effect could be obtained by evaluating other dimensions than valence or even modifying the image content and context. However, the investigation into content independent visual characteristics provides a wide usage for the EmoTune filter.

Overall, throughout Chapter 5, we have presented different approaches analyzing the appearance and structure of images under various aspects. We analyzed visual similarity over different domains on a large-scale and showed that similar meanings can be identified from the image structure even if their visual representation differs. Further, we explored the global appeal of pictures and, based on a huge diversity of ratings from social online behavior, found that aesthetically pleasing images are typically favored by people. Finally, we investigated in the emotional effect of images towards the observer and how simple modifications can influence the mood of a person. We evaluated the direct connection between images and an observer. Thereby, we found that the appearance of an image itself largely influences how its pictured content is transported to the observer, e.g., slight color modifications can actually influence the way an image is perceived. Changing the appearance of an image can even change its meaning towards more positive or rather negative. In the next chapter, we will make use of the immense expressive range of images and present methods to illustrate natural language with visual data.

6 Text Illustration

Whereas we previously focused on the meaningful analysis of the visual appearance of images, the main emphasis of this chapter lies in exploring the semantic connection between natural language and images in order to find a suitable illustration to a given text. Thereby, the main challenge consists in obtaining a semantically close translation from the textual description to a corresponding visual representation and, thus, to bridge the challenging semantic gap between those sources. We will demonstrate how basic textual processing employing techniques from NLP (Chapter 3) builds a foundation to generate suitable visual representations. Further, the previous chapter emphasized that the appearance of an image itself largely influences how its meaningful information is transported to the observer. Thus, we do not only aim for a semantically close visual representation to a given textual description, but also incorporate the appearance of images to present visually coherent picture stories in different visual styles.

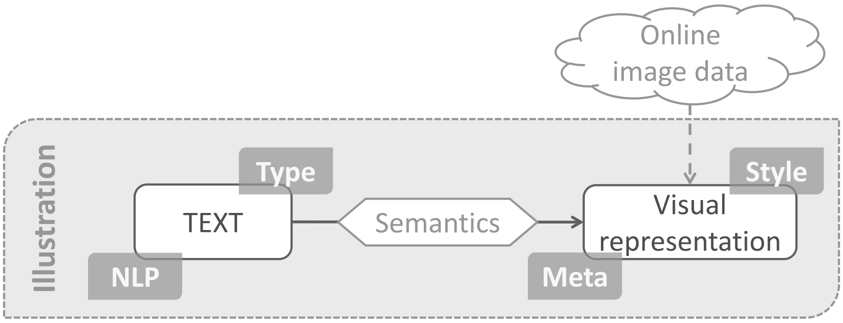


Figure 6.1: *Overview. This chapter focuses on illustrating text with a semantically close visual representation. Text is processed employing NLP. Visual data with associated meta information is retrieved from online collections. We present texts of different types and illustrations in various styles.*

Figure 6.1 illustrates the general approach of the work presented in this chapter. An input text is processed using techniques from NLP. We retrieve visual data from online repositories and make use of associated meta data to find a semantically relevant visual representation to the text. We present a method to generate 3D scenes (Section 6.1), an approach to illustrate different text types with relevant images (Section 6.2), and a framework that optimizes over semantic relevance as well as visual coherence along the storyline to create picture stories with style (Section 6.3).

6.1 Visualization of Descriptive Texts

Instead of building upon images, this section employs 3D scenes as visual representation. Although not aiming at illustrating text with images, the framework presented in this section serves as an example of how natural language can facilitate the creation of virtual environments. In other words, based on a textual description, a visual representation will be created in the form of a representative 3D scene.

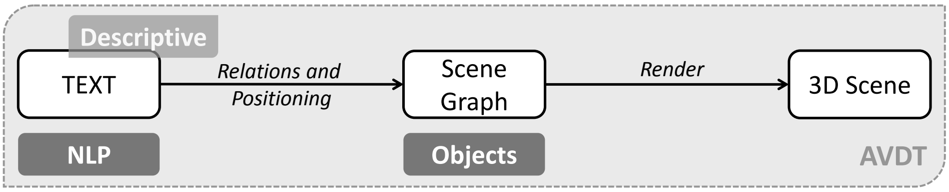


Figure 6.2: *High-level concept. From a descriptive text, information about relations and positions is extracted and associated with objects in a scene graph. The objects are linked with 3D models and rendered into a 3D scene.*

The high-level concept of our framework for the *Automatic Visualization of Descriptive Texts (AVDT)* is illustrated in Figure 6.2. An arbitrary descriptive input text is parsed using methods from NLP and relevant information about units, relations, and positions is extracted and associated with objects in a scene graph. The objects are linked with appropriate 3D models and, to correctly arrange the 3D models, the spatial dependencies of the entities are evaluated before they are finally rendered into a virtual environment. We state that:

> $\mathcal{H}_{AVDT}$: *Simple textual descriptions enable realistic arrangements of 3D objects in a virtual scene.*

As stated in $\mathcal{H}_{AVDT}$, we will show that extensive linguistic analysis of natural language employing methods of Natural Language Processing (Chapter 3) as well as adequate refinement allows us extract relevant information about existent units as well as their dependencies from the given text and, thus, to correctly arrange the 3D models in a virtual scene leading to natural looking virtual environments. The work presented in this section is based on the following publication: [SSDL11].

6.1.1 Introduction

Realizing thoughts in a 3D scene is still a very time-consuming and difficult process. Typical scene modeling tools tend to be overwhelming at first sight. Before starting to model the scene, the user has to familiarize himself with the supporting graphics software, i.e., learning all the menus and buttons and finding out, how to tweak

A **table** is standing *on* the **carpet**. On the *left* side of the **carpet** stands a **sofa**. A **couch** is in *front* of the **carpet**. *Behind* the **carpet** is a **shelving**. The **television** is standing *on* the **shelving**. On the *left* side of the **shelving** is a **loudspeaker** and a **subwoofer** is standing on the *left* side of the **loudspeaker**. Another **loudspeaker** is standing on the *right* side of the **shelving**. On the *left* side of the **television** is a **statue**. *On* the **table** stands a **bowl**. *In* the **bowl** lie **2 apples**.

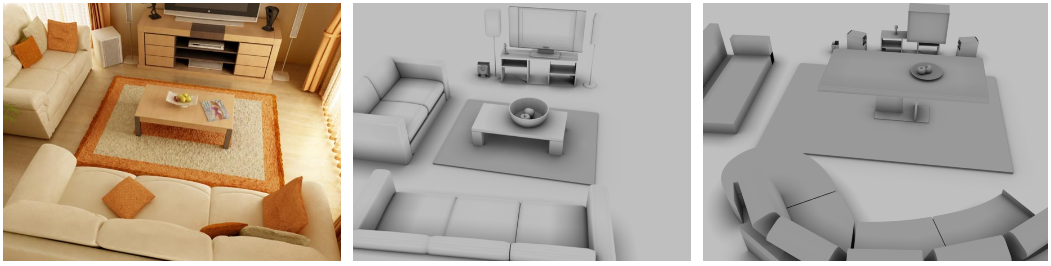


Figure 6.3: *Input text (top) describing common living room photo (bottom left) and resulting indoor visualizations by AVDT using different models (bottom middle and right).*

parameters. After that, the task of actually creating the visualization still remains. Therefore, a language-based approach, in which virtual environments are described and created directly through natural language and which does not rely on special graphics applications, simplifies the process of 3D scene generation.

In this section, we present a framework which automatically generates 3D scenes from natural language. More precisely, an almost arbitrary descriptive text serves as a basis for creating a specific virtual environment. The description consists of the objects occurring in the scene, as well as the spatial relationships between them. Our system concentrates on creating stage setups and, on this account, focuses on the key issues of Information Extraction (IE) as part of NLP, semantics, and graphical representation of a given text. Figure 6.3 shows an example photograph, a short text describing the setup in the picture, and two result outputs created by our AVDT system using different models. This indoor example will be compared to an outdoor result in Section 6.1.5.

AVDT System Overview. An overview of the AVDT pipeline is illustrated in Figure 6.4. The pipeline is based on the two processing elements *Automatic Scene Graph Generation* (Sec. 6.1.3) and *Natural Arrangement of Objects* (Sec. 6.1.4) each containing several components. First, a descriptive text is parsed and tagged by the IE process (Sec. 6.1.3). This extracted information is then gathered and refined (Sec. 6.1.3) which mainly refers to collecting information about real world objects in the text as well as spatial relations that indicate a connection between those entities. Furthermore, additional information (e.g., type, quantity, or position) is added creating new data elements, which we call *Part of Spatial Information* (POSI). A directed graph is then built by filtering unnecessary POSIs out of the text, whereas the remaining ones represent the nodes of the graph (Sec. 6.1.3). Every POSI representing an object is associated with an appropriate 3D model, which is stored

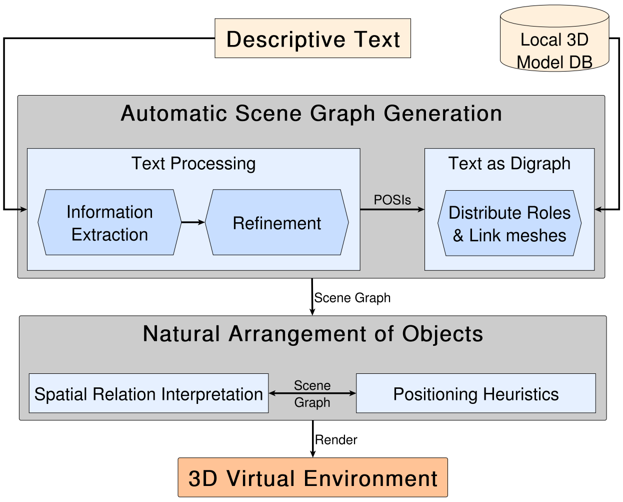


Figure 6.4: *AVDT system overview. A descriptive text is analyzed, extracted information refined and transformed into a directed graph. Then, object-to-object relations are evaluated, combined with positioning heuristics and rendered as 3D scene.*

in a local model database. For our processing, we assume that the 3D models we retrieve from Google 3D Warehouse[3] are modeled correctly and contain a clearly defined front side. The resulting scene graph builds the basis for the second part of the systems pipeline. The dependencies in the graph and the spatial relations are evaluated by first calculating the location of every POSI in the graph (Sec. 6.1.4) and, in a second step, applying positioning heuristics (Sec. 6.1.4) for increasing the "natural" look of the scene. Finally, the resulting virtual environment is rendered by a physics engine into a realistic stage setup enabling the user to fly around or move objects.

6.1.2 Previous Systems on Language-based 3D Scene Generation

Already quite a few projects have investigated the field of natural language input for creating virtual environments. Thereby, several systems have been developed. In the following, we present the ones that are most related to our work and motivated the development of our system.

The SHRDLU program [Win71] was one of the earliest systems that was able to understand and evaluate natural language. User interaction was allowed via simple english dialogs about a small blocks world shown on an early display screen. SHRDLU was primarily a language parser with the ability to use semantic information and context to interpret natural language input. However, the usable vocabulary for interaction was rather limited and the amount of referenced objects was restricted to a pre-existing environment.

Another system has been created by Adorni et al. [ADMF83]. Natural Language Input for Image Generation (NALIG) is able to interpret simple italian phrases of the form *<subject> <preposition> <object> [<reference>]*. The system disambiguates descriptions by defining primitive relationships between objects, represented as taxonomical rules, e.g., *H_SUPPORT(A,B)*.

In 1996, Clay et al. [CW96] implemented the Put system that uses a combination of linguistic commands and direct manipulation to correctly arrange and constrain rigid objects within a virtual scene. Although Put shares the intention of our system to ease the 3D scene creation process, it is limited to pre-existing objects and their spatial arrangements. Besides, it allows only a small subset of english expressions and uses a rigid syntax to formulate placement instructions.

One of the most well-known projects in the field of language-based 3D scene generation is WordsEye, created by Coyne and Sproat [CS01]. It generates static scenes out of a user-given text. An entered text consists of simple sentences that describe positions of objects and their orientations, colors, textures, and sizes. Although WordsEye realistically visualizes natural language input, the interpretation of spatial relationships often fails and the structure of the input is rather restricted. In order to generate "natural" looking scenes, the user is also required to use parameters for arranging objects, which takes time and effort. Contrary, in our approach we disregard colors and textures because we want to focus on correctly interpreting spatial relations without the need of user interaction, and keep the input more flexible. Therefore, we make extensive use of several Natural Language Processing methods. As WordsEye is closely related to our work, we discuss differences in more detail and show several examples in Sec. 6.1.6.

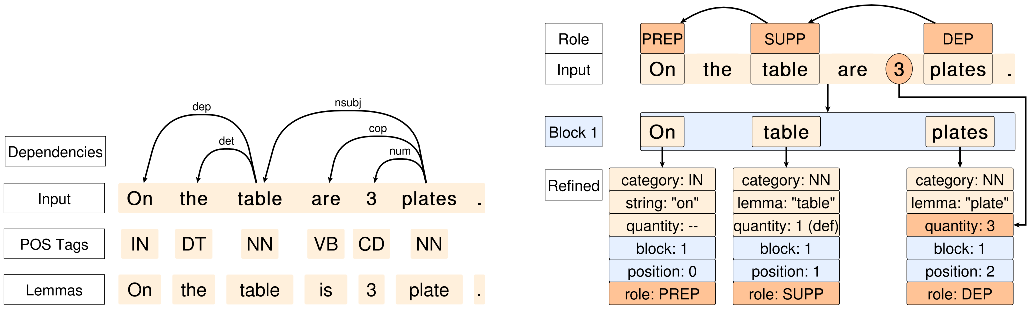
6.1.3 Automatic Scene Graph Generation

This section describes the first part of the AVDT pipeline on which our whole system is based. It consists of analyzing an arbitrary descriptive text and transforming it into a directed graph representation containing all scene relevant data.

Text Processing

As shown in Figure 6.5, first basic information about syntax and structure of the text is extracted in a pre-processing step (Figure 6.5(a)). Then, in order to build a data structure with all relevant information for 3D scene generation, refining of the extracted textual information follows (Figure 6.5(b)).

Information Extraction using GATE. As mentioned in Section 3.4.1, information extraction forms an important part of NLP and produces structured output from



(a) Information extracted by GATE

(b) Refinement by AVDT

Figure 6.5: *Text processing in AVDT. Basic information extracted by GATE (a) is refined by our system (b). Thereby, collected elements are labelled with their specific POS tag, their lemma, quantity, block index, and their position and role within a block.*

unseen documents [CL96]. Thus, in order to extract grammatical information, our system pre-processes an input text using an open source architecture called *GATE (General Architecture for Text Engineering)* [CMBT02]. GATE is a modular infrastructure that consists of various NLP software components including an IE system called ANNIE which is detailed in Section 6.2.3. Basic steps like tokenization (Sec. 3.2.1), sentence splitting (Sec. 3.2.2), and part-of-speech (POS) tagging (Sec. 3.2.2) are performed by ANNIE. The relevant results we use from GATE for further processing are visualized in Fig. 6.5(a) and listed in the following.

- **POS Tags.** From the part-of-speech (POS) tags annotated by ANNIE, we mainly use nouns (tagged with NN, NNS, NP, or NPS) and prepositions (IN). The nouns form the basis for finding adequate 3D models and the prepositions serve for interpreting spatial relations between objects. Furthermore, cardinal numbers (CD) and coordinating conjunctions (CC) are used.

- **Lemmas.** Additionally, GATE provides a morphological analyzer which adds a token with its lemma form to each word (Sec. 3.2.1). We apply the lemmatization to nouns only in order to find their singular form. This helps us in identifying an accurate 3D model for our object. A search based on the plural form of a noun would rather result in a whole set of objects. But, as the quantity of an object is calculated separately, the plural form does not serve our system.

- **Dependencies.** In order to derive relationships (Sec. 3.2.2), our system also uses the plugin for the Stanford Parser integrated in GATE [CMBea10] to extract the Standford Dependencies [dMMM06]. Such dependencies represent grammatical relations between words in a sentence and are designed to ease the understanding of those relationships.

Refinement of Textual Information. For the purpose of creating a data structure that stores all the relevant information needed for creating a 3D scene, the next step consists in refining the extracted textual information. This mainly means structuring the input text into interpretable *blocks*, adding useful meta-data concerning quantity and spatial dependencies, and filtering out useless information without spatial contribution. Especially nouns and prepositions are collected for further processing because a preposition can determine a spatial *relation* between two *objects* (nouns). The results of this process are visualized in Figure 6.5(b). Out of the refinement, new data elements arise and we call them *Parts of Spatial Information* (POSI).

- **Blocks.** For enabling the interpretation process, we segment the text into smaller blocks by splitting sentences at punctuation marks and coordinating conjunctions. Every block consists of a preposition describing a spatial relation and two nouns that embody objects. Thus, it represents an object-to-object relationship. The block index as well as the position of a collected element within the sentence are stored (Fig. 6.5(b)).

- **Objects.** Real world objects, embodied by nouns, are fundamental for creating a virtual environment out of natural language input. We use the lemma form instead of the normal string. In case of plural nouns, refining also includes adding information about the object quantity. By default, AVDT initializes every detected object with a quantity of 1. If a quantity occurs in front of a noun, the number is replaced.(Fig. 6.5(b)).

- **Relations.** In grammar, a preposition is a part-of-speech indicating a relation between a noun and other words in a sentence. As we focus on identifying prepositions describing spatial relations in static scenes, we refer to Landau et al. [LJ93] who proposed a list of spatial prepositions in English. In order to ease the interpretation process (Sec. 6.1.4), we collected the prepositions describing static relations, grouped them according to their semantic meaning, and chose an arbitrary representative out of each class. *Under*, *underneath*, *below*, *beneath*, e.g., form a group with the representative *under*. Due to very small and insignificant differences in the meaning of the spatial prepositions within a group, we interpret them similar to the representative one. Furthermore, some spatial relations are hidden in compounds of a preposition and a prepositional phrase, e.g., "*on the left side*". In general, such a part of a sentence is used with another prepositional phrase construction like "*of the table*". For achieving a homogeneous spatial-relation-evaluation system, AVDT identifies spatial components like "*left*" and treats them as normal prepositions (Fig. 6.6).

- **Roles.** In order to indicate the purpose of an object within a block, it gets associated with a specific tag which is defined based on the sentence structure. Therefore, we use the dependencies extracted from GATE as a basis. As noted in Section 3.2.2, dependency parsers link heads with dependents. In contrast, we derive a structure where the head object is directly connected with its dependent object (Fig. 6.5(b)). Therefrom, we retrieve the two roles of

Input:	On	the	left	side
POS Tags:	IN	DT	NN	NN
Output:			left	
Retag:			IN	

Figure 6.6: *Prepositional phrases are identified and spatial components, such as "*left*", are extracted and refined.*

a *supporter* and a *dependent*. A supporter is specified as the noun following a preposition and not being related to another noun. Subsequently, a dependent refers to a noun that is grammatically related to a supporter.

- **Ordering.** Saving position information for the collected elements within a sentence is important as their chronological order within a block is not always the same. In order to keep our system as flexible as possible concerning the input technique, a user is allowed to enter his text in various ways. For example "*On the table is a vase.*" is accepted as well as "*A vase is on the table.*". Because AVDT interprets each block as a sequence with the following order,

  Preposition → Supporter → Dependent

  the system checks the elements of every incoming block for their correct arrangement. Therefore, sentences like "*A vase is on the table.*" are automatically reordered which leads to a robust and less sensitive input interpretation.

The resulting meta-data enriched POSIs can now be used to build a directed graph representation.

Representing Text as a Directed Graph

The illustration in form of a directed graph eases the process of calculating spatial relations because it clearly presents the dependencies between the retrieved POSIs and is simple to traverse. The noun POSIs, either acting as a supporter or a dependent in their blocks, are processed into the nodes of a directed graph. Furthermore, we identify nouns that refer to the same object and link them to the same 3D model.

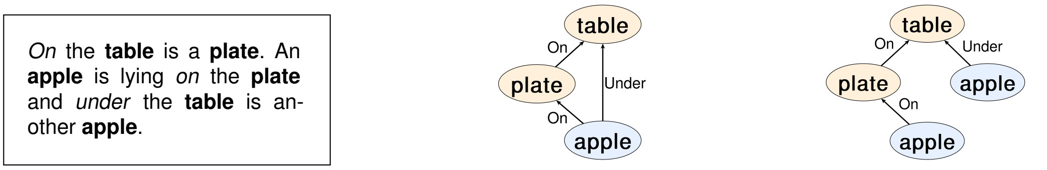
Graph Structure. While the root represents the origin of the scene, i.e., a root element is the only independent POSI in the graph, the leaf level contains all dependents that do not support any other object. Consequently, any other POSI occurring in the graph represents an object that supports another element but is also a dependent at the same time.

```
1) IF !( $POSI_i$.exists )
       LINK $POSI_i$ WITH new mesh

2) IF ( $POSI_{i_{Sup}}$.lemma == $POSI_{Sup}$.lemma )
       LINK $POSI_{i_{Sup}}$ WITH $POSI_{Sup}$.mesh

3) IF ( !( B( $POSI_{i_{Dep}}$ ) == B( $POSI_{Dep}$ )) && ( $POSI_{i_{Dep}}$.lemma == $POSI_{Dep}$.lemma ))
       LINK $POSI_{i_{Dep}}$ WITH new mesh
   ELSE ABORT BECAUSE Repetition

4) IF ( $POSI_{i_{Sup}}$.lemma == $POSI_{Dep}$.lemma )
       LINK $POSI_{i_{Sup}}$ WITH $POSI_{Dep}$.mesh

5) IF (( B( $POSI_{i_{Dep}}$ ) FOLLOWS B( $POSI_{Sup}$ )) && ( $POSI_{i_{Dep}}$.lemma == $POSI_{Sup}$.lemma ))
       LINK $POSI_{i_{Dep}}$ WITH $POSI_{Sup}$.mesh
       LINK $POSI_{Sup}$ WITH
                    ( $POSI_{Sup}$ FROM $POSI_{i_{Dep}}$ )
   ELSE LINK $POSI_{i_{Dep}}$ WITH new mesh

6) IF ( cycle )
       LINK $POSI_{i_{Dep}}$ WITH new mesh
```

Table 6.1: *Atomic rules link incoming* POSI_i *either with mesh of an existing* POSI *(same lemma) or with a new 3D model.*

Link with 3D Meshes. As nouns within a descriptive text often refer to the same real world objects, we invented several atomic rules (Table 6.1) in order to ascertain that POSIs that refer to the same object are linked to the same 3D object. We also link POSIs that use the same noun but refer to a different object in the real world with a new, separate copy of the same 3D model. Considering the sample text "*On the table is a plate. Beside the plate is a spoon.*". Obviously, both sentence blocks include the noun POSI "*plate*". Whereas in block 1 the plate implies the role of the dependent, it appears as a supporting POSI in block 2. Our rules for solving linguistic ambiguities define an incoming POSI, that has to be processed, as $POSI_i$. Already evaluated supporter or dependent POSIs are marked as either $POSI_{Sup}$ or $POSI_{Dep}$. Since the POSI-related block plays an important role, it is illustrated by *B(value)*, where the value is either an incoming, supporting, or dependent POSI. Furthermore, the word *cycle* in rule 6 refers to a linguistic cycle (Fig. 6.7).

All rules compare the lemma form of an incoming POSI with the lemma form of the existing POSI. Therefore, the "IF" clauses are true only if the lemmas of the two POSIs are the same. Rule 1 links a new incoming and not yet existing POSI with a new mesh. Rules 2 and 3 consider repeated supporters or dependents. We thereby assume that a supporter may serve for various dependents (2), whereas various dependents may depend from different supporters (3). Rules 4 and 5 resolve textual dependencies. This considers the fact that a POSI may act as a supporter as well as a dependent at the same time. Although such POSIs, containing different roles (as a supporter and a dependent POSI), have to point to the same mesh, it is important that they keep their information about the spatial relation (the supporting POSI of the former dependent is saved). As can be seen in rule 4, POSIs that act as supporter and dependent at the same time are created by linking the dependent POSI to the 3D model of the supporter POSI. Furthermore, the supporter POSI receives information about the supporter of the dependent POSI. This way, both POSIs are clustered under the same mesh keeping all relevant dependency information at the same time.



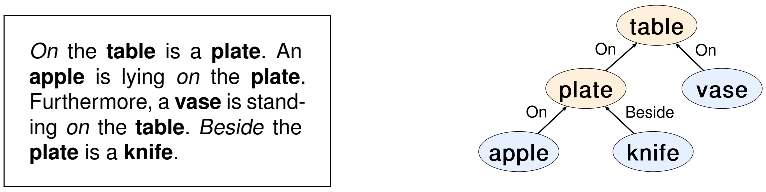
On the **table** is a **plate**. An **apple** is lying *on* the **plate** and *under* the **table** is another **apple**.

(a) Input (b) Cycle (c) Second dependent

Figure 6.7: *By creating an additional dependency (c), the linguistic cycle (b) is dissolved and a directed graph is achieved (c).*

Finally, rule 6 is responsible for linguistic cycles which may be used in order to create a stack of different objects. In this case, we want all those objects to obtain their own 3D model. The example in Fig. 6.7 shows that a cycle is dissolved by creating an additional dependency for achieving a directed graph structure. We create a second dependent POSI for the "apple" and link it to the table. This results in a clean hierarchical structure that is easy to be evaluated afterwards.

Clean Representation. By deleting remaining duplicates, we achieve a clean hierarchical representation of the analyzed text that can be seen as a directed graph with one root. An example is visualized in Fig. 6.8, where the AVDT system determines that the POSI, representing the table, is the root of the graph, since it is the only POSI that does not depend on another one. Consequently, the vase and plate POSIs represent dependents that also function as supporters. The remaining leafs of the graph illustrate dependents.

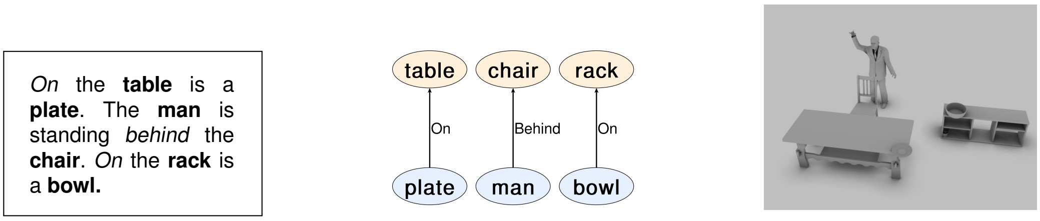


On the **table** is a **plate**. An **apple** is lying *on* the **plate**. Furthermore, a **vase** is standing *on* the **table**. *Beside* the **plate** is a **knife**.

(a) Input (b) Clean representation

Figure 6.8: *After duplicates are deleted, a clean directed graph representation is achieved (b).*

Unconnected Text. Another aspect refers to the problem of unconnected text descriptions. In case of incoherent textual information, our system is able to generate multiple digraphs that represent such texts. Multiple graphs are dissolved by placing an evaluated graph at a prior one (Fig. 6.9).


On the **table** is a **plate**. The **man** is standing *behind* the **chair**. *On* the **rack** is a **bowl.**
table
chair
rack
On
Behind
On
plate
man
bowl

(a) Input (b) Multiple graphs (c) Resulting scene

Figure 6.9: Unconnected text descriptions (a) create multiple independent graphs (b). Every graph is evaluated separately and situated at a prior calculated graph. The resulting scene is displayed on the right (c).

Representing a given arbitrary text as a directed graph significantly eases the process of calculating spatial relations between two related POSIs. For instance, in order to calculate the final position of the apple in the example mentioned above (Fig. 6.8), one must only regard the position of the supporting plate POSI. Due to the hierarchical structure of the graph, the plate already contains its final position and, therefore, the location of the apple can easily be computed as described in the following section.

6.1.4 Natural Arrangement of Objects

As we focus on analyzing grammatical syntax and dependencies in natural language and correctly mapping the results to objects and their relations within static scenes, any further parameters associated with noun POSIs, except information about quantity, are ignored. Aiming at enabling the user to easily depict a scene while receiving a basic nice looking result, the interpretation of spatial relations should not only be correct and accurate. Rather, it should also be capable of delivering a virtual scene with “natural” arranged objects, i.e., like one would expect it in the real world.

Spatial Dependencies Interpretation

As already mentioned in Sec. 6.1.3, we group prepositions concerning static spatial relations and chose a representative for each class. The representatives are all interpreted uniquely. Our system creates an axis-aligned bounding box (BB) for each object endorsed in the graph. Next, AVDT calculates the position of a bounding box, depending on the saved spatial relation and the final position of the supporter saved within the currently viewed POSI. The final location is then stored in a transformation matrix, which is combined with a rigid body later on. Finally, all rigid bodies are processed by a physics engine and the final scene is rendered.

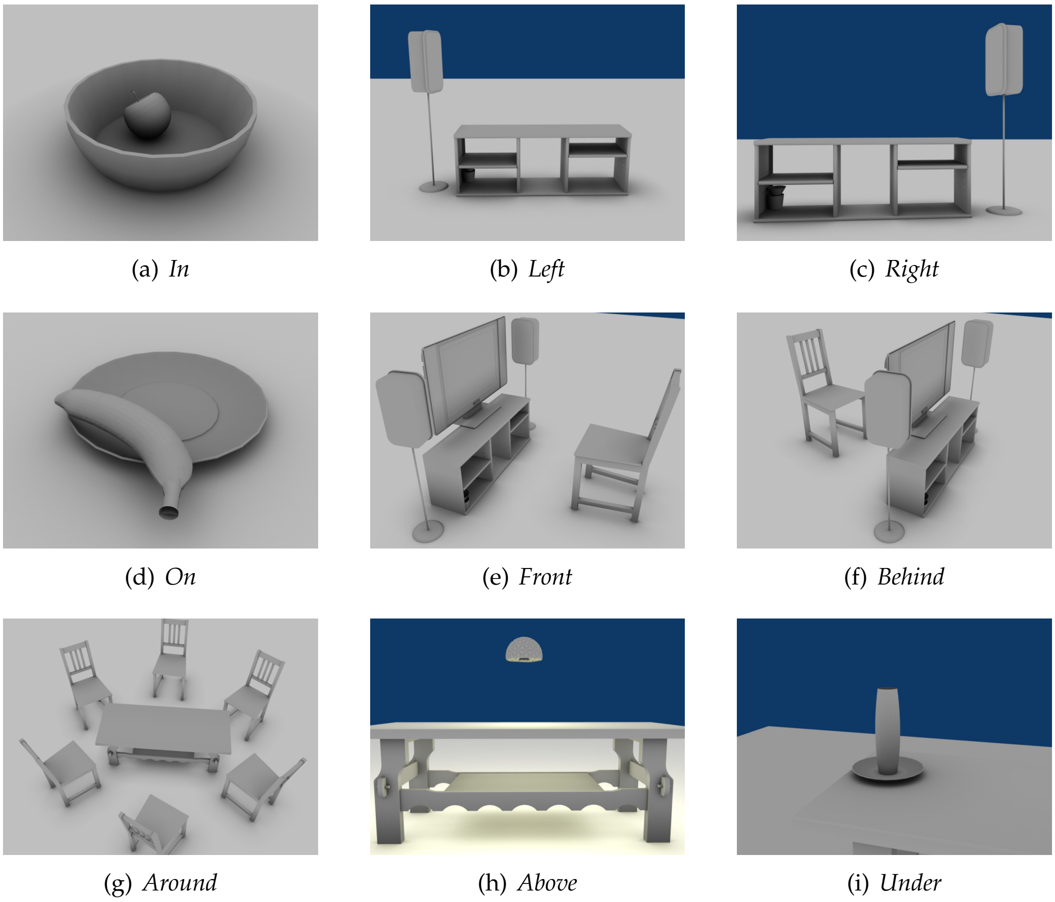
The pseudo-code shown in Table 6.2 illustrates how the spatial relation *beside* is calculated. Every dependent is placed randomly at one of the four sides of the bounding box of its supporter. Some more examples of relations we implemented and their resulting renderings are visualized in Fig. 6.10. The spatial relation *in*, for example, is evaluated by placing the dependent on the bottom of the supporters model. Because an object does not stand exactly on the center within another entity, random coordinates are used in order to vary the position of the dependent. If our system detects the spatial dependency *on*, the dependent is placed on top of the supporters bounding box. The position on the surface is not fixed and the system calculates random coordinates over the top surface of the supporter. However, in order to prevent a dependent from floating in the air or falling down when physics are activated, the computed coordinates do not cover the entire surface of the bounding box. Further, the spatial relations *left*, *right*, *behind*, and *front* are interpreted as fixed on a specific side of the bounding box. Our system evaluates the relationship *above* by applying no contact between the dependent and the supporter. Therefore, the depending object is placed in the "air" above the center of the supporter. Subsequently, the spatial preposition *under* is processed by situating the dependent below the supporter. This also requires changing the position of the supporting object as well as altering the locations of other possibly participating entities. However, the correlation *around* differs as being interpreted as a uniform distribution of one or more dependents around their supporter, placed on a circular orbit. This example especially illustrates that AVDT easily can be extended in order to increase the positioning possibilities.

```
FOR ( random side of supporters BB )
    CALCULATE height position for dependent
    IF ( random side not yet used )
        PLACE dependent
    ELSE IF ( other side free )
        PLACE dependent
    ELSE
        PREVENT collision PLACE dependent
```

Table 6.2: *Pseudo-code of preposition "beside"*

Positioning Heuristics for Natural Appearance

For improving the "natural" look of a 3D virtual environment, heuristics are applied within AVDT. These rough rules are used while a spatial relation is evaluated and calculated.



(a) *In*

(b) *Left*

(c) *Right*

(d) *On*

(e) *Front*

(f) *Behind*

(g) *Around*

(h) *Above*

(i) *Under*

Figure 6.10: *Renderings for some object–to–object relations.*

Distance Heuristic. The first heuristic refers to the distance of dependents to their supporter and calculates the distance ratio based on:

$$\underset{dim_{Dep}>dim_{Sup}}{H_{dist}} = \frac{(dim_{Dep}-dim_{Sup})}{c_d}$$

This means that in case the dependent has a bigger dimension (dim_{Dep}) than its supporting object (dim_{Sup}), the distance is evaluated from the difference of both dimensions. The result is divided by a constant c_d for normalizing the calculated distance to a realistic one. The larger the constant is chosen, the smaller the space between a dependent and its supporter gets. Whereas the maximum distance is achieved for $c_d = 1$, a value of $c_d = 1.7$ seems to create realistic stage setups. Based on this heuristic, a bed which is placed near a table is likely to be further away than a chair (Figure 6.11).

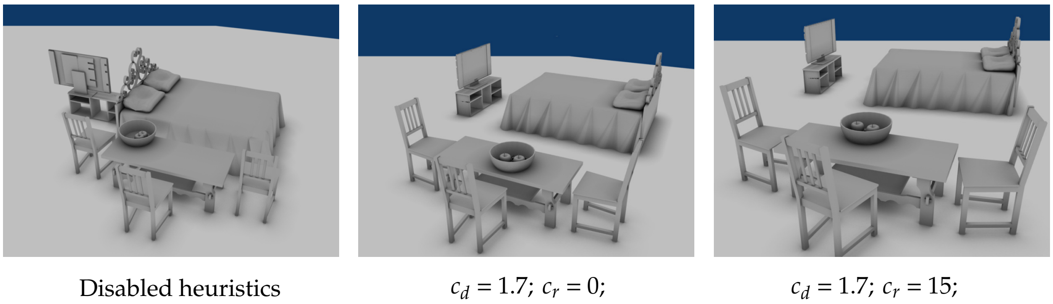


Figure 6.11: *Visualizations for different heuristic positioning parameters show different looks above same stage setup.*

Rotation Heuristic. For further increasing the realism of a 3D scenario, we use a second positioning heuristic, which is used to ensure that every dependent is facing its supporter in the scene. For example, an armchair that is standing on the left side of a table should face the supporting table that is located on its right side. Moreover, this heuristic applies a little random rotation to all of the occurring objects within a scene to achieve an "untidy" appearance. The rotation heuristic is defined as:

$$H_{rot} = \cos(\alpha + (\xi * c_r)) + \sin(\alpha + (\xi * c_r)) * \vec{V}_r$$

AVDT uses rotation quaternions, which are applied to the transforms of the axis-aligned bounding boxes. By passing an appropriate angle α as well as a vector describing the rotation axis, AVDT adjusts a dependent on its supporter. Further, a random number $\xi \in [-1,1]$ combined with a rotation constant $c_r \in [0,360]$ is used to destroy the perfect alignment of a dependent to its supporter and results in an untidy scene. In general, one would never arrange one's furniture such that it is perfectly aligned in the room ($c_r = 0$). The same principle applies in AVDT. Every object is randomly rotated in 2D space (3D for preposition *in*) in order to increase the naturalness within a virtual environment (Figure 6.11).

By applying positioning heuristics on the results of the position calculation, our system is able to further increase the naturalness within a created virtual environment, without using any extra parameters, e.g., size or distance specifications. Also, by varying the different parameters of the heuristics, one can achieve very different results for the same input text.

6.1.5 Results

We present some results for different input texts. As previously described, every POSI has been assigned with an adequate 3D model and rigid body. Also, every spatial relation has been evaluated and combined with positioning heuristics. Finally, the resulting virtual environment is rendered interactively, offering the user the opportunity to fly around and manipulate objects within the scene.

In *front* of a **cottage** is a **tree**. On the *left* side of the **cottage** are 2 **trees** and 3 **trees** are growing on the *right* side of the **cottage**. *Behind* the **cottage** is another **tree**. A **bench** is standing on the *left* side of the first **tree** and in *front* of the **bench** is a **man**. In *front* of the first **tree** is a **fence**. An **oldtimer** waits in *front* of the **fence**. On the *left* side of the **oldtimer** is a **lantern** and another **lantern** is on the *right* side of the **oldtimer**.

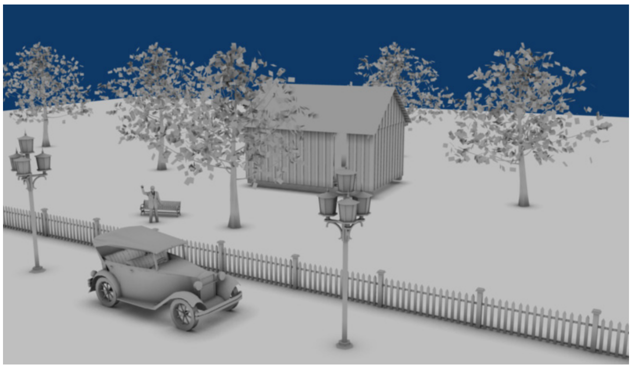


Figure 6.12: *AVDT result of an outdoor scene (right) based on a descriptive input text (left).*

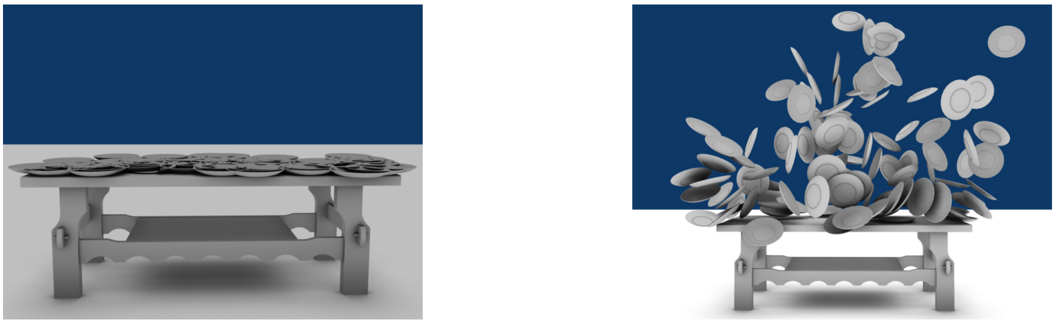


Figure 6.13: *AVDT allows interaction after rendering.*

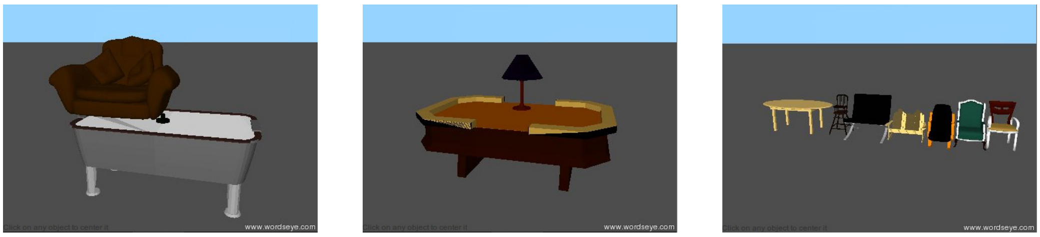
We compare an example of indoor scene visualizations (Figure 6.3) of AVDT to an outdoor scene result (Figure 6.12). The indoor examples show the same scene with different models (bottom middle and right) and are based on a simple text (top) describing a photo of a common living room (bottom left). As can be seen, the various objects are nicely placed in the space and are correctly arranged according to their spatial dependencies. Besides, the adapted models (bottom middle) result in a quite satisfying interpretation of the real world picture (bottom left). As the capabilities of the AVDT system are not limited to the generation of indoor scenes, the proposed algorithms and mechanisms are also able to create natural looking landscape scenarios. The visualization shown in Figure 6.12 gives a nice example of an outdoor scene rendering in which the arrangement of the several objects leads to a realistic impression. Overall, those examples demonstrate that our presented system is capable of translating descriptive texts into realistic arrangements of objects in 3D scenes as indicated by our initial statement $\mathcal{H}_{AVDT}$.

Due to the included physics engine, AVDT allows the user to actively manipulate the stage setup after rendering. Figure 6.13 gives a little impression on how the physics engine can influence the appearance of a 3D scene.

6.1.6 Discussion and Future Work

After the comparision to WordsEye, some further limitations are discussed serving as a basis for future work at the same time. Although our work already creates nicely looking 3D scenes from natural language, several research areas could increase the capabilities as well as the usability of AVDT.

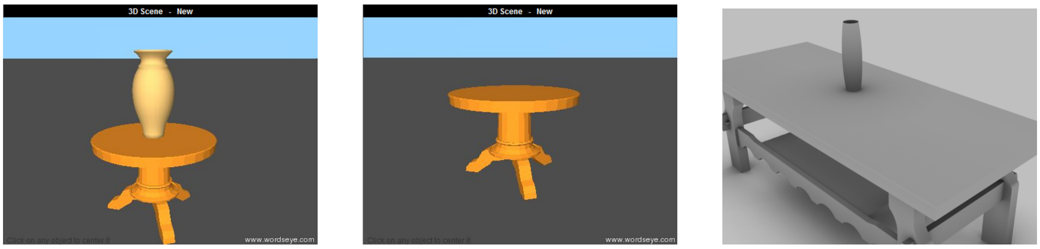
Comparision to WordsEye. As already mentioned in Section 6.1.2, WordsEye [CS01] is one of the most well-known projects in the field of language-based 3D scene generation. Natural language is visualized in a nice and realistic way. Thus, we want to compare some main differences concerning our work. In order to ease the modeling process, we mainly focus on an automatic realistic placement of objects. Thus, we treat every class of object-to-object relations differently and interpret spatial relations correctly, whereas WordsEye often needs additional user interaction and parameter tuning for realistic spatial arrangements. Some examples for failing placements in WordsEye are shown in Fig. 6.14 whereas our correct ones were already mentioned in Fig. 6.10. Besides, contrary to WordsEye, AVDT is capable of dealing with linguistic cycles. Furthermore, whereas WordsEye requires a defined pattern for user input, our system allows more comfortable input because of syntax reordering. An example is given in Fig. 6.15. Although our system allows for more flexible linguistic input, WordsEye, on the other hand, provides colors, textures, etc. As those attributes add a more realistic appearance to a scene, a further step in AVDT will consist in interpreting and including more context such as adjectives, and adverbs.


Click on any object to center it
www.wordseye.com
Click on any object to center it
www.wordseye.com
Click on any object to center it
www.wordseye.com

(a) *Left* (b) *Above* (c) *Around*

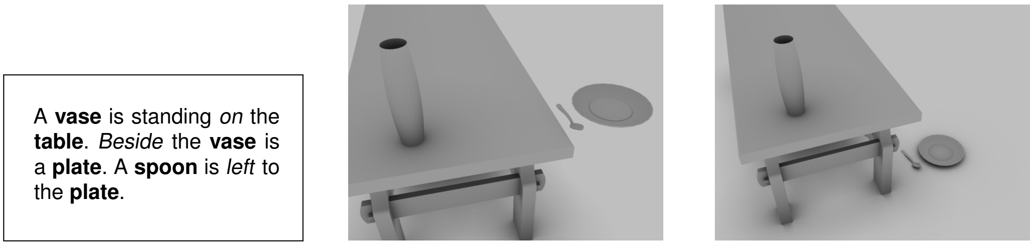
Figure 6.14: *Failing spatial interpretations in WordsEye.*

Positioning. Although the positioning system is stable, misplacements of objects may still occur (Fig. 6.16). This can be solved by developing a method that automatically adjusts the position of each object by traversing the graph backwards. At this point, a user can solve this problem by iterating the scene once again or by manually repositioning the objects. Further, wrong placements appearing due to limitations of a bounding box can be avoided by implementing object-tight


3D Scene - New
www.wordseye.com
Click on any object to center it
3D Scene - New
www.wordseye.com
Click on any object to center it

(a) WordsEye (b) WordsEye (c) AVDT

Figure 6.15: *Comparison to WordsEye concerning syntax stability. Input (a):* "The vase is on the table." *Input (b):* "On the table is a vase." *Both input texts result in (c) by AVDT.*



A **vase** is standing *on* the **table**. *Beside* the **vase** is a **plate**. A **spoon** is *left* to the **plate**.

(a) Input (b) Wrong placement (c) Result

Figure 6.16: *Lack of collision tests may lead to wrongly placed objects (b). At this stage, the misplaced objects fall to the ground due to the underlying physics engine (c).*

bounding boxes or switching to triangle-based collision detection. Unfortunately, this would increase computation time. Besides, as described in Sec. 6.1.4, AVDT does not evaluate any extra parameters for the purpose of creating natural looking virtual environments. Therefore, the AVDT system cannot evaluate spatial relations like "*On the second plate is a spoon.*" Furthermore, although our system already improves the appearance of a created 3D scene by applying positioning heuristics, it has no knowledge about how objects are used in real life, which complicates the determination of the orientation of objects and might also lead to wrongly interpreted locations. By combining the used NLP techniques with common-sense knowledge, as introduced in ConceptNet [HSA07], the positioning of objects could be significantly eased and enhanced.

Text Processing. Concerning text, several passages are filtered out intentionally. For example, our system does not allow repetitions since they are generally not used to describe a scene, at least in the normal linguistic usage. Moreover, AVDT evaluates cyclic expressions like "*On the table is a vase. A flower is in the vase. The flower is under the table.*" by generating a second dependent object for the flower. Both cases, repetitions and cyclic text segments, are recognized and filtered out or corrected by the rules described in Sec. 6.1.3. Anyway, due to the very complex nature of natural language, our system is not capable of handling all kinds of different text. Further enhancements in the amount of natural language processing could be achieved by extending the information extraction process. Using linguistic databases like WordNet [Mil95] for enlarging the semantic analysis of a text, or WordNet-Affect [SV04] for filtering unnecessary nouns like feelings or emotions, could improve the natural language understanding of AVDT.

Natural Appearance. Finally, for further raising the naturalness of a scene, the interpretation of adverbs or adjectives that indicate attributes of objects could be added. Besides, it is not only important to position and orient objects realistically, but also to visualize their "behaviour". This includes depicting poses for (humanoid) objects by evaluating verbs and/or common-sense knowledge, or using the incorporated physics engine for physical simulations, e.g., skeletal dynamics, fluids, or surfaces.

6.1.7 Conclusion

The work we presented in this section further extends the border of language-based 3D scene generation. Our AVDT system enables a user to quickly generate virtual environments by using natural language as input. Starting from a descriptive text, relevant information about objects and spatial relations are gathered and refined. The findings are used to link retrieved entities to appropriate 3D models as well as deriving a directed graph representation of the text. With the aid of that digraph, spatial relations between objects are evaluated. The resulting locations and models are finally assembled in an interactive virtual environment.

Especially the development of a rule-based text interpretation module, as well as the clear and easy to traverse hierarchical graph representation and, also, the accurate interpretation of spatial relations in combination with positioning heuristics improved the success of creating virtual environments close to a user-given input text. Several results illustrated how often the intention of an underlying descriptive text is already properly visualized by the AVDT system and how objects are realistically arranged as indicated by our initial statement $\mathcal{H}_{AVDT}$. This work can be further extended and one can imagine numerous applications in which there is need for transferring spatial ideas in visual communication.

Overall, in this section, we demonstrated that simple descriptive texts can support the generation of virtual scenes and employed 3D objects as type of visual representation. In the rest of this chapter, we will show that arbitrary input texts can be illustrated with existing photographs. Thereby, our focus lies on the semantic relevance as well as the visual coherence between natural language and visual representation in the form of images.

6.2 Illustrating Text from Online Photo Collections

The previous section employed 3D scenes as form of representation to visualize descriptive texts. In this section, we also rely on a similar linguistic analysis but, in contrast, we now build upon images to illustrate written natural language. We present a first system to semi-automatically illustrate a given, short text with semantically close pictures we retrieve from online photo collections.

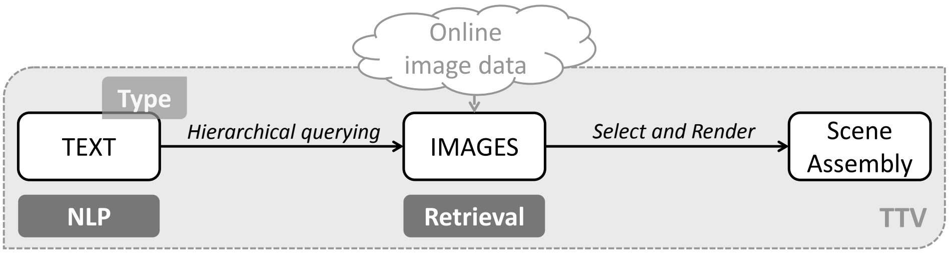


Figure 6.17: *High-level concept. A short text is parsed and relevant images are retrieved from online photo collections using a hierarchical algorithm. The final images are selected in a user-assisted process and rendered into a scene assembly.*

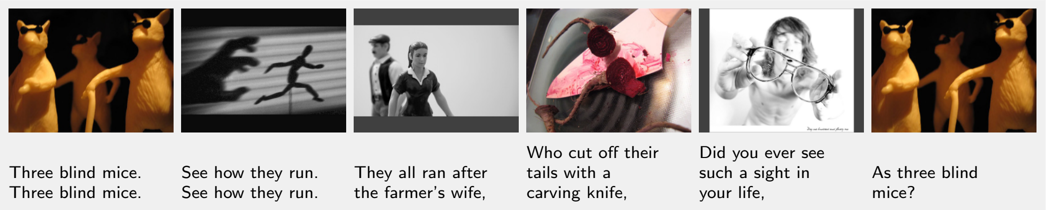
Figure 6.17 shows the high-level concept of our *Text-to-Video (TTV)* pipeline. An arbitrary input text is parsed and decomposed into suitable units to construct meaningful search terms. Using these search terms, we develop a hierarchical querying algorithm to retrieve a set of relevant candidate images from online photo collections. We state that:

> $\mathcal{H}_{TTV}$: *Hierarchically querying online photo collections results in semantically relevant images with high precision to a given text snippet.*

We will evaluate our hypothesis $\mathcal{H}_{TTV}$ on different types of texts in a small user study and demonstrate promising initial results. The final images are then selected in a user-assisted process based on which we automatically create a storyboard or a photomatic animation. The material of this section is based on the following publication: [SRC^{+}10].

6.2.1 Introduction

As motivated in Section 1.1, when aiming to tell a story, it is common to use natural language although pictures have at least a similar expressive range to describe particular content or evoke certain emotions. However, telling a story with images is an immense challenge as creating images or even videos is very time-consuming and the outcome largely depends on skill. Thus, in this section,



Three blind mice. Three blind mice.

See how they run. See how they run.

They all ran after the farmer's wife,

Who cut off their tails with a carving knife,

Did you ever see such a sight in your life,

As three blind mice?

Figure 6.18: *Semi-automatically generated storyboard with images retrieved from Flickr for the nursery rhyme "Three blind mice" [BGBG62].*

we present a framework that simplifies the process of telling a story with pictures, but eliminates the need to create the images yourself. More precisely, we aim at generating visualizations driven by natural language, augmenting written text semi-automatically by a sequence of images obtained from online photo collections. As the system itself may not be able to evaluate the semantic content within a resulting image sequence, the user shall be integrated as a controlling instance and obtain the opportunity of manual adjustment. Figure 6.18 gives an example for the semi-automatical illustration of a nursery rhyme [BGBG62].

TTV System Overview. As already mentioned, the main task when aiming at a relevant illustration is to obtain a semantically close translation from natural language to image sequences. As currently neither the semantics of written text nor of images can be automatically extracted with sufficient success rates by freely available tools, we provide a user-in-the-loop solution.

Rather than semantically analyzing the text, individual sentences are parsed to extract the functional description for the individual words. In order to determine semantically matching images we rely on the tagging of images in photo-community sites such as Flickr. From the extracted parse trees, we formulate optimized search queries to obtain as specialized images as possible for each part of a sentence. Typically, for each query a set of images is returned. The ultimate choice of which image to include in the final output is left to the user. We provide automatic means to select images with similar color distribution along the sentences and mechanisms to reuse images for similar text parts or to handle the protagonist of a story. The images can be presented in various formats, e.g., as storyboard or as slide animation. The quality of the image sequence largely depends on the available tagged imagery and the semantic complexity of the input text. Issues that are not correctly resolved by the parser will typically yield unsatisfying results but can often be corrected with little user intervention.

Figure 6.19 outlines an overview of our *Text-to-Video (TTV)* pipeline which mainly consists of three parts. The information extraction part consists of automatically parsing and segmenting a given input text into parts which we refer to as *POTs (Part-of-Text objects)* (Section 6.2.3). For each POT we construct an optimized search

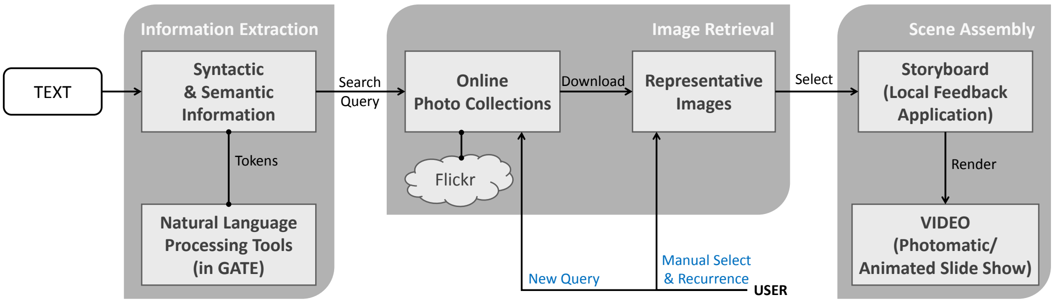

TEXT
Information Extraction
Syntactic & Semantic Information
Tokens
Natural Language Processing Tools (in GATE)
Search Query
Image Retrieval
Online Photo Collections
Flickr
Download
Representative Images
Select
Scene Assembly
Storyboard (Local Feedback Application)
Render
VIDEO (Photomatic/ Animated Slide Show)
New Query
Manual Select & Recurrence
USER

Figure 6.19: *Text-to-Video (TTV) system overview. After information is extracted from the input text, a query is submitted to an online photo collection to find representative images. The retrieved pictures are then assembled in a storyboard. Processes providing user interaction are indicated in blue.*

query (Section 6.2.4). The image retrieval part (Section 6.2.2) consists of automatically querying the online collection Flickr and retrieving a set of candidate images for each POT. Finally, the scene assembly part (Section 6.2.5) consists of automatically or semi-automatically picking the most representative images (*shots*) per POT. We present results on a nursery rhyme, fairy tale, screen play, short story, and a news article (Section 6.2.6).

6.2.2 Image Retrieval

As previously mentioned, online photo collections like Flickr provide a tremendous amount of imagery. Images on Flickr are attributed by titles, tags, and texts by users in an informal way yielding a so called *Folksonomy* [GT06] (as opposed to the more formal ontology). Flickr allows for a full text search in each of these three categories. Each query will potentially return a set of images. In the field of information retrieval, the quality of retrieved documents can be measured in terms of the precision of the returned results (Section 3.4.3). This measure (Eq. 3.3) can be transferred to our problem with images being the retrieved information:

$$precision = \frac{|\{relevant\ \ images\}| \cap |\{retrieved\ \ images\}|}{|\{retrieved\ \ images\}|} \tag{6.1}$$

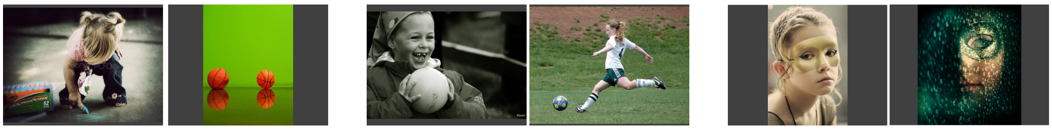
Thus, one can measure the precision of the answer with respect to a query by manually counting the number of matching images (Eq. 6.1). Note, that the recall measure can not be used as the knowledge about how many images for a query exist in all online photo collections in the web is not given.

Table 6.3: Compound query results for fairy tales. Precision (Equation 6.1) in percent and number of total successful queries in parenthesis. A query was counted as successful, if at least one matching image was retrieved.

Search Tokens	Queries	Full Text Search	Title Search	Tag Search
Combined Nouns	25	21% (25)	21% (20)	36% (24)
Nouns & Adjectives/ Adverbs	25	17% (25)	25% (19)	36% (21)
Nouns & Verbs	25	8% (25)	13% (17)	16% (11)
Nouns & Averbs	20	5% (12)	10% (10)	12% (5)
Nouns & Adverb Stems	20	8% (9)	3% (7)	14% (5)
Nouns & Verbs	20	8% (20)	13% (13)	15% (6)
Nouns & Verb Stems	20	10% (19)	14% (12)	25% (11)

Basic Tokens and Stemming

In order to obtain an image which matches the semantics of a POT, we need to assemble a compound query. By far the most frequent category of words in Flickr tags are nouns (about 90%) while verbs, adjectives and adverbs are found less often. Most often, querying for a single noun ignores the information of the remainder of the POT. By combining the nouns and attributes of one POT by conjunction, more and more specific images can be retrieved, such that they finally match the desired semantics (Fig. 6.20). Thus, combining nouns or adding adjectives, adverbs, or verbs can help in retrieving a more specialized collection of images even though the precision for the query per se might drop (Table 6.3). In particular, the precision for queries including verbs is low. But most often the retrieved images show the action represented by a sentence much better. Fig. 6.21 demonstrates this with the retrieved images for the queries *cocoon* (noun), *cocoon emerges* (noun & verb), and *cocoon emerge* (noun & stemmed verb). As shown in the second half of Table 6.3, the precision for adverb or verb queries can also be improved by querying for the stem rather than the inflected form.



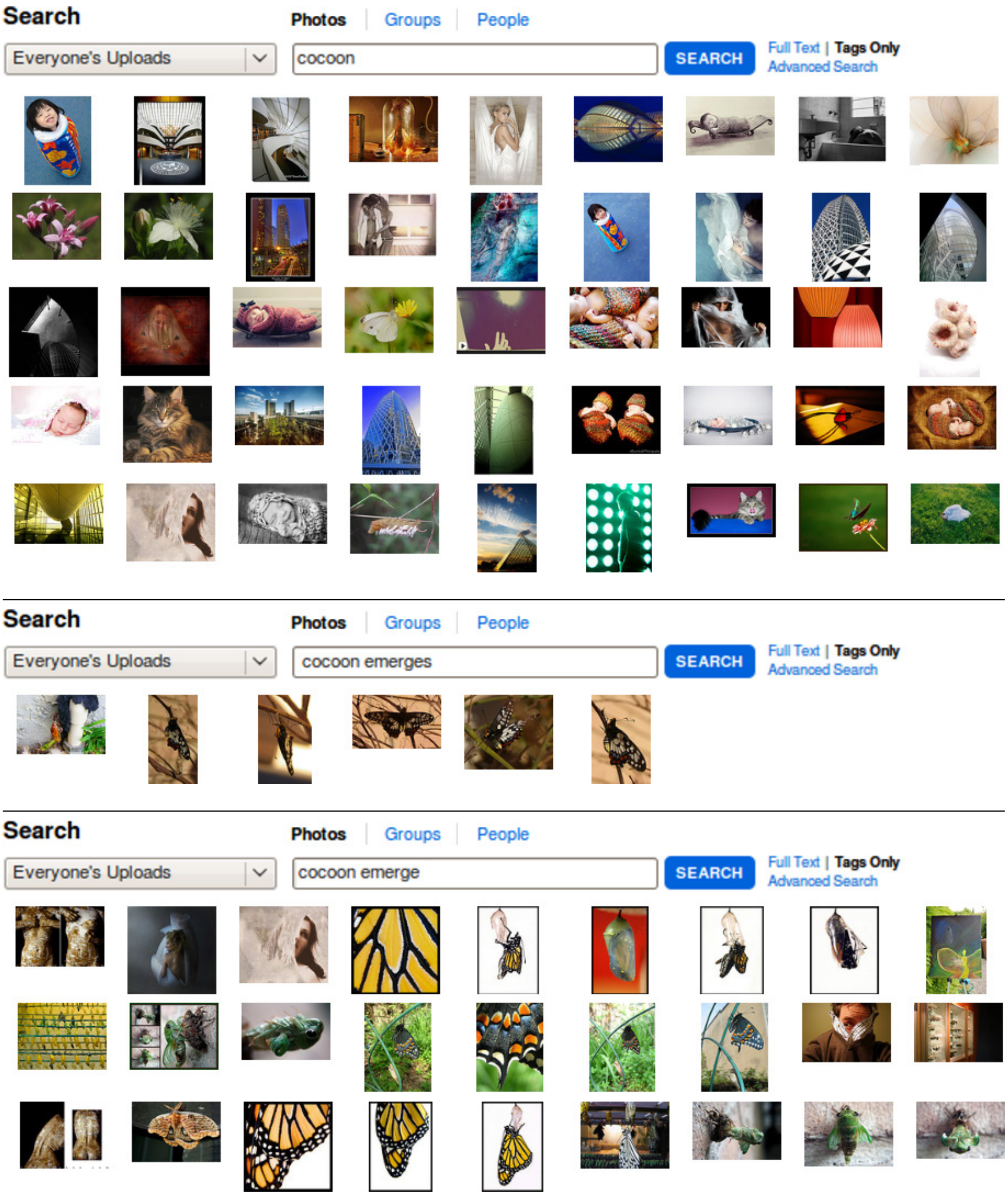
girl ∨ ball

girl ∧ ball

girl ∧ ball ∧ beautiful

Figure 6.20: *Successive specialization of a shot for "the beautiful girl at the ball".*

The highest precision is typically obtained searching in tags, but verbs and adverbs are rarely found in tags while they are more frequent in full text or title search. We will use these insights in Sec. 6.2.4 to assemble an optimal query for each POT.



cocoon

cocoon emerges

cocoon emerge

__Figure 6.21:__ Improving the correlation of shot and sentence action by combining nouns with verbs (middle) or verb stems (bottom). Flickr search results (Accessed March 6, 2010).

6.2.3 Text Analysis

As indicated in Figure 6.19, analyzing the text creates the fundamental step in our TTV pipeline. An arbitrary input text is automatically parsed and basic information is extracted by the system. Thereby, each word is assigned with a token describing its part-of-speech (Section 3.2.2). Besides, in order to find appropriate images that match the story, the text has to be split into suitable smaller parts whereby the chosen size of the text part affects if relevant images can be found. Thus, incorporating information about the extracted tokens, semantic text segmentation is performed after the IE process. For the resulting text segments, the extracted tokens are used later in the pipeline (middle part in Figure 6.19) to generate meaningful queries which will be described in Section 6.2.4. Further details of our semantic text analysis procedure were presented in [Sch10].

Information Extraction using GATE

As mentioned in Section 3.4.1, the process of extracting information from unseen documents to produce structured output [CL96] forms an important part of NLP. Thus, in order to form versatile queries for image retrieval, we extract syntactic and semantic information from the story. Therefore, similar to the text processing performed by the previously presented AVDT framework (Section 6.1.3), we again use the text engineering architecture GATE [CMBT02] as it provides an automatic application flow. When parsing a document, GATE converts it into a single unified model of *annotations* [BTMC04]. Annotations encoding the data read and produced by GATEs processing resources as well as the originally formatting information from the document are associated with each document.

Basic Annotations from ANNIE. The modular infrastructure GATE is, beneath others, distributed with an IE system called *ANNIE (A Nearly-New Information Extraction System)*. ANNIE annotates the input text with tokens and creates an associated annotation set out of the extracted information for every token. Those annotation sets are then returned from ANNIE in form of a list for further processing. The components from ANNIE that are relevant for this thesis are illustrated in Figure 6.22. "Tokenizer" and "Sentence Splitter" are required to annotate each word or symbol with a part-of-speech tag by the "POS Tagger".

- **Tokenizer.** As outlined in Sec. 3.2.1, the aim of tokenization is the division of a text into single tokens. The *English Tokenizer* included in ANNIE comprises a normal tokenizer and a transducer to adapt the generic output of the tokenizer to the requirements of the later executed English POS tagger [GRCH96]. For example, a negative construct like "don't" is converted by the transducer from the three tokens "don", " ' ", and "t" into the two tokens "do" and "n't".

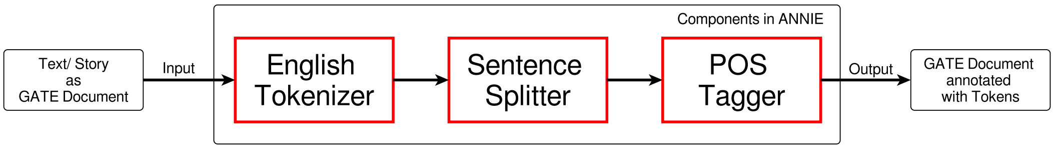


Figure 6.22: *GATE pre-processing with ANNIE components to get basic annotations. Tokenizer and Sentence Splitter are required to annotate each word or symbol with a part-of-speech tag by the POS Tagger.*

- **Sentence Splitter.** As indicated in Sec. 3.2.2, the task of a sentence splitter is to divide a text into its sentences. Thereby, the challenge consists in finding the punctuation marks belonging to a sentence instead of a word (e.g., as abbreviation). The ANNIE sentence splitter therefore makes use of a list of abbreviations to look up the words followed by a punctuation dot which aids in filtering them from full stops marking the end of a sentence [GRCH96].

- **POS Tagger.** As described in Sec. 3.2.2 a POS tagger classifies the words within a text. ANNIE uses a tagger proposed by Hepple [Hep00] which is a variation of the Brill tagger [Bri95] and produces a part-of-speech tag as annotation for each word or symbol [GRCH96]. Overall, the POS tags used by the Hepple tagger are similar to the Penn Treebank tags, whereof the ones relevant to this thesis are listed in the Table 3.1. The ANNIE POS tagger uses a default lexicon and ruleset (the result of training on a large corpus taken from the Wall Street Journal), which can be manually modified if necessary [GRCH96]. The tokens, created by the English Tokenizer beforehand, are then matched with the lexicon and added to the annotation set of the according token if a match is found.

The tokens determined by this pipeline (Fig. 6.22) form the basis for generating our queries in order to retrieve images.

Additional Tokens from Stemming. Furthermore, we discovered the usage of verb stems as a significant improvement concerning the query answers (see Section 6.2.2). GATE provides the *Snowball Stemmer* as plugin [GRCH96] which functions as a wrapper for the freely available Snowball stemmers [1] and is a tool based on the Porter stemmer for English which has been described in Section 3.2.1. It can easily be executed on the input text after some ANNIE processes were run and additionally annotates each token with its stem. In this TTV system, we apply stemming to verbs and adverbs only.

[1] `http://snowballstem.org/`

Semantic Text Segmentation

A reasonable segmentation of the text is necessary to enable a semantic comparison between a text segment and a picture. An example sentence with relevant POS tags is given in Figure 6.23.

“*Over there is*	a	**blue**	**car**	and	a	**green**	**bicycle**	*that looks exactly like mine!*”
		JJ	NN	CC		JJ	NN	

Figure 6.23: *Example sentence with part-of-speech tags relevant for text segmentation.*

The middle part of this sentence mentions “a blue car” as well as “a green bicycle”. Intuitively, it is clear that the car is meant to be blue and the bicycle should be a green colored one. However, querying for all these tokens at once might result in mixed up colors for the car and bicycle items which can actually change the meaning of the text. However, as the IE process extracts nouns (NN), adjectives (JJ), but also coordinating conjunctions (CC) we make use of this information to perform a meaningful segmentation of the text. In the example, splitting the text at the coordinating conjunction “and” correctly attaches the adjectives to their corresponding nouns.

Choosing the size of a text part largely influences how good the retrieved images might match. Too large segments might describe too complex situations for which no matching images are available in the web. In contrast, too small text parts might lead to many images that rarely display the described content. A suitable segment size is needed that balances the number of query results and their matching quality. To derive suitable sets of tokens from meaningful text segments, we perform the following two steps:

- Starting with sentences as the most general entities, they are split by punctuation marks or coordinating conjunctions (CC) in order to receive smaller segments, the POTs, which could be clauses, the entities of an enumeration or similar.
- Within these POT objects we query ANNIE tokens, in particular, the different types x of nouns (NNx), verbs (VBx), adjectives (JJx), and adverbs (RBx), as they contain the most significant information. For the same reason, we remove auxiliary verbs.

The next step is to find an appropriate image for each POT. In analogy to the filmmaking process, we call such a representative image a *shot*.

6.2.4 Forming the Query

Shots for the individual POTs are obtained by submitting proper queries to the online image collections. Using the list of tokens extracted for each POT (Sec. 6.2.3),

POT:	“the *poor* ***girl*** *bore* it *patiently*”
Noun list:	***girl***
Token list:	*bore*, *poor*, *patient*
Priority list:	$Q = [\ (\textbf{girl} \wedge \textit{bore} \wedge \textit{poor} \wedge \textit{patient})\ \vert$ $(\textbf{girl} \wedge \textit{bore} \wedge \textit{poor}) \vee (\textbf{girl} \wedge \textit{bore} \wedge \textit{patient}) \vee (\textbf{girl} \wedge \textit{poor} \wedge \textit{patient})\ \vert$ $(\textbf{girl} \wedge \textit{bore}) \vee (\textbf{girl} \wedge \textit{poor}) \vee (\textbf{girl} \wedge \textit{patient})\ \vert$ $(\textbf{girl})\ \vert$ $(\textit{bore} \wedge \textit{poor} \wedge \textit{patient})\ \vert$ $(\textit{bore} \wedge \textit{poor}) \vee (\textit{bore} \wedge \textit{patient}) \vee (\textit{poor} \wedge \textit{patient})\ \vert$ $(\textit{bore}) \vee (\textit{poor}) \vee (\textit{patient})\ ]$

Figure 6.24: *Example forming a hierarchical query from the POT "the poor girl bore it patiently". Tokens are sorted into queue Q with decreasing priority. (Reprint from [Sch10]).*

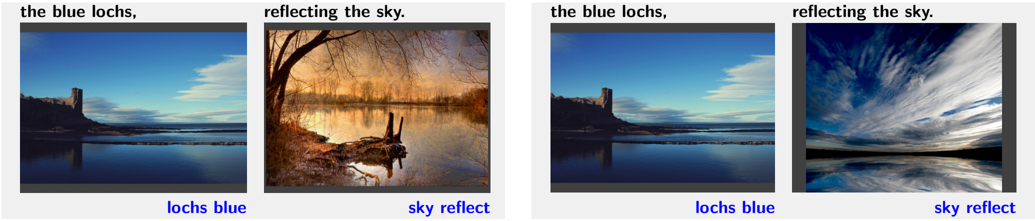
we formulate an appropriate query that results in a high precision and sufficiently many images to choose from. A conjunction of all n_t token T_i in a POT will produce the most specialized query. However, if too many constraints are added it might happen that only an unsatisfactory number of images is reported. We therefore assemble the query iteratively with the goal to find a sufficiently large set of images which are as specific as possible.

Following the evidence of Section 6.2.2, we create a priority list Q of compound queries for a shot. The first query is the conjunction of all n_t tokens, followed by a disjunction of conjunctions formed by all possible subsets containing $n_t - 1$ tokens, etc., until we end with a disjunction of all n_t individual tokens. For $n_t = 3$, we would assemble the following list of queries:

$$Q = \{(T_1 \wedge T_2 \wedge T_3), ((T_1 \wedge T_2) \vee (T_1 \wedge T_3) \vee (T_2 \wedge T_3)), (T_1 \vee T_2 \vee T_3)\}$$

Due to their importance in the syntactic analysis and their frequency in the image tags, we treat nouns in a special manner. Queries combining multiple nouns yield the highest precision and, therefore, we first treat the conjunction of all nouns as a single token in the algorithm outlined above, and in a second step create a list for each noun separately and append them. An example illustrating this procedure is given in Figure 6.24. The exctracted noun and adjective tokens (“girl”, “poor”) as well as the stem forms of the verb and adverb tokens (“bore”, “patient”) are sorted into a priority queue. The list starts with the highest priority (first line) decreasing downwards.

Based on the priority list, the system issues a sequence of queries, accumulating the downloaded images. The process is stopped as soon as the number of downloaded images for one shot exceeds a user defined threshold, e.g., 30. For each entry of the list, we first perform a query on Flickr tags and then perform a title search. This way, we were able to download 30 images for almost all of our example shots and due to the structure of the priority queue ensured that the most specialized images are always at the top of the image set.



(a) Original images

(b) Color consistency enforced

Figure 6.25: *Color consistency enforced on the selection for one query. The queries are indicated in blue.*

6.2.5 Scene Assembly

We now retrieved a set of candidate images per shot, sorted by relevance, from which we can automatically select the highest ranked image for each shot. Alternatively, this selection can be performed by the user. User selection will, for example, be necessary if the semantics, style, or composition of the highest ranked image does not match the user preference. To simplify the selection process, we provide the user with additional tools, one dealing efficiently with *recurring queries*, and another handling *color consistency* between neighboring shots.

More details on our scene assembly part as well as the aggregation of the images in the feedback application and the resulting video output were presented in [Sch10]. Thereby, an extension to separately handle the *main character*, i.e., the protagonist of a story, has been provided and will shortly be outlined in this section.

Recurrence. It is often desirable to use the same image for similar shots (see first and last image of Figure 6.18). Our system therefore reuses by default the selection results for shots with the same query. If the user chooses a different image for one shot, all other shots in this category are updated accordingly. Beyond a literal match, the user can manually group multiple queries together which will then be represented by the same image.

Color Consistency. Online images typically vary largely in style and color. Similarity across queries can be increased by sorting the image sets for neighboring shots by color similarity. Mehtre et al. [MKNM95] indicate that it is sufficient to perform a coarse comparison to exclude severe color missmatches in image retrieval. We therefore compute for each image its mean RGB color vector. Given the current representative image I_A for shot A, we select for a neighboring query B the representative image I_B that minimizes the Euclidean distance of the mean color vectors. For the next query C, the comparison is carried out with respect to I_B,

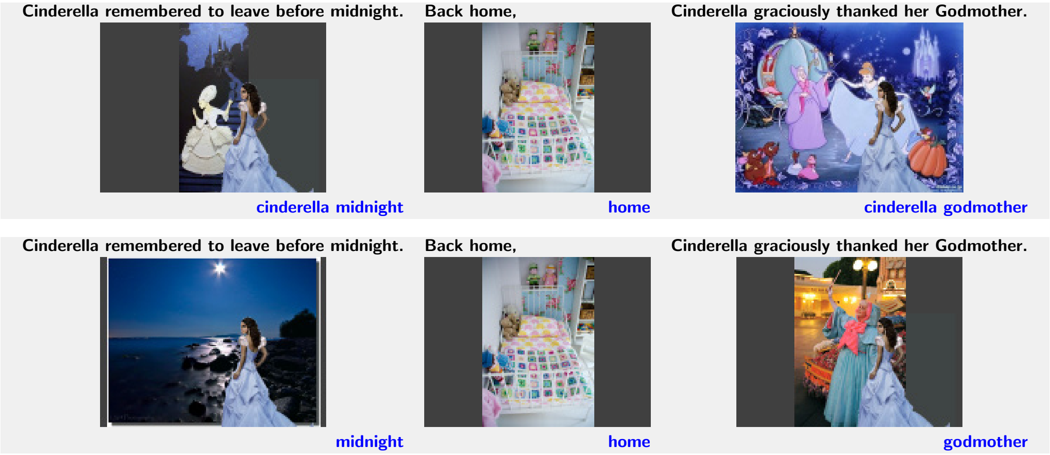

Cinderella remembered to leave before midnight.
Back home,
Cinderella graciously thanked her Godmother.
cinderella midnight
home
cinderella godmother
Cinderella remembered to leave before midnight.
Back home,
Cinderella graciously thanked her Godmother.
midnight
home
godmother

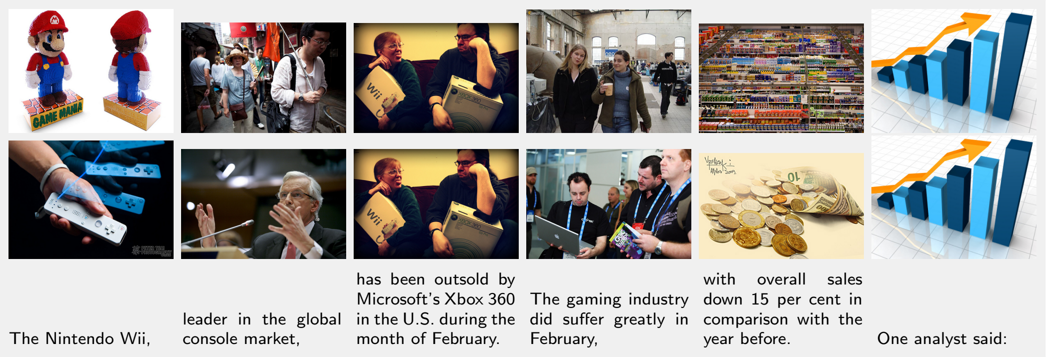
Figure 6.26: Additional layer for protagonist handling. Ignoring recurrence of the main character leads to different appearances of the individual (top row, background layers). Adding an additional protagonist layer and removing the individual from the POT textual query leads to a similar appearance of the protagonist within different environments (bottom row). (Images from [Sch10], Fig. 5.2 b) & c)).

and so on. The user is free to indicate whether or not the color matching constraint should be applied and, if so, into which direction the color should be propagated. The improved results are shown in Figure 6.25(b). Alternatively, color consistency can be achieved by processing the selected images (e.g., converting them to sepia or black-and-white representations, applying color style transfer techniques [NN05]).

Main Character Handling. Often, multiple text parts describe an individual within different settings, e.g., varying environmental situations. However, it is nearly impossible to retrieve images with the same character in such different situations. Thus, in an extension to the TTV pipeline, a recurring main character can be handled separately [Sch10]. Thereby, if existent, the name of the protagonist is found automatically by simply selecting the most frequently occurring noun in the story. A set of images is then queried for the protagonist, presented to the user for selection and cutting. As shown in Figure 6.26, the resulting protagonist image is then treated as an additional layer on top of the shot by the TTV system. Additionally, to avoid duplicate appearances of the individual, the protagonist is removed from the textual queries for the POT. Overall, this method leads to a coherent appearance of the protagonist within different environments.

Table 6.4: Results on different text types. For each of the submitted queries, the precision for the first (column 2) and the first 10 (column 3) retrieved images is evaluated.

Text type	Precision #1	Precision #10	#Words	#Queries	#User interactions	Query time
Nursery rhyme	40%	46%	53	8	5	00:02:33
Fairy tale	80%	60%	538	73	45	00:33:18
Screen play	80%	55%	230	41	28	00:20:59
News	60%	49%	147	19	12	00:12:02
Short novel	40%	38%	967	150	113	01:27:08



The Nintendo Wii,

leader in the global console market,

has been outsold by Microsoft's Xbox 360 in the U.S. during the month of February.

The gaming industry did suffer greatly in February,

with overall sales down 15 per cent in comparison with the year before.

One analyst said:

Figure 6.27: Automatically generated visual story (top row) and manually improved version (middle row) of a news article about the Wii controller (bottom row).

6.2.6 Results

At this stage, each POT has been assigned a shot. The resulting image set can simply be represented as a storyboard (see Figures 6.18, 6.27) or presented in the form of a video as an animated slide show, where text and image transitions are achieved by a constant motion.

We have applied our system on various text types: the nursery rhyme "Three Blind Mice", the fairy tale "Cinderella", a part of the screenplay to "Braveheart", a news article about the Wii controller, and the short novel "Animal Farm". In general, we observed high context-sensitive precision in our tests, considering the actual meaning of the sentence rather than just the queried tokens. This can be seen in Table 6.4, which shows our results for a range of text types. For each query, 30 images were downloaded and the context-sensitive precision for the first and the first 10 images was evaluated. Because of the very good sentence related precision of the first reported images, the required user interaction to construct a semantically close storyboard is moderate.

We think that the variation in the text types correlates with the typical spectrum of submitted Flickr images. The content of the selected nursery rhyme and short

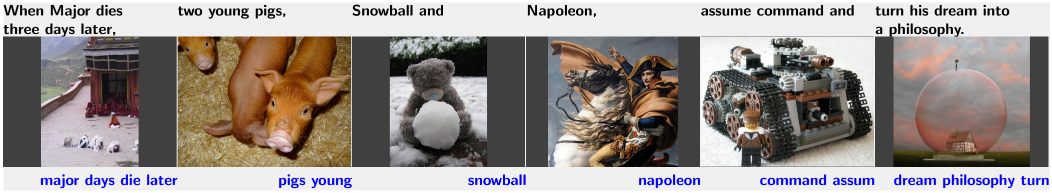


Figure 6.28: *Semantically mismatching images due to metaphorical character names in the "Animal Farm" novel.*

novel is slightly more abstract than the content of the other categories. However, in Section 6.3, we will present a framework that is capable to automatically illustrate even highly abstract creative texts.

In Fig. 6.27, we compare completely automatically generated results against the manually optimized selection. While quite a number of shots received a decent representative automatically, a few clicks were necessary to obtain the final selection where semantic errors and deviations in style have been removed.

In general, we were surprised by how often the retrieved images for our generated queries match the intention of the original text. However, Figure 6.28 clearly shows the limits of our approach, namely dealing with word sense ambiguities. The images retrieved for the two pig characters *Snowball* and *Napoleon* from the novel "Animal Farm" do not depict pigs, but the literal or most frequent meaning of the words. A solution for this problem could be the usage of sophisticated lexical semantic analysis, such as named entity recognition (Section 3.4.1). Besides, our parsing is limited to group only tokens that are adjacent in a sentence. Dependency graphs (Section 3.2.2) could be employed to assemble queries for the non-connected parts of a sentence.

6.2.7 Conclusion

The system presented in this section can be seen as one of the first steps towards creating a movie based on a textual input. After parsing the text, the system automatically generates queries and retrieves images from the community site Flickr. Most often, a set of representative images is found automatically due to our hierarchical querying approach. This hierarchical algorithm is capable of retrieving Flickr images with high precision to given text snippets, which was also indicated by our initial observation $\mathcal{H}_{TTV}$. Finally, after a few user interactions, a reasonable storyboard is produced. Thereby, initial means were presented to support the user in the selection of, e.g., similar appearing images between neighboring text parts based on their RGB color information, or same images for recurring text parts. Overall, we demonstrate promising initial results on several types of texts.

By enhancing the semantic analysis of the text, e.g., by using WordNet or by considering syntactic as well as semantic relations between objects, the quality of the (semi-)automatically generated storyboards might be further increased. Additionally, an automatic classification of the retrieved images into sets of similar images might facilitate the manual selection process. One natural extension to the presented system is to animate the retrieved images to better visualize the action in a story – as a next step towards creating full-fledged movies.

Overall, the focus of the presented illustration pipeline lies in assembling semantically relevant image sets to a given text snippet. The final selection is left to the user whereby some automatic means provide support in this process. The next section builds upon this idea of illustrating text with images from online collections but extends the approach to the automatic illustration of complete storylines. Thereby, semantic relevance and, in addition, visual coherence will be considered to generate a meaningful picture story along the input text.

6.3 Auto – Illustrating Creative Text with Style

The semi-automatic illustration pipeline presented in the previous section focused on assembling relevant image sets to a given text snippet. Initial automatic means support the user in the final selection of, e.g., similar appearing images between neighboring text parts. Aiming for a similar goal, this section proposes a framework that also illustrates given text with relevant images. However, in addition to semantic relevance, we now focus on ensuring visual coherence along the entire storyline. As shown in Chapter 5, the visual appearance plays a key role in how pictured content is transported to an observer. Thus, to create a visually coherent and meaningful picture story, we build on these findings and present an approach to automatically generate meaningful illustrations in different visual styles.

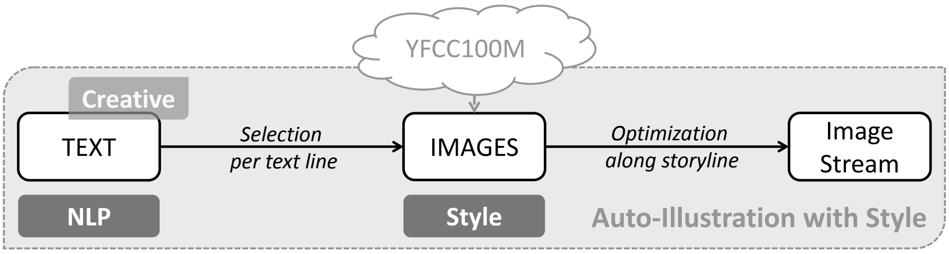


Figure 6.29: *High-level concept. Information is extracted from a text and relevant images in a given style are assembled for each text line from a large image collection. Optimization along the storyline results in a semantically relevant and visually coherent image stream.*

Figure 6.29 shows the high-level concept of our optimization based framework to automatically produce both semantically relevant and visually coherent illustrations for poems and songs. Based on such creative texts, a linguistic analysis is performed to extract suitable textual information and images in a given style are assembled for each text line from a large collection of annotated images (YFCC100M [Tho14]). Optimization along the storyline then results in a semantically relevant and visually coherent picture story depicting a particular visual style. We hypthesize that:

> $\mathcal{H}_{ST}$: *Illustrating text with a specific style strengthens the visual coherency of a semantically meaningful illustration.*

We evaluate our hypothesis $\mathcal{H}_{ST}$ and variations on our optimization procedure in a user study (Section 6.3.6) and show that style is the strongest of our incorporated visual features to ensure visual coherence. We demonstrate our method on a selection of 200 popular poems and songs collected from the internet and operate on around 14M Flickr images (Section 6.3.4). Finally, we present two applications, identifying textual style, and automatic music video generation (Section 6.3.8). The work presented in this section is based on the following publication: [SBL17].

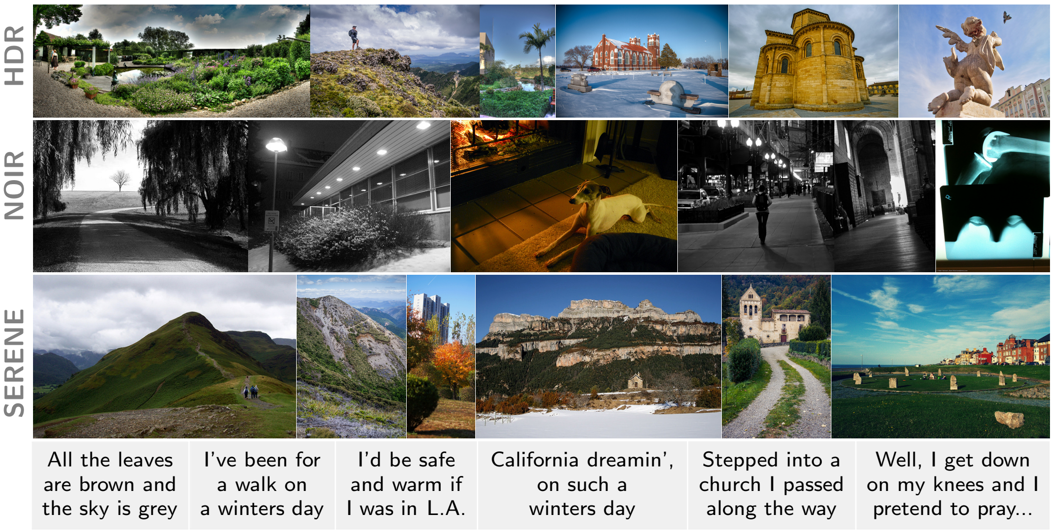

HDR
NOIR
SERENE
All the leaves are brown and the sky is grey
I've been for a walk on a winters day
I'd be safe and warm if I was in L.A.
California dreamin', on such a winters day
Stepped into a church I passed along the way
Well, I get down on my knees and I pretend to pray...

Figure 6.30: *Automatically generated illustrations of the first text lines (bottom) of the song "The Mamas The Papas - CALIFORNIA DREAMIN" in three different styles (left).*

6.3.1 Introduction

When an artist creates a poem or song they weave their story carefully, selecting words that produce a vivid visual story in our minds. In this section, we explore the goal of automatically illustrating such creative pieces of art with images. As artist's compositions are intended to be highly emotional and lyrical, we aim to select images that are aesthetically pleasing and highly stylistic according to the style of the artwork, e.g., we might illustrate a poem about love with images predicted to be "romantic" but a heavy metal song in "horror" style. Although these kinds of texts are quite challenging due to their high level of abstraction, they often display beautiful language, lending themselves well to our goal of auto-illustration with style.

Additionally, the nice repetitive structure of such creative texts perfectly suits our approach. They typically consist of a sequence of text lines whereby each line consists of textual information that semantically belongs together and often transports a specific mental image as part of the entire story. Thus, we exploit this underlying structure which claims to be illustrated by a meaningful picture per line of text and, compared to our previous illustration approach, prevents the need for further text segmentation (Section 6.2.3).

Furthermore, as such texts often deal with a certain theme, we incorporate the global idea as well as allowing for visual adaptions to content changes. Some example outputs of our pipeline for different styles are shown in Figure 6.30. The resulting sequences can easily be synchronized with a song to generate a music video.

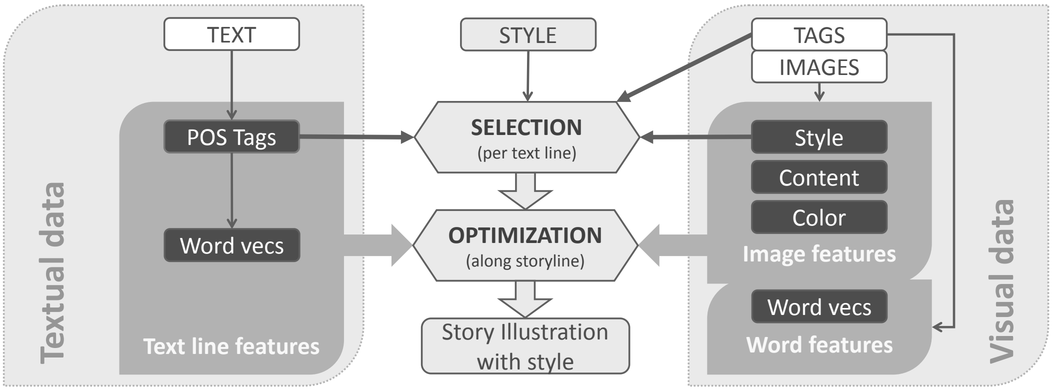


Figure 6.31: Overview. Given a large collection of annotated images, we automatically illustrate texts in two steps. First, for each text line, linguistically relevant images are selected that match important words and depict a certain style. Then, based on this collection, we optimize along the storyline to match style and coherence between selected photos.

Auto-Illustration Overview. An overview of our processing pipeline is illustrated in Figure 6.31. Given a large collection of annotated internet photos, we would like to auto-illustrate poems and songs with images reflecting a user-specified style. These illustrations should tell a visual story of the text, matching both the specified image style and demonstrating visual coherence between selected images. This is achieved by first selecting images from the collection that match important words from the input story, depicting the specified style. Suitable candidate images are selected in this manner for each text line exploiting the underlying structure of creative texts. Compared to our previous illustration approach (Section 6.2) where the focus was to directly query for the most precisely matching images for a particular text snippet, we now initially select a broader set of pictures for each text line to obtain enough image candidates for further processing. Next, a coherent image sequence is generated to illustrate the story by optimizing style scores and consistency between successive images along the text lines. Consistency is measured using a combination of textual and visual coherence scores related to image content and color.

In order to select good images for illustration, we extract a number of visual and textual features for comparing text to images (Section 6.3.3). In particular, image tags are used in combination with parsed words and word2vec embedding representations [MSC+13] (Section 3.3.3) to better match not only the syntactic, but also the semantic meaning of images and text. To encourage semantic similarity of image content between subsequent images, the response vector of a deep neural network pre-trained for image classification is utilized, judging content similarity between two images as the distance between the corresponding representations. Finally, the style of each input picture is predicted based on the work of Karayev et al. [KTH+14] which will be outlined in Section 6.3.2. We use the predictions for 20 different image styles to tell a picture story in a particular illustration style. Each of these criteria can now be used to both assemble a selection of candidate images per text line and then to optimize for consistency along the story using global discrete energy optimization (Section 6.3.5).

The novel combination of considering textual semantic search, content similarity, style classification, and discrete optimization allows us to generate picture story illustrations with controllable style, even for challenging abstract text types such as poems and song lyrics.

6.3.2 Related Work on Visual Style

More recently, studying the visual style of images and 3D scenes became a popular research direction. Previous work investigating in stylistic elements of 3D objects has so far focused on furniture. However, making use of deep learning, special interest has arisen in manipulating images by transferring artistic style from a painting to a photograph or recognizing a particular image style. Overall, visual style comprises color concepts (Section 5.3.2) as well as aesthetic aspects of appeal (Section 5.2.2). However, style combines and extends those concepts (with features like texture, composition, or geometric aspects) leading to a high diversity of various combinations that distinguish different visual styles.

3D Style of Furniture. Approaches investigating in the style of 3D scenes have mainly focused on furniture [LHLF15, LKWS16, HLK^{+}17]. Liu et al. [LHLF15] explore the compatibility between pieces of furniture from a stylistic point of view. They have proposed to learn a metric based on geometric features of the different objects to predict the compatibility between them allowing for stylistic coherent arrangements of furniture in a scene. Hu et al. [HLK^{+}17] aim to identify and group style-defining elements for furniture models to distinguish between certain styles and characterize high-level descriptions like "European". In order to transfer the geometric style of furniture items to other models in a 3D scene, Lun et al. [LKWS16] have proposed a method which allows users to specify a style by an exemplar shape and have focused on preserving the functionality of the target objects when transferring the style from the initial item to other furniture objects. However, we focus on the style of pictures rather than 3D objects.

Image Style Transfer. Generally, transferring a visual concept from one image to another has already been of high interest in research for some time. As described in Section 5.3.2, previous work on example-based color transfer has focused on transferring the color concept from a given reference picture to another visual source. However, meanwhile, deep learning methods such as convolutional neural networks even allow to transfer an artistic style from a painting to a captured

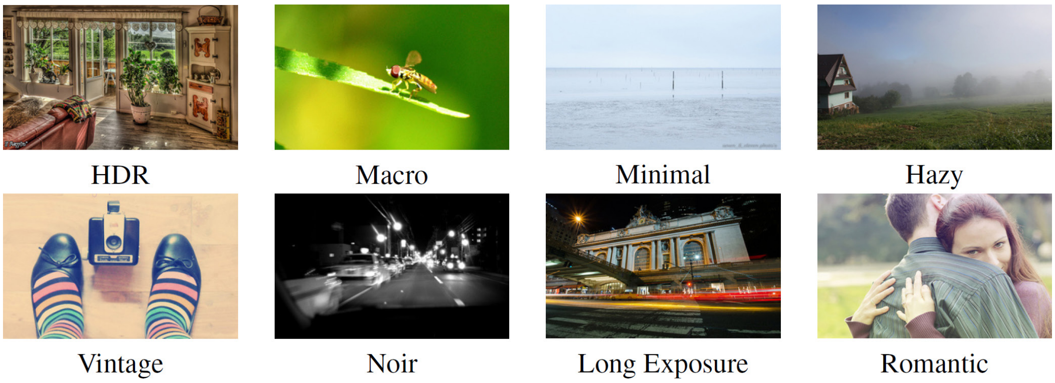


Figure 6.32: *Example photographs with style labels of the 20 style classes recognized on 80K Flickr images by Karayev et al. (Examples from [KTH+14], Fig. 1, left part).*

photo [GEB16a, JAFF16]. Thereby, visual style extends the color concept by other features like texture or composition. "Neural Style Transfer", a line of work which has been introduced by Gatys et al. [GEB15, GEB16a, GEB+16b] has demonstrated exciting results in creating fascinating images. They give photographs an artistic style which they typically derive from artworks, in particular, pieces of art that have been painted by famous artists like Vincent Van Gogh or Pablo Picasso. A key idea of their approach lies in applying the style of one image to the content of another one. As the representations of content and style are separable in the CNN, both can be manipulated independently of each other allowing to synthesize new images [GEB16a]. To model content and style features in a CNN, they make use of the 19-layer VGG network [SZ14]. Further, Gatys et al. [GEB+16b] improve the work on neural style transfer and differentiate between spatial, color, and scale control. This allows them to reduce some previous failure cases like modifying the sky with a ground texture. Additionally, their factorization into the mentioned aspects enables combining those aspects from different images and derive new interesting styles. Following this line of work, Johnson et al. [JAFF16] have focused on the transformation task, i.e., when an input image is manipulated and changed into a differently appearing output image and have managed to increase performance while maintaining quality.

Style Recognition. A method recognizing particular visual styles in paintings or images has been introduced by Karayev et al. [KTH+14]. Compared to previous work that identifies particular color concepts for transfer (Section 5.3.2) or aesthetic appeal in images (Section 5.2.2), they aim to explore a broader range of visual styles. They consider painting styles like "cubism" or "expressionism", but also study a variety of other types, e.g., *genres* like "noir" or "vintage", *atmospheres* like "hazy" or "sunny", *moods* like "melancholy" or "serene", *optical techniques* like "macro" or "long exposure", or even *composition styles* like "minimal" or "geometric". In order to explore styles like the latter ones, they have collected data from Flickr Groups, e.g., "Geometry Beauty" or the "Film Noir Mood" group. To classify painting style, they assemble a data set from WikiPaintings[2]. They evaluate several image features comprising color, composition, content, and low-level statistics. Finally, their approach predicts style in an image using features from a pre-trained multi-layer deep network [KSH12], fine-tuned to predict image style on 80K Flickr photographs depicting 20 different styles as well as 85K paintings labeled with 25 art styles. Figure 6.32 shows examples for some styles their method recognized on Flickr photos. We utilize the model they trained on such Flickr images.

Our work. Similar to previous auto-illustration approaches (Section 4.2.3) and our previously presented text illustration method (Section 6.2), we also retrieve images that match an input text for our auto-illustration method we present in this section. However, we also optimize for two additional storyline features:

1. We select images for illustration according to a particular story style.
2. We attempt to select a visually coherent set of images for illustration.

In order to illustrate stories according to a particular style, we rely on previous work for style recognition in images, in particular, the approach introduced by Karayev et al. [KTH+14] and described above. We use the model they fine-tuned on 80K Flickr photographs predicting 20 different styles on images.

6.3.3 Feature extraction

To support our algorithm that selects images to illustrate text, we extract a set of features from images and their associated tags as well as from lines of text. The features allow matching both content and style between individual lines of text and images, and between pairs of images selected to illustrate a sequence of text lines.

Text Features

The semantic mapping between the given text and the images is established by matching several textual features. The language features are extracted by applying **parsing** to determine potentially relevant words from the input text while **image tags** are used directly. However, **word-vector representations** are generated for both, extracted words and image tags.

[2] www.wikiart.org

Parsing. A line l of text T is first analyzed by tokenizing and parsing to determine part-of-speech (POS) labels for each word [BKL09] and a lemmatizer based on WordNet [Fel98, BKL09] improves performance. Details on these basic NLP techniques are provided in Section 3.2. Then, the extracted *nouns*, *verbs*, *adjectives*, and *adverbs* are gathered in a set $\tau(l)$. POS parsing enables us to select the most relevant words for each matching task. For candidate retrieval, we consider the subset $\tau_{NV}(l) = \{w_1 ... w_A\} \subseteq \tau(l)$ of extracted nouns w_N and verbs w_V to obtain a broad and large enough image set. Later, the word-vector representations for text lines are computed based on the entire set, $\tau(l)$, to capture additional meaning.

Image Tags. All images considered for illustration have user associated tags. Thus, for each image I we store its associated list of tags $\kappa_I = \{w_1 ... w_K\}$ in an inverted file table (see Section 3.4.2), making it efficient to access all images with tags matching words $\tau_{NV}(l)$ from a text line l.

Word-vector Representation. In addition to directly matching between words, we exploit recent work that maps words to vectors (word2vec) and has been introduced in Section 3.3.3. Based on a continuous skip-gram model this approach provides a mapping of phrases into a 300d vector space [MCCD13, MSC$^+$13]. As mentioned, the mapping comprises a large number of syntactic and semantic word relationships while compressing semantic similarity. For a single word w, we obtain its word2vec representation $V(w)$ and, for a set of words, we average the vector representations of all words in the set. We calculate average word vectors for tags of a text line $\tau(l) = \tau_l$ as V_{τ_l} and for image tags κ_I as V_{κ_I}.

Image Features

Features are extracted from the images for identifying **style**, **content**, and for ensuring **color** consistency between selected images.

Style Feature. Estimates for 20 style-classes are extracted using the previously described method of Karayev et al. [KTH$^+$14] which classifies image style using a convolutional-neural-network approach (Section 6.3.2). This estimate is used to consider only images that match the specified style, and to ensure consistency between images selected for illustration. We assume that an image I matches a certain style, sty, if its prediction score for this style is greater than 0.5 and define the style constraint as $I_{sty} > 50\%$.

Image Content Feature. To compare the visual content of two images we make use of deep learning results, in particular, the VGG 16-layer model [SZ14]. As

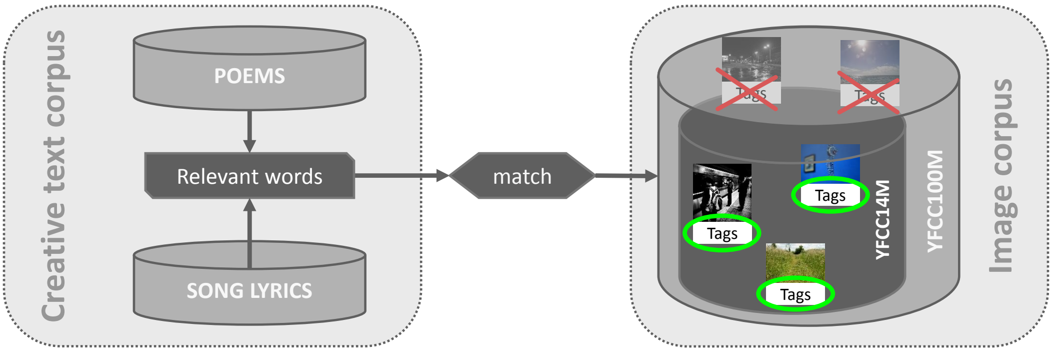


Figure 6.33: *Corpora overview. A list of representative words is assembled from the creative texts to match relevant words with image tags in the YFCC100M data set and assemble a suitable subset of images for our image corpus.*

indicated in Section 4.1.3, this deep convolutional network has demonstrated high accuracy on image classification. We use a pre-trained model which has been trained to recognize 1000 classes from the ImageNet Challenge. For our image content representation we use the 4096d features extracted from the first fully-connected network layer.

Color Feature. As color significantly influences our visual impressions of images and is not well represented using pre-trained CNN features, we incorporate a simple RGB color histogram feature for evaluating image similarity. We extract 256 bins per color channel for each image.

6.3.4 Corpora

To demonstrate our approach, we make use of publicly available data. We crawl a set of 200 famous poems and song lyrics for our creative text corpus, and make use of the YFCC100M Yahoo Flickr dataset for our image corpus. An overview of creating the corpora is given in Figure 6.33. Compared to our previous illustration approach (Section 6.2) which directly queries the Internet, assembling such large corpora now allows us to preprocess the data and extract the mentioned textual and visual features on a large scale.

Creative Text Corpus. We assemble a set of creative texts, namely songs and poems to demonstrate our approach. Therefore, we obtained lyrics to 100 songs from *SongLyrics.com* [3]. This website provides lyrics to top songs for every year from 1950 to 2011. In order to cover different music styles, we crawled the lyrics from

[3] `http://www.songlyrics.com`

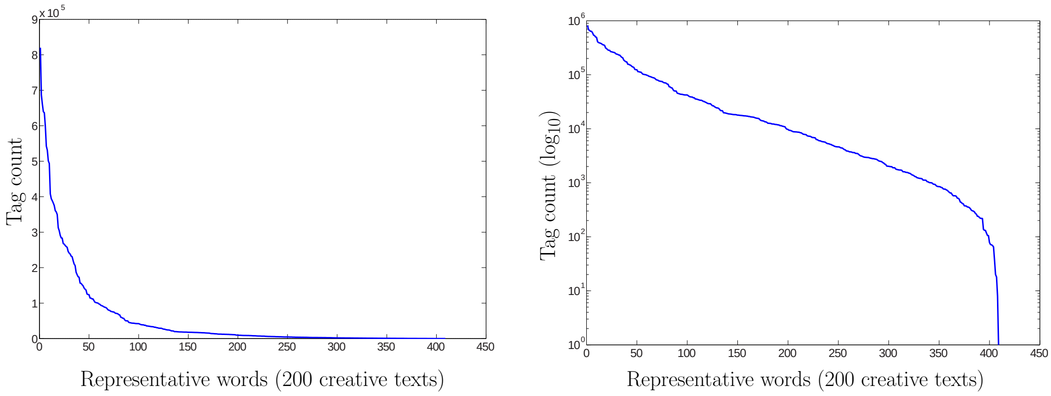


Figure 6.34: *Frequency count of representative word list in YFCC100M photo tags. Left: Frequency of a word found in tags of the photos corresponds to the number of retrieved image results by querying the word. Right: Log-plot of distribution. The representative words $w_{\mathcal{R}}$ and the corresponding frequency counts f_{tags} are listed in Appendix A.*

the "Top 100 Songs of All Time" category. For poems, the *Best Poems Encyclopedia* [4] website provides best poem texts for various categories, e.g. "life" or "love". To retrieve a broad set across categories, we downloaded the poems in their "Top 100 Poems" category which covers famous poems of all time.

Image Corpus. For generating illustrations, we make use of the "Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M)" [Tho14] recently published by Thomee et al. [TSF+15]. It contains about 100 million Flickr images and videos with associated meta-data and which are licensed under creative commons. We consider the image portion, around 70% of which have associated textual tags. We select images that are potentially relevant to our text corpus by filtering out images that are not tagged with relevant words, i.e., nouns and verbs appearing in our creative text corpus. Text pre-processing is used to lemmatize, remove stop words (Section 3.2.1), and select words that occur in both the poems and song lyrics, leaving us with a representative set of over 400 words $w_{\mathcal{R}}$. Figure 6.34 shows the distribution of the number of found images in the YFCC100M set for all words of the representative word list indicating that several words occur rather often whereas the larger part occurs less frequently. However, the frequency f_{tags} of the resulting words mentioned in the tags reached up to 820079 for the most frequent word "music". Besides, for around 60 words more than 100000 images do match. The complete list of relevant words and the corresponding frequency counts are provided in Appendix A. As illustrated in Figure 6.33, only Flickr images whose tags match at least one of the words in this representative list are selected, leaving us with a subset of 14 million images. Table 6.5 indicates the amount in our YFCC14M

[4] http://100.best-poems.net

Table 6.5: *Style in YFCC14M. For all 20 provided styles, the table indicates the amount of images within the YFCC14M subset with $I_{sty} > 50\%$ for a certain style. "Detailed" contains the highest amount and half of the styles hold at least more than 300K images.*

Style	> 50%	Style	> 50%	Style	> 50%	Style	> 50%
Detailed	674K	Hazy	335K	HDR	296K	Texture	219K
Bright	519K	Geom. Comp.	334K	Romantic	259K	Sunny	201K
Horror	498K	Bokeh	327K	Melancholy	234K	Pastel	184K
Depth of field	408K	Serene	324K	Long Exposure	227K	Macro	135K
Noir	352K	Minimal	303K	Vintage	224K	Ethereal	87K

subset with a prediction greater than 50% for each style, typically yielding several hundred thousand images that are likely to depict that style.

6.3.5 Selection and Optimization

Given the feature vectors computed for each database image and the similarity measures defined in Sec. 6.3.3, the selection process for auto-illustration first computes a suitable candidate set of images matching to each text line. Then, based on those pre-selected candidate image lists, an optimization step estimates the best sequence of images for illustration, maximizing text and style match scores, as well as cohesion in content, color, and style along the illustration.

Selecting Candidate Images

For every line in a text, a set of candidate images is selected that semantically matches the text line and corresponds to the specified illustration style. Thereby, instead of querying for a rather small set of the most precisely matching images (Section 6.2.4), we now first select a broader set of pictures for each text line to obtain enough image candidates for the optimization over multiple priors.

Specifically, given the POS analysis, candidate selection is performed for each text line l by including each image $I \in$ YFCC14M into the set of candidate images I_{cand_l}, if the following condition is fulfilled:

$$(\kappa_I \cap \tau_{NV}(l) \neq \emptyset) \land (I_{sty} > 50\%) \Rightarrow (I \in I_{cand_l}) \tag{6.2}$$

This means we select all images I whose tags κ_I match at least one noun or verb present in the text line $\tau_{NV}(l)$ and, at the same time, depict the requested illustration style with a style prediction score greater than 50%. A higher threshold might thin out the candidate list too drastically. Sorting this list according to style score, the top 1000 candidate images per text line form a suitable basis for our optimization phase. Figure 6.42 shows example text lines accompanied by their top candidate images constrained by style.

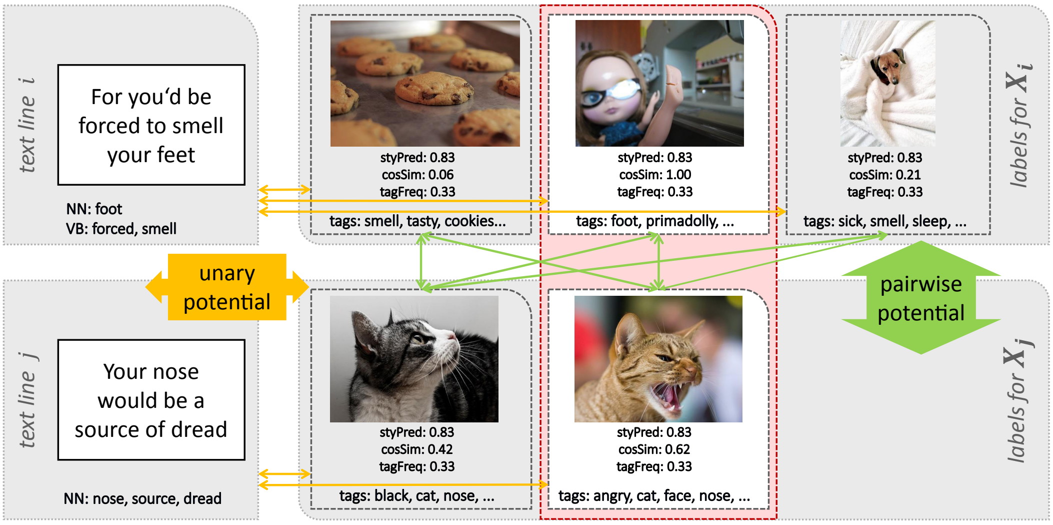


Figure 6.35: *Optimization structure for two successive text lines. The images in the middle (red framed) are finally picked. ("J. Prelutsky - BE GLAD YOUR NOSE YOUR FACE","depth of field")*

Storyline Optimization

Given the candidate image sets for each text line, we would like to select a final set of images to illustrate our story that are: 1) semantically relevant to the story, 2) good representatives of the selected style, and 3) visually coherent along the story line. Figure 6.35 visualizes the underlying structure for two successive text parts, each connected with a small subset of potential images for selection. The task of choosing the best image from each subset, while preserving semantic relatedness and visual consistency can be described as a discrete optimization problem with pairwise variables and formulated as an energy with unary and pairwise terms.

Thus, from an assembled and presorted set of images per text line, the goal is to select one image for each position, X_i, of the corresponding text line $i \in \nu$. We formulate this as minimization of an energy function E with image labels for a text line i along all lines ν (nodes) and over consecutive text pairs ε (edges):

$$E(X) = \sum_{i \in \nu} U(X_i) + \sum_{i,j \in \varepsilon} P(X_i, X_j) \tag{6.3}$$

The unary potential function U measures semantic and stylistic relatedness between a text line and a potential image for illustrating that line. The pairwise consistency terms P describe the interaction potential between pairs of images.

Unary terms: Two types of unary terms measure semantic (text) relatedness between the text lines and images, s_{freq} and s_{sem}, combined in a weighted sum (Eq. 6.4). In order to turn the similarity into a cost for the minimization, we calculate $1 - \sum weightedUnaryTerms$.

$$U(X_i) = 1 - (\lambda_1 s_{freq}(x_i) + \lambda_2 s_{sem}(x_i)) \tag{6.4}$$

- **Tag frequency.** s_{freq} computes the overlap between all nouns and verbs extracted from the text line and the tags associated with an image: $s_{freq} = \frac{1}{A}\sum_{a \in A} \xi_a$, with $\xi_a = \begin{cases} 1, & w_a \in \tau_{NV} \text{ occurs in } \kappa_I, \forall a \in A \\ 0, & else \end{cases}$
- **Semantic.** s_{sem} calculates the word2vec similarity between the text line words τ and the image tags κ using cosine similarity between the average representation vectors V_τ and V_κ as $s_{sem} = \text{cossim}(V_\tau, V_\kappa) = \frac{V_\tau \cdot V_\kappa}{||V_\tau|| ||V_\kappa||}$ (Section 3.3.3). This allows for similarity comparisons beyond exact word matches.

Pairwise terms: Between all possible candidate image pairs for each successive text line, a pairwise energy term is computed that is minimized to obtain a globally consistent illustration such that the storyline flows smoothly along the illustration. This pairwise potential is defined as a weighted sum (Eq. 6.5) of three types of consistency: style d_{sty}, color d_{col}, and content d_{cont}.

$$P(X_i, X_j) = \mu_1 d_{sty}(x_i, x_j) + \mu_2 d_{cont}(x_i, x_j) + \mu_3 d_{col}(x_i, x_j) \tag{6.5}$$

- **Style.** d_{sty} computes image to image coherence as the normalized Euclidean distance between the 20d style vectors of successive candidate pairs.
- **Content.** d_{cont} is obtained by the Euclidean distance between the l2 normalized CNN feature activation vectors to encourage smoothness between what successive images in our illustration depict (Figures 6.41, 6.38).
- **Color.** d_{col} is calculated as the Euclidean distance between RGB color histograms computed between successive pairs of candidate images.

To minimize E, an NP-hard problem, we use the "sequential tree-reweighted message passing algorithm" (TRW-S) proposed by Kolmogorov [Kol06] whose main property is that the value of the bound is guaranteed not to decrease and, thus, at least a "local" maximum of the bound is retrieved. The weights of the parameter sets $w_U = (\lambda_1 | \lambda_2)$, $w_P = (\mu_1 | \mu_2 | \mu_3)$ will be discussed in the following section.

6.3.6 Human Evaluation

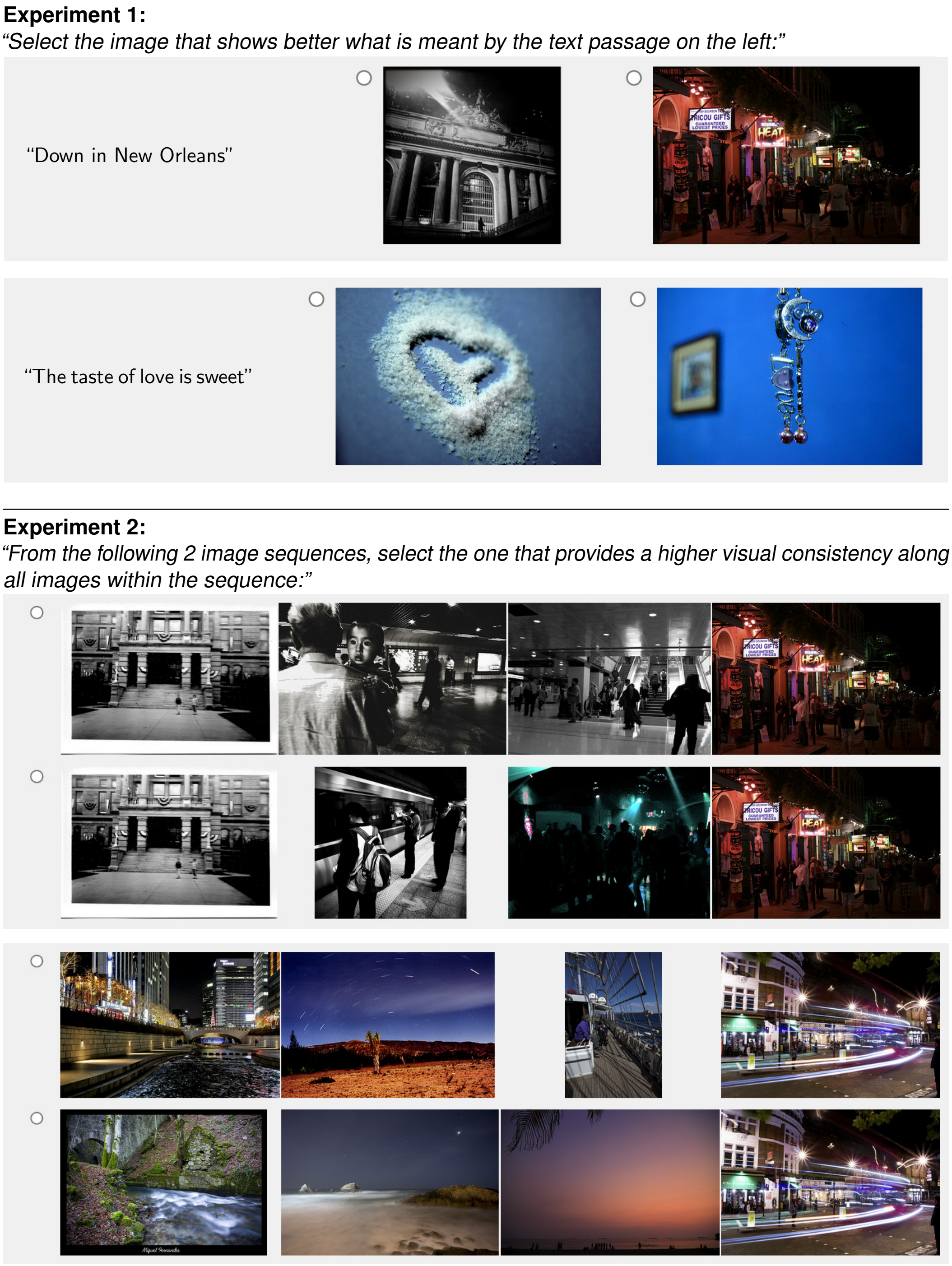
Aligning a story with appropriate images in a pleasant style is a subjective task, especially for abstract texts such as poems and songs. Thus, in order to obtain more general ratings from a wide variety of people, we performed experiments on Amazon Mechanical Turk (AMT) to measure the quality of our method (Section 2.3.1). First, we tear apart the relative contributions from various pieces of our system. We perform an experiment on the semantic connection between text and images, the unary terms, and evaluate the pairwise terms which regulate visual consistency along image sequences. Finally, we validate the quality of resulting illustrations.

Experiment Data Set

Our experiments are designed based on the assembled data described in Section 6.3.4, consisting of about 200 creative texts. In total, 20 styles can be used for illustration. We randomly select 110 text-style combinations consisting of half poems and half song lyrics and ensure that every style is represented. Due to the style constraint, some text lines may result in only a few or even no image responses for a requested style. Thus, we only use text lines l with enough image candidates ($\#I_{cand_l} > 1000$) to optimize over 1000 image labels and at least one word present in the pre-trained Google word2vec representation ($\exists V_{\tau_l}$), ensuring that we can perform a proper parameter set evaluation. Finally, to evaluate visual consistency along an image sequence, the number of image responses for each of the succeeding text lines should also be large enough. Thus, we only accept consecutive text parts T_q with M succeeding lines $T_q = \{l_{q_1}...l_{q_M}\}$ such that all consecutive lines l_{q_m} fulfill the word2vec and candidate set requirements.

User Study

As already mentioned, people may have different internal rating systems, especially for subjective tasks. Thus, to measure relative contributions, we formulate our first experiments (Exp. 1, Exp. 2) as binary forced-choice tasks. Each pairwise preference test is designed as a data-pair selected by two parameter sets controlling different portions of the features contributing to the optimization and is presented to 5 Turkers. Depending on the type of study, data within a pair either consists of two images compared to a text line or two image sequences for evaluating visual consistency. Examples for both types of experiments are shown in Figure 6.36 and results are listed in Table 6.6. Randomized ordering and positioning are used to negate click biases. Tasks resulting in rather unclear (2–3)-decisions out of 5 Turkers are filtered out afterwards as they are not suitable to detect trends. Based on the derived best parameter settings, we let Turkers rate the quality of final illustrations along the according text streams in a third experiment (Exp. 3) on a 5pt Likert scale. Examples for the third experiment are shown in Figure 6.37.



Experiment 1:
"Select the image that shows better what is meant by the text passage on the left:"

"Down in New Orleans" ○ ○

"The taste of love is sweet" ○ ○

Experiment 2:
"From the following 2 image sequences, select the one that provides a higher visual consistency along all images within the sequence:"

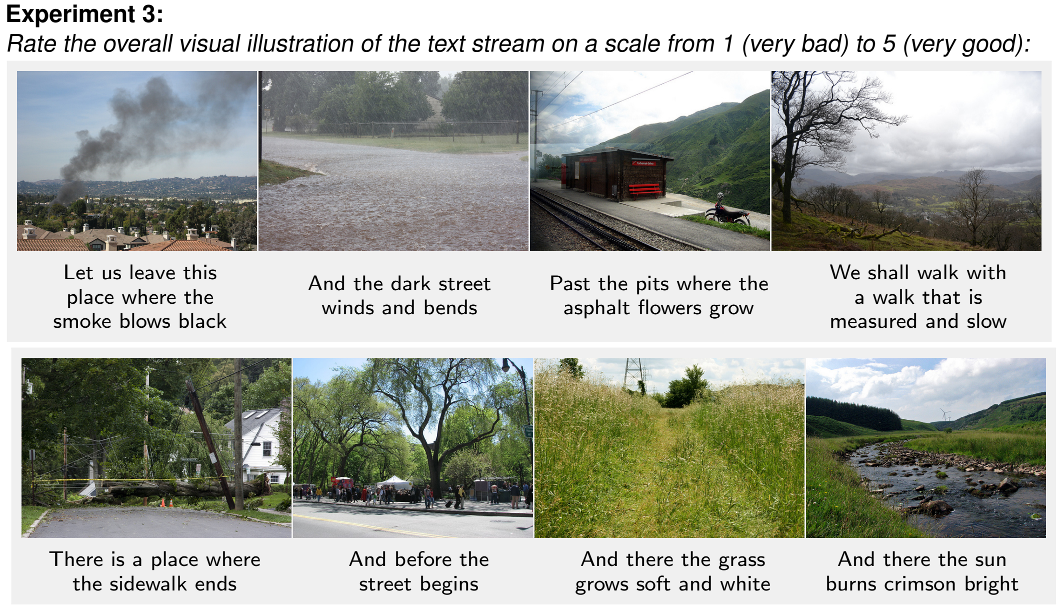
○

○

○

○

Figure 6.36: *Examples from AMT user studies for pairwise preference tests. Exp. 1 (top): Evaluation of semantic text-to-image relation. Exp. 2 (bottom): Two example pairs of image sequences to evaluate visual consistency along the image streams. Depending on the experiment, the task is to select the preferred image (top) or image sequence (bottom).*



Experiment 3:

Rate the overall visual illustration of the text stream on a scale from 1 (very bad) to 5 (very good):

Let us leave this place where the smoke blows black

And the dark street winds and bends

Past the pits where the asphalt flowers grow

We shall walk with a walk that is measured and slow

There is a place where the sidewalk ends

And before the street begins

And there the grass grows soft and white

And there the sun burns crimson bright

Figure 6.37: *Two examples from AMT user studies to evaluate the quality of final text stream illustrations (Exp. 3). The task (top) is to rate each sequence on a 5pt Likert scale.*

Experiment 1: Text-to-image semantic. Our first experiment evaluates the semantic connection between text and images, represented by the unary terms in the optimization. We formulate our hypotheses $H_{T\leftrightarrow I}$ as:

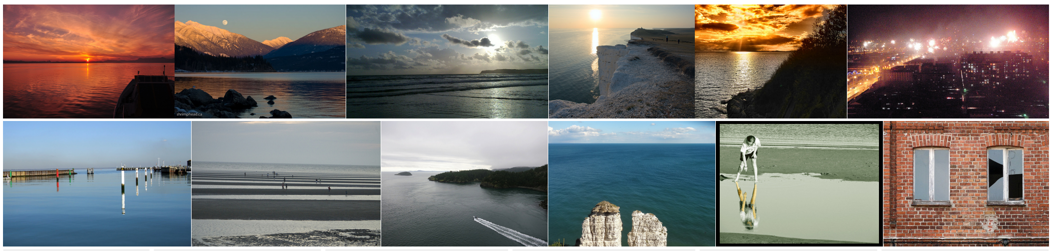
$H_{T\leftrightarrow I}$:	• Both, word vectors and tag frequency are relevant in the unary terms (tagFreq > 0, wordVec > 0). • The positive influence of the word vectors is higher than of the tag frequency (wordVec >= tagFreq).

The requirements described in Section 6.3.6 result in around 2110 text–image pairs. We compare binary contributions of the unary features, e.g. only wordVec $w_U = (0|1)$ against all in (1|1). Figure 6.36 (top) shows some examples of the tasks we gave to the Turkers consisting of a short text line and 2 images. Overall, including only the wordVecs has been preferred over only tagFreq (Table 6.6). Combining tagFreq and wordVecs has been selected over using only the one or the other feature with a 67% preference indicating that both terms are needed.

Experiment 2: Consistency along storyline. Our second experiment focuses on the visual coherence between successive images. We present pairs of sequences containing 4 images per stream to provide a reasonable evaluation set, and, without the underlying text lines to focus on the visual coherence. We evaluate the contribution of our visual features, the pairwise terms, to the optimization, formulating the

Table 6.6: *User study results. Diffferent style groups S are identified due similar impact. S_{cont}: higher impact of content than color (e.g. "sunny", "hazy"), S_{col}: styles largely connected to color (e.g. "noir", "vintage"), S_{abst}: abstract styles (e.g. "minimal").*

	Exp.1: Unary contributions			**Exp.2: Pairwise contributions**			
	tagFreq < wordVec	wordVec < all	tagFreq < all	col < cont	cont < sty	col < sty	noPE < allPE
S_{cont}	68%	60%	78%	85%	58%	73%	71%
S_{col}	54%	63%	60%	26%	71%	55%	74%
S_{abst}	69%	60%	80%	84%	58%	60%	47%



The sea is calm tonight.

The tide is full, the moon lies fair

Upon the straits; on the French coast the light

Gleams and is gone; the cliffs of England stand,

Glimmering and vast, out in the tranquil bay.

Come to the window, sweet is the night-air!

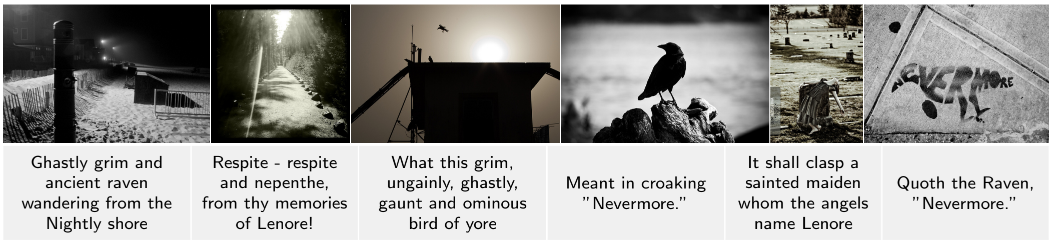
Figure 6.38: *Poem "M. Arnold - DOVER BEACH" in: "sunny" (top), "minimal" (bottom).*

hypotheses $H_{I_{seq}}$ as:

$H_{I_{seq}}$:
- All three features are necessary (style, content, color > 0).
- Highest preference results are retrieved for relation: color $\lesseqqgtr$ content < style.

The constraints described in Section 6.3.6 lead to a dataset of about 1000 image sequence pairs. Based on the outcome of Exp. 1, we set $w_U = (1|1)$ and compare binary contributions between the pairwise feature terms to retrieve relations between them. Figure 6.36 (bottom) shows some tested sequences. Results are shown in Table 6.6. Style was always preferred over the other features to ensure coherence which confirms our initial observation $\mathcal{H}_{ST}$. In this experiment, we identified different groups S of styles. Styles like "sunny" profit from a higher contribution of content than color (S_{cont}, 85% preference). Other styles, e.g. "noir" largely depend on color being preferred 74% over content (S_{col}). Rather abstract styles like "minimal" are not as suitable for auto-illustration as (0|0|0) was preferred 53% over (1|1|1) (S_{abst}).

The results of the study demonstrate the importance of distinguishing between certain styles. Thus, based on the performed binary experiments and the obtained relations, we experimentally obtained different parameter sets relating the proportions of the features for different groups of styles and, similar to experiment 2, tested them against all weights set to 1. Most of the styles worked best for a weight set of $w_U = (.8|1), w_P = (1|.5|.2)$, e.g. "hazy" 80%. Contrarily, for "horror" (1|1|1), i.e.,



Ghastly grim and ancient raven wandering from the Nightly shore

Respite - respite and nepenthe, from thy memories of Lenore!

What this grim, ungainly, ghastly, gaunt and ominous bird of yore

Meant in croaking "Nevermore."

It shall clasp a sainted maiden whom the angels name Lenore

Quoth the Raven, "Nevermore."

Figure 6.39: Poem "E. A. Poe - THE RAVEN" illustrated in style "noir". Note "nevermore" presented in form of a raven and the different selections for "croaking" and "quoting".

setting all weights to 1, was preferred in average 92% over partial combinations. However, very color depending styles worked better combining the features with $w_U = (.8|1), w_P = (1|.2|.5)$, thus setting color > content, e.g. "noir" was preferred 90% over all set to 1 and "vintage" 80%.

Experiment 3: Text illustration. Based on the previously derived parameter settings, we let Turkers rate the quality on a subset of 45 illustrations along text lines on a 5pt Likert scale from 1 (very bad) to 5 (very good). Some tested examples are shown in Figure 6.37. The subjective outcome of our system makes it challenging to obtain scores rating the overall quality. However, for many styles the resulting mean μ was around 4 indicating good quality, e.g., "long exposure" $\mu = 4.0$ ($\sigma = .91$), "noir" $\mu = 3.9$ ($\sigma = 1.07$). We tended to find that results in the style group G_{cont} had stronger decisions (smaller σ-values). "Minimal" only obtained a top-two box acceptance of 37%. However, non-abstract styles obtained top-two box acceptance rates between 70% and 80%, indicating high acceptance of our results, e.g., "sunny" 75%.

6.3.7 Results and Discussion

The main challenge of our approach is to balance semantic relevance with producing an illustration that both depicts the requested style and demonstrates strong visual coherence along the illustration. Figure 6.39 provides a consistent visual appearance of the style "noir" while preserving the meaning of the underlying text lines, even distinguishing between the raven "croaking" and "quoting".

The weak nature of tags and polysemy makes this problem highly challenging, e.g., in Figure 6.40, all images are tagged with "cloud" although the last image (bottom row) does not show a cloud. However, sometimes our method works surprisingly well, e.g., in Figure 6.36 (Exp. 1) the text "The taste of love is sweet" shows an amazing result with our features, providing an idea of taste. Overall, the nice repetitive structure of creative texts allows us to search an image for each text line instead of forcing text splits that can lead to wrongly combined words.

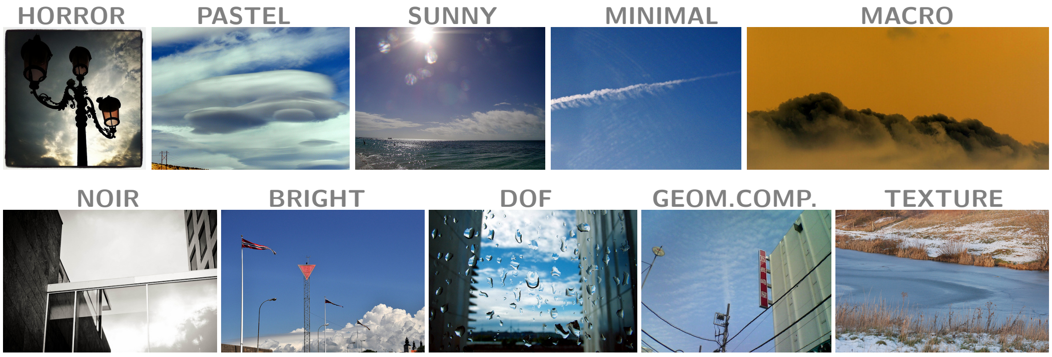


Figure 6.40: *Poem "W. - I WANDERED LONELY CLOUD". The text line "*I wandered lonely as a cloud*" is presented in different illustration styles. All image tags include the word "cloud" although the last image (bottom row) does not depict a cloud.*

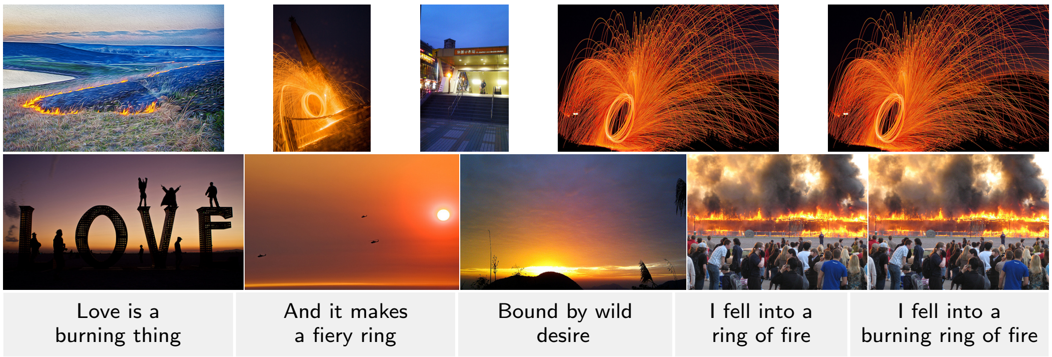

Love is a burning thing
And it makes a fiery ring
Bound by wild desire
I fell into a ring of fire
I fell into a burning ring of fire

Figure 6.41: *Song "J. Cash - RING OF FIRE" illustrated with the styles "long exposure" (top) and "sunny"(bottom). Recurrence is provided for similar texts parts.*

Additionally, restricting the candidates per text line by the style prediction attribute makes it even more challenging to create a semantically relevant illustration. In Figure 6.30, the line "Well, I get down on my knees and pretend to pray" works nicely for the first and last row with the styles "hdr" and "serene", but provides a rather strange X-ray image of a knee for "noir" as there was no better candidate available. Overall, as shown in Section 6.3.6, the way we combine our textual features generally supports selecting suitable images illustrating the meaning of the text.

Further, recurring lines in poems as well as refrains in songs often serve as stylistic device to strengthen the main message by repetition. Thus, if similar content is described in text lines, the images should provide similar or even identical visualizations, e.g., Figure 6.41 visualizes recurring images for lines with similar descriptions. Often, such creative texts deal with a main theme. We capture its global idea by first providing pre-selection of similar candidate lists by recurrence,

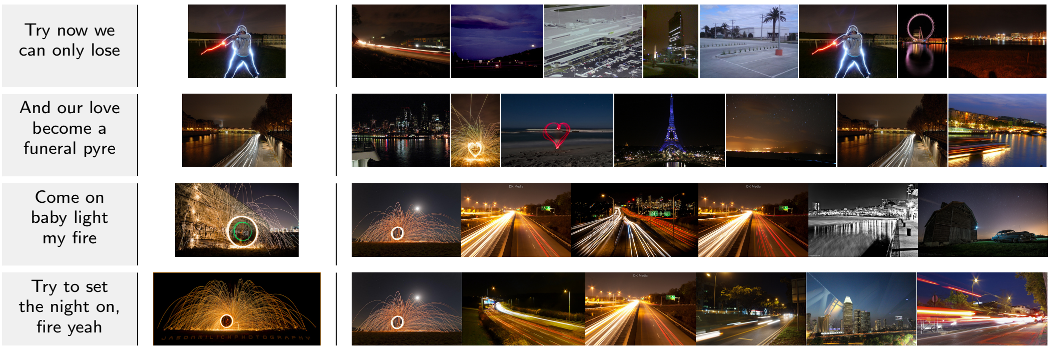

Try now we can only lose
And our love become a funeral pyre
Come on baby light my fire
Try to set the night on, fire yeah

Figure 6.42: Example for candidate lists along part of song "The doors - LIGHT MY FIRE" in illustration style "long exposure". Left: succeeding text lines. Second column: selected optimization result. Right: candidate images per text line.

and, then, the content feature selects images with highest content similarity along the storyline. Figure 6.42 shows an example of candidate lists sorted by highest style prediction (right) along succeeding lines (left) and estimated optimization results (middle). We can observe that the images are selected properly with high coherence in their visual appearance and style along the sequence.

Overall, not every style is similarly suitable to illustrate text. In our user study (Section 6.3.6) we identified different style groups and retrieved lowest acceptance rate for abstract styles. However, even for such styles we were able to retrieve nice results, e.g., in Figure 6.38 for the abstract style "minimal". Additional examples for different styles are shown in Figure 6.41 or in Figure 6.43. Longer and more extensive illustration results are presented in Appendix B.

Finally, rather than restricting our framework to creative texts only, we enable the input of new text data of arbitrary type and length. For a selected style, the text is then parsed and illustrated visually consistent by our system with one image per text line out of the YFCC14M images.

6.3.8 Applications

Our system can enable applications such as identifying the style of a text using our candidate selection process or automatically generating a music video by illustrating complete songs in different styles.

Text Style from Candidate Selection. Restricting the number of retrieved images for text lines by the style constraint $I_{sty} > 50\%$ (Section 6.3.5) leads to interesting insights about the connection between text and styles. We calculated the average number of retrieved images for the text lines in a story relative to the available

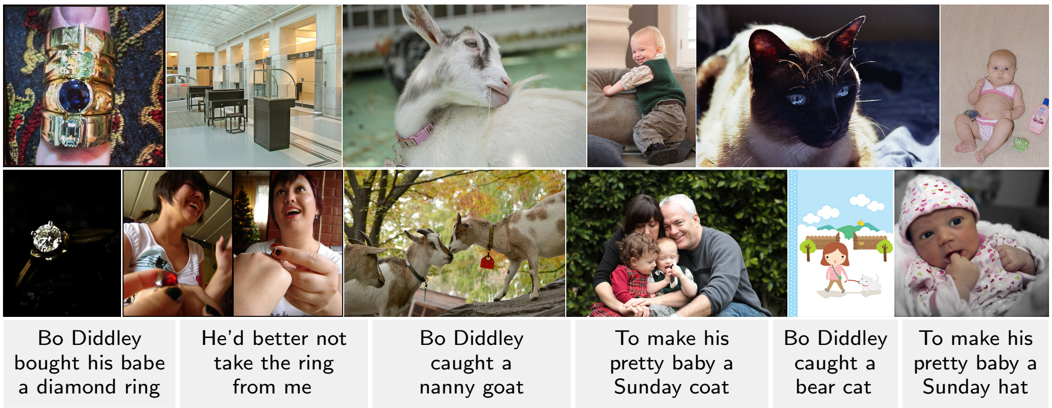

Bo Diddley bought his babe a diamond ring
He'd better not take the ring from me
Bo Diddley caught a nanny goat
To make his pretty baby a Sunday coat
Bo Diddley caught a bear cat
To make his pretty baby a Sunday hat

Figure 6.43: *Song "B. Diddley: BO DIDDLEY". Highest average style-image-responses are obtained for the styles "pastel" (top) and "romantic" (bottom).*

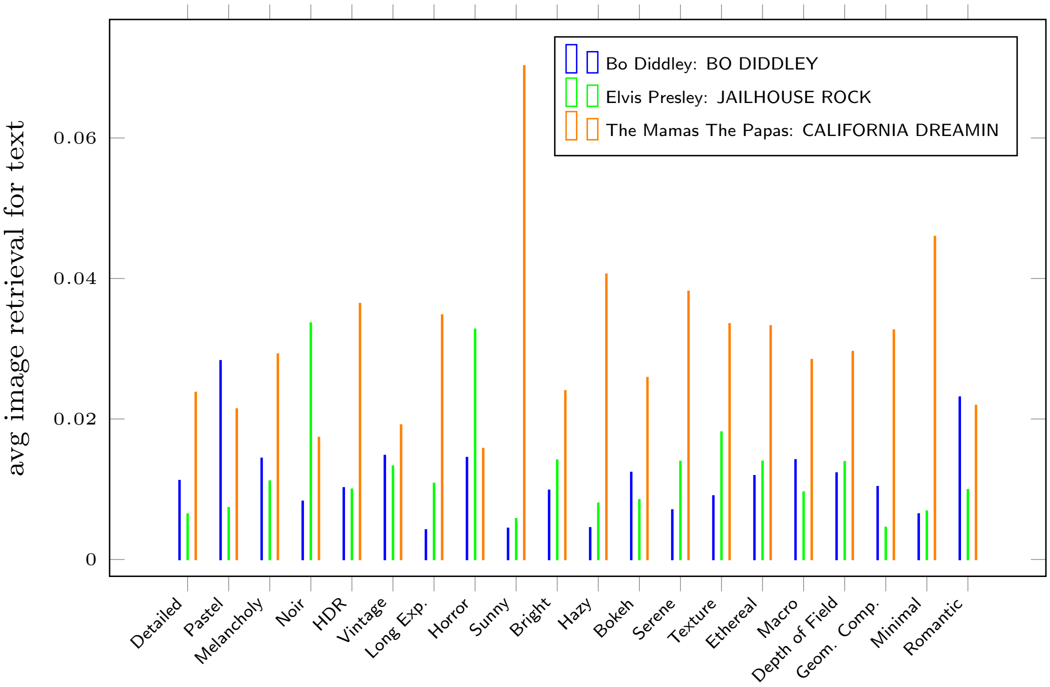


Figure 6.44: *Examples for average style-image-responses for some song lyrics for all 20 styles. Highest peaks indicate main moods of the text.*

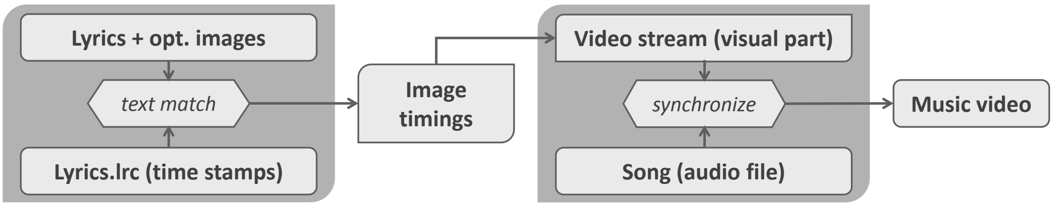


Figure 6.45: *Music video generation pipeline.*

amount of images for a certain style $\#I_{sty}$ in the YFCC14M subset. Thus, for a story text T and a certain style sty, the average style-image-response T_{sty} is calculated as shown in Equation 6.6. Note, that only text lines $l \in L$ with number of candidate images $\#I_{cand_l} > 0$ are considered.

$$T_{sty} = I_{ret}(sty) = \frac{\sum_{l \in L} I_{cand_l}(\text{sty})}{\#l} \cdot \frac{1}{\#I_{sty}}, \quad \text{with } \forall l \in L, \#I_{cand_l} > 0 \tag{6.6}$$

Figure 6.44 demonstrates this idea, showing the resulting average style-image-response of stories for all provided styles. Some texts have a peak in one or two styles, e.g., "The Mamas The Papas: CALIFORNIA DREAMIN" in "sunny". Others even have peaks in connected styles, e.g., "E. Presley - JAILHOUSE ROCK" in "noir"+"horror" or "B. Diddley: BO DIDDLEY" (Figure 6.43) in "pastel" + "romantic". Interestingly, these styles do seem to indicate the main moods of the texts as they often work better as styles with a lower image-style-response.

Music Video Generation. Further, we enable the generation of simple music videos. An overview is given in Figure 6.45. For that purpose, our system outputs a file listing the song lyrics and selected images links. Additionally, the audio version of the song and its ".lrc"-file are needed. The LRC format [Ss12], is usually used in karaoke to align song lyrics with the music. The structure simply consists of the text lines of a song and its associated time-stamps. Optionally, additional ID tags indicating artist and song meta information might be attached.

For the music video generation, we start by matching the text lines between the ".lrc"-file and our "lyrics2images"-file. To be robust against spelling errors, we compare the stems of the words text line to obtain the time-stamps for the images. Our tool converts them into duration timings for each image. A video stream is then simply generated by displaying the images in their proposed ordering and duration timings similar to a slide show. This video stream is joined together with the audio file resulting in a music video. The synchronization is already provided by the image timings. However, beat detection could improve synchronization in future work.

6.3.9 Conclusion

Given a preferred style the presented pipeline automatically generates an illustration for the entire storyline of a poem or song lyric. Our framework optimizes both semantic relevance and visual coherence while selecting images, exploiting recent advances in convolutional neural networks for image and style classification. The generated sequences have been evaluated in a user study, indicating that combining multiple features improves over simpler image selection processes. Similar to our initial statement $\mathcal{H}_{ST}$, the user study revealed that style was the preferred feature to ensure visual coherence compared to the other attributes. Highest acceptance rates have been obtained for non-abstract styles. Finally, we also demonstrate applications to story style classification and music video generation.

7 Conclusion

The main objective of this work is the exploration of image semantics as well as the semantic relatedness between visual and textual sources. Therefore, this thesis presented various approaches to explore the visual appearance of an image in terms of meaningful similarities, aesthetic appeal, and emotional response of an observer. Based on gained insights as well as extensive textual processing we further tackled the semantic gap between textual and visual sources which allowed us to create semantically close visualizations as well as visually coherent picture stories in different visual styles. As semantics is a highly subjective term and implicates a strong connection to people, we exploited highly diverse large-scale online data provided by users and analyzed subjective effects in various user studies.

This chapter concludes the present thesis with a summary of the main outcomes, a discussion on semantic aspects considered during this work, and, finally, gives an outlook on future work that arises from our developed methods.

7.1 Summary

This section briefly summarizes our most important results on image analysis (Section 7.1.1) and text illustration (Section 7.1.2) and discusses the semantic perspectives we have considered throughout this thesis (Section 7.1.3).

7.1.1 Image Analysis

In order to approach visual semantics, in Chapter 5, we focused on analyzing the visual structure as well as the appearance of images under different aspects, namely meaningful similarities, aesthetic appeal, and emotional response:

Meaningful Similarities. Based on the concept of local self-similarities, our GPU-based approach (Section 5.1) efficiently finds similar images in large datasets across visual domains exploiting the local image structure. We organized our implementation such that only a one-time generation of the descriptors is required for all images. The descriptors are saved into a data base that can easily be extended and, further, enables direct efficient matching comparing the spatial arrangements of the descriptors. This setup allows us to search thousands of images in only a few seconds. We applied our implementation to online search results we retrieved for simple textual queries and demonstrated with an additional clustering step that our approach can be utilized to identify various meanings for ambiguous words. Those meaningful similarities can be used to extend the tag-lists of similar images and improve text-based search, or to locate an object within a real-world scene.

Aesthetic Appeal. To explore the meaningful connection between the appearance of pictures and human beings, we investigated in the highly subjective aesthetic appeal of images. In Section 5.2, we presented an approach to learn aesthetic rankings from a broad diversity of user reactions in social online behavior. We showed that users tend to favor aesthetically pleasing images and derived a score from those multi-user agreements in online platforms, in particular Flickr, which we validated in a user study. This score rates how visually pleasing a given photo is and forms the input for our deep learning network. Our developed method is capable of resolving aesthetic relations on a highly fine-granular level without the need to design specific hand-crafted features and enables a variety of applications, namely, resorting photo collections, capturing the best shot, and a smooth prediction along a video stream.

Emotional Response. In order to gain even deeper insights into the influence of the appearance of images towards people, we explored how simple changes in basic global image modifications are capable of actually changing the emotional perception of an observer (Section 5.3). Therefore, we collected empirical data on images associated with emotion labels and analyzed the valence ratings of the different modifications and their strengths. We revealed specific ranges where the tuning of brightness, saturation and color temperature work most efficiently or might produce unintended results. Thereby, we observed that those modifications can sometimes even change the semantics of an image, for example, Figure 5.37 shows an example were a too strong modification of the color temperature produces an impression of being close to fire. From all those findings, we derived our EmoTune filter which allows for almost linear control by combining specific modes, and demonstrated successful application to both images and videos.

7.1.2 Text Illustration

Further, we explored the semantic connection between natural language and images. In Chapter 6, we presented several methods to bridge the semantic gap between textual and visual sources with the aim at illustrating text:

3D Visualizations of Descriptive Texts. Starting with 3D scenes as visual representation, we demonstrated in Section 6.1 how text processing can be employed to arrange 3D models utilizing information extracted from descriptive texts. Based on an extensive textual analysis, we focused on identifying relevant units and their dependencies within the textual description. Our derived object-to-object relations are then used to resolve the spatial dependencies between the objects and to correctly arrange the 3D models within the scene. Overall, we showed that employing natural language can facilitate the creation of virtual environments.

Relevant Illustrations from Online Photo Collections. In order to further explore the semantic connection between textual and visual sources, we exploited the huge amount of online available imagery associated with meta information to find visual representations for arbitrary texts. In Section 6.2, we presented a system that illustrates a given text by retrieving relevant images from online photo collections. Therefore, we employed methods from NLP to parse the input text. We extracted relevant information, constructed meaningful textual search terms, and developed a hierarchical algorithm to query online photo collections. With our querying method we were able to retrieve relevant results with high precision and, thus, semantically close images to a given text snippet. We compared the average precision of the retrieval results on different types of texts and observed variations between the categories. The final images are selected in a user-assisted process and presented in the form of a storyboard. Our system allows to create a photomatic animation from the resulting illustration and, thus, can be seen as one of the first steps towards creating a movie based on a textual input.

Illustrating Creative Texts with Style. Finally, in Section 6.3, we demonstrated an approach to illustrate complete storylines with semantically close images and incorporated their appearance to obtain not only relevant but even visually coherent picture stories in different styles. Therefore, we employed creative texts as such artist's compositions are intended to be highly emotional and claim for visualizations that are aesthetically pleasing and highly stylistic according to the style of the artwork. Our presented framework optimizes both, semantic relevance and visual coherence while selecting images, and exploits recent advances in convolutional neural networks for image and style classification. Although such creative texts are quite challenging due to their high level of abstraction, we demonstrated that our combination of features is capable of generating image sequences in a consistent visual appearance in a specific style while preserving the meaning of the underlying text lines. Further, we presented applications to identify the style of a text based on our candidate selection process, or automatically generating a music video by illustrating complete songs in different styles.

7.1.3 Semantic Challenge

In general, as presented in Chapter 2, understanding semantics is challenging as the meaningful interpretation is very subjective and usually differs between people. For machines it is even worse to approach high-level meanings as they often need to be grasped from the context and the semantic gap between different sources can be almost insurmountable. This thesis presented several techniques to approach semantics. In summary, we mainly explored "semantics" from two perspectives:

- *Visual semantics* of single images analyzing their appearance under different aspects (Chapter 5)
- *Semantic connection* between textual and visual sources with the aim at creating semantically close illustrations based on textual descriptions (Chapter 6)

As mentioned throughout this work and relating to these two perspectives, semantics is highly subjective and largely depends on the meaningful interpretation of people. Thus, there was a strong necessity to relate human knowledge to this work. In addition to several user studies to verify the quality of our results, we have addressed this semantic challenge by exploiting large amounts of visual data previously associated with human textual information. From huge online photo collections we queried a diversity of visual data based on the requirements of the various presented tasks. The big bottleneck of the existing noise in the retrieved data, as image tags are not always reliable, has accompanied us all the time.

In particular, to approach visual semantics, we exploited such Flickr images to demonstrate how the visual structure of images can be utilized to derive meaningful similarities which form visual clusters and are capable of identifying various meanings of ambiguous queried words (Section 5.1.5). Similarly, we used image data previously labeled with emotional tags to approach the emotional influence of images (Section 5.3.3). In addition to the tagging process, we observed a favoring behavior of users in social online platforms rating images they like. Thus, we exploited this highly diverse and meaningful meta information that reflects the behavior of a huge amount of users to derive a score of aesthetics and learn rankings of pleasing images (Section 5.2). Furthermore, our investigation in the emotional effect even directly connected an observer to an image as we explored to which amount slight modifications in the global visual appearance can actually change the emotional perception of a person (Section 5.3). We have even observed changes in image semantics through these modifications.

Further, we investigated in the semantic connection between textual and visual sources with the aim at illustrating texts. Thereby, the main challenge is to bridge the semantic gap, i.e., to find a visual representation that shows the same meaning as described in the textual description. In addition to making use of large image data bases, we employed methods of NLP to analyze the given text as well as the meta data associated with the images to obtain a semantically strong connection on the textual level. Thus, harvesting in multiple knowledge sources in the form of texts and image databases allowed us to create relevant visual representations for simple text elements (Section 6.2) as well as complete storylines (Section 6.3). Our developed hierarchical querying algorithm (Section 6.2.4) demonstrated successful retrieval of images with high semantic precision from web-scale online photo collections to given text snippets and, therefore, was also used to obtain relevant image data based on associated emotion labels (Section 5.3.3). Further, in Section 6.3, we extended our illustration task to complete storylines. Thereby, the main challenge was to balance semantic relevance with producing an illustration that both depicts the requested style and demonstrates strong visual coherence along the illustration. However, the way we combined our textual features still supported selecting suitable images illustrating the meaning of the text (Section 6.3.6). Besides, although the style feature restricted the available amount of imagery for the illustration, at the same time, it strengthened the visual coherence of the appearance and, thus, the visual semantics of the picture story.

Overall, we have demonstrated the strong necessity to directly or indirectly integrate human beings when exploring semantics due to its immense subjectivity and have presented relevant contributions to tackle this challenging task with its high level of abstraction. Examining this intriguing subject is an important further step in depicting the real world in machines.

7.2 Future work

This thesis presented a complete pipeline with promising results to automatically translate natural language into meaningful visual representations with the focus on still imagery and visual semantics. We believe that this work paved the way for *language based dynamic video generation* and raised new challenges on the way from Text–to–Video (Figure 1.1).

Our techniques on analyzing the high-level meaning of images demonstrate significant achievements to approach visual semantics from a computational perspective. Overall, the present thesis can be seen as a foundation for *intriguing visual storytelling*. Several extensions claim to be explored to further pursue this line of research and will be pointed out in the following.

Besides, the present work considerably contributes to bridge the semantic gap between text and images. Thus, we additionally propose new ideas for future projects that arise from connecting natural language with the visual world as well as integrating visual semantics into the interaction between humans and machines.

Combine Visual Storytelling with Aesthetics and Emotion. In Section 6.3, we presented an automatic approach to illustrate texts with a user selected style. A straightforward extension would be to directly integrate the techniques developed in Chapter 5 into the illustration system, more precisely, to create aesthetically pleasing and emotionally affective visual stories.

Selecting pictures due to their aesthetic appeal (Section 5.2) along a story could intensify rather beautiful appearing styles like "romantic" or "sunny"and result in an even more pleasing image sequence. However, one has to take into account that an additional constraint further restricts the image set during the selection process and the already huge initial amount of images must be further increased.

In contrast, the EmoTune filter (Section 5.3) can be applied directly on top of the outcome, i.e., the images of the final sequence. First of all, modifying the emotional influence of an image stream could emphasize the effect of rather negative or more positive styles. For this purpose, the emotional direction of the single styles needs to be explored. Furthermore, controlling the emotional effect of a complete picture story would allow to directly relate the visual story to the feelings of the observer. Generally, when jointly tuning a sequence of images rather than individual pictures, an even stronger impact is expected and could be certainly useful in advertising, e.g., to positively influence humans towards a certain product.

Adaptive Auto-Illustration of Books with Style. Further, although we focused on creative texts like poems and song lyrics, our method presented in Section 6.3 can also be applied to arbitrary input texts which are suitable for illustration with a specific style. Given such a style, we demonstrated the capabilities of machines to produce corresponding results by automatic means. The next reasonable step would be to expand from self-contained texts towards a larger scale like complete books or movie scripts. However, especially longer texts typically change the mood along the story and would benefit from automatic means that adapt the visual style according to the storyline. To enable such an adaptive style alignment of images, a further processing step is required on the text side. The mood of the text needs to be recognized beforehand, e.g., through the identification of affective words. Work in NLP that has investigated in this direction can be utilized. For example, an affective extension has been built by Strapparava and Valitutti [SV04] upon WordNet (Section 2.1.2), and a set of affective words has been collected by Stevenson et al. [SMJ07]. Additionally, the ranges and boundaries of parts within the large text establishing a certain mood need to be found. Such a text part can be a passage within the text, but also a complete section, or chapter in the book. As soon as the primary moods are identified for those text parts, the style of the corresponding image streams can be set automatically and the images can be selected due to the recognized style along the according text parts. Similarly, if the emotion established by a text part can be identified, e.g., through sentiment analysis of the natural language [Liu12, SPW$^+$13], the image streams can be modified by applying our EmoTune filter presented in Section 5.3 and mentioned previously.

Besides, having aligned an image sequence along a book, the pictures can serve as keyframes that can be matched with a film version of the story and support the recently upcoming book–movie alignment task [ZKZ+15, TBS15]. Overall, applying such a large-scale adaptive approach to complete movie scripts further supports the main idea of this thesis, namely a full-fledged movie generation.

Text–to–Video: Language-based Dynamic Video Generation. In Chapter 6, we demonstrated the potential of starting from an arbitrary text to produce a photomatic animation or music video. This represents an important step towards the crucial idea of this thesis, namely to create a stand-alone Text–to–Video system (Figure 1.1). However, in the long run, several extensions wait to be fulfilled to achieve the automatic generation of a full-fledged video.

To realize the creation of dynamic videos, a potential idea would be to exploit moving images provided by Flickr instead of still pictures. However, the available amount seems rather small. For example, within the YFCC100M data set presented in Section 6.3.4 only around 80K are moving images and, therefrom, only 53% are associated with tags resulting in about 42K useable moving images which is not enough data to support visualizations for arbitrary texts.

Another approach would be to use the final picture stream as a foundation and build an animation upon it. The basic picture stream could be mapped into a multi-layer setup, e.g., similar to a stage in theater, by detaching the foreground region [STZ+16] from the background and handling them separately. Extracting foreground objects or persons, e.g., by combining salient region detection [CZM+11] with the GrabCut approach [RKB04], enables reusing them throughout the story. As indicated in Section 6.2.5, we made initial attempts to allow recurrence of the protagonist, i.e., the main character of the story. Inspired by the theater setup, a further big improvement would then be the transformation of the foreground into a 3D scene setup by converting the 2D foreground items into 3D objects. An approach to match images with 3D objects has been presented by Tasse et al. [TD16] and could be integrated. Additionally, our work presented in Section 6.1 can be employed to correctly place the 3D objects in the scene and could be updated to establish a certain style of the models by considering previous work on style of 3D furniture items described in Section 6.3.2. Finally, to enable moving objects and characters, a next challenge would be to learn the movements according to the actions described in the text. For example, in the sentence “a man walks to the tree”, not only the items “man” and “tree” need to be identified, but also the action of “walking” towards the “tree”. To already animate the 2D items, one could build upon the work introduced by Xu et al. [XWL+08].

A further challenge consists in establishing a smooth appearance or transition between the background images of the scenery. To obtain a consistent color appearance, the color consistency approach introduced by Park et al. [PTSSK16] can be applied on top of the image sequence. Besides, the success of learning methods recently enabled the generation of new content, e.g., the StackGAN is capable to produce graphics content from text [ZXL+16]. The results are promising, however, incorporating meaningful aspects like the aesthetic pleasingness or even the emotional effect could strongly enhance their visual outcome. Anyway, a similar approach could be employed to generate missing frames between two existing pictures with an additional blending step on top to smoothen the transition. This approach could also be applied to the foreground 2D snippets to generate a smooth animation or, generally, to any other situation where frames need to be generated.

When setting up such a new scenery, another important aspect is the lighting. However, as indicated in Section 5.3, several color tones or combinations are related to certain emotional stimuli. Thus, as mentioned previously, extracting the emotion transported by a piece of text allows to modify the appearance of the scene.

Finally, to further ease the storytelling process for the user, it is straightforward to allow for speech as input in addition to written text. In doing so, speech recognition can be employed to transform the spoken input into machine-readable text [HDY+12, DHK13]. Generally, adding an audio layer like speech or sound to the output would make the resulting video output more lively as indicated by our music video application (Section 6.3.8). One could build upon learning approaches connecting visual data and sound which have arisen recently [AVT16, CUF16].

Visual Semantics in Human-Computer Interaction. Meanwhile, affective computing became a popular research direction in the field of human-computer interaction (HCI). Valitutti et al. [VSS04] stated:

> "There is a wide perception that the future of human-computer interaction is in themes such as entertainment, emotions, aesthetic pleasure, motivation, attention, engagement, etc." [VSS04]

Providing machines with an "emotional intelligence" has been denoted by Picard [Pic00] as giving computers the capability of recognizing and responding to the users emotional state combined with sensing and reasoning about the environmental context. Several parts of this thesis can provide valuable contributions to this line of research.

First of all, linking natural language and images allows to establish a text-based interface to the visual world which can support tasks like machine-based localization of objects within our daily environment (Section 5.1.5). Further, visual feedback illustrating the natural language input (Chapter 6) can be helpful to verify that the meaning of the interpretation of the machine remains the same.

Further, research on NLP progresses in exploring the connection between natural language and affective information which is highly relevant for HCI to give machines an affective understanding of the users emotional state [VSS04]. Thereby, our EmoTune filter (Section 5.3) can be easily employed on the machine-side to modify visual feedback accordingly to the sentiment input towards a similar or even the contrary direction of the users feelings and, to some extent, "react emotionally". Similarly, meaningful visual feedback can be provided through aesthetically pleasing images (Section 5.2) or pictures in a particular visual style (Section 6.3).

Overall, we believe that, when building intelligent systems that support human beings, the integration of visual semantics is a crucial step to provide machines with a reasonable visual feedback and push the meaningful communication between humans and machines to the next level.

A Frequency List from Creative Texts

The following list presents the relevant words we extracted from our creative text corpus to assemble a suitable image corpus as described in Section 6.3.4. This list has the form of a frequency list (see Section 3.4.2) and displays the complete set of 408 relevant words $w_{\mathcal{R}}$ and their corresponding frequencies f_{tags} in the YFCC100M photo tags. The list is sorted by frequency in descending order.

Table A.1: *Frequency list of relevant words $w_{\mathcal{R}}$ from creative texts and their corresponding frequencies f_{tags} in YFCC100M photo tags.*

No.	Word	Frequency
	$w_{\mathcal{R}}$	f_{tags}
1	music	820079
2	city	687049
3	party	661219
4	people	639694
5	water	636765
6	summer	599420
7	sky	542511
8	night	531225
9	street	500975
10	live	493598
11	sea	408175
12	spring	394323
13	blue	388669
14	car	382249
15	red	374965
16	show	360719
17	river	357076
18	tree	349476
19	rock	312632
20	sign	304232
21	bird	291224
22	light	283785
23	road	283252
24	dog	268180
25	house	266220
26	mountain	262585
27	bridge	259529
28	band	257299
29	dance	245674
30	fall	241035
31	girl	238404
32	train	232460
33	sun	230847
34	day	219835
35	baby	210286
36	woman	206025
37	may	187026
38	hotel	180363
39	school	174451
40	running	173061
41	home	156502
42	man	155756
43	ice	152018
44	wall	149136
45	world	142880
46	walk	139324
47	fire	136886
48	love	125622
49	rain	124208
50	window	123952
51	life	114412
52	town	114134
53	air	111979
54	dancing	111969
55	game	106889
56	run	102462
57	country	101579
58	work	100730
59	track	100066
60	ship	97728
61	wood	96542
62	state	93718
63	stone	92755
64	glass	91390
65	west	89529
66	pride	88318
67	war	87475
68	cold	85215
69	morning	81511
70	face	79985
71	rise	79177
72	play	76616
73	smile	76222
74	moon	75184
75	wine	75145
76	boy	72678
77	child	71132
78	evening	70844

No.	Word	Frequency
	$w_{\mathcal{R}}$	f_{tags}
79	rose	69939
80	cross	67459
81	door	66664
82	leaf	60185
83	book	59496
84	crowd	57411
85	wild	57042
86	ball	52685
87	ride	50225
88	sweet	50067
89	cool	50002
90	star	45804
91	time	45037
92	dress	44516
93	folk	43743
94	sound	43500
95	place	43305
96	point	42909
97	peace	42413
98	movie	42304
99	engine	42289
100	hair	42203
101	shot	41014
102	mirror	39617
103	lady	39102
104	drink	38763
105	power	38691
106	drive	37485
107	kiss	37421
108	death	35625
109	see	35542
110	smoke	34996
111	dawn	34674
112	silver	34058
113	line	33721
114	hand	33630
115	wind	33101
116	room	32043
117	year	31991
118	land	30969
119	eye	30551
120	head	29582
121	dead	29288
122	friend	29105
123	chair	29072
124	king	27819
125	end	26700
126	heart	26619
127	wheel	25677
128	mary	24987
129	fast	24419
130	good	24322
131	god	23981
132	singing	22967
133	battle	21769
134	week	21697
135	mother	21484
136	close	19787
137	body	19729
138	living	19596
139	son	19323
140	set	19117
141	meet	18877
142	side	18798
143	ring	18714
144	make	18571
145	bell	18403
146	cell	18328
147	gun	18287
148	daughter	18286
149	drag	18116
150	burning	17894
151	warm	17873
152	swing	17730
153	bag	17687
154	change	17598
155	broken	17507
156	back	17392
157	lost	17331
158	half	17325
159	number	17163
160	bed	17096
161	company	16884
162	hip	16866
163	babe	16597
164	dream	16547
165	hope	16439
166	break	16368
167	looking	16082
168	way	15837
169	inside	15707
170	soul	15553
171	record	15492
172	father	14805
173	look	14287
174	stop	14124
175	money	13996
176	round	13862
177	dirty	13455
178	strike	13433
179	met	13151
180	sleep	12664
181	wife	12633
182	seat	12587
183	foot	12541
184	still	12278
185	song	12248
186	waiting	12216
187	shine	12049
188	front	11998
189	spirit	11946
190	stand	11797
191	gift	11784
192	sing	11450
193	lot	11445
194	sister	11252
195	sad	11107
196	closed	10988
197	burn	10895
198	fine	10151
199	alone	9796
200	corner	9719
201	brother	9370
202	darkness	9338
203	start	9240
204	vision	9027
205	distance	8849
206	watch	8794
207	mile	8773
208	watching	8757
209	shoe	8708
210	tide	8644
211	found	8636
212	laugh	8467
213	miss	8426
214	heaven	8201
215	hero	7954
216	talk	7887

No.	Word	Frequency
	$w_{\mathcal{R}}$	f_{tags}
217	shade	7833
218	pain	7781
219	deep	7428
220	honey	7337
221	wonder	7269
222	move	7249
223	today	7101
224	save	6897
225	help	6894
226	win	6831
227	hell	6806
228	arm	6327
229	fear	6300
230	hour	6128
231	moment	5995
232	devil	5985
233	word	5856
234	lord	5786
235	well	5750
236	made	5696
237	laughing	5653
238	sight	5457
239	like	5334
240	felt	5290
241	left	5246
242	taste	5128
243	die	5107
244	bone	4941
245	sit	4901
246	stair	4760
247	keep	4693
248	care	4654
249	lead	4631
250	part	4619
251	loud	4606
252	kind	4472
253	pray	4416
254	right	4395
255	beat	4239
256	past	4115
257	thinking	4058
258	voice	3885
259	step	3868
260	better	3831
261	thought	3749
262	buy	3731
263	call	3718
264	think	3636
265	name	3625
266	turn	3619
267	get	3509
268	hurt	3491
269	grow	3489
270	doe	3380
271	going	3249
272	mind	3158
273	leaving	3128
274	everything	3068
275	thing	3010
276	born	2974
277	fell	2959
278	kill	2951
279	wear	2940
280	seen	2936
281	wish	2885
282	dig	2853
283	leave	2827
284	gone	2799
285	truth	2796
286	pale	2762
287	used	2741
288	use	2709
289	yes	2636
290	take	2550
291	tell	2517
292	wait	2487
293	saw	2469
294	hide	2366
295	comfort	2324
296	beyond	2219
297	desire	2185
298	sweat	2040
299	blow	2029
300	burned	2019
301	remember	2000
302	smell	1989
303	believe	1987
304	wan	1901
305	shake	1898
306	hold	1895
307	something	1793
308	stay	1765
309	tale	1752
310	shoulder	1730
311	come	1725
312	lover	1721
313	cry	1699
314	nothing	1607
315	loving	1600
316	write	1594
317	reach	1560
318	find	1555
319	feel	1544
320	bit	1508
321	lie	1494
322	thousand	1453
323	please	1403
324	let	1387
325	knee	1351
326	none	1304
327	need	1294
328	lip	1244
329	never	1231
330	done	1215
331	tear	1204
332	fate	1200
333	shame	1190
334	carry	1123
335	sent	1121
336	forget	1120
337	give	1113
338	say	1053
339	know	1044
340	try	1014
341	loved	1006
342	got	1001
343	caught	992
344	held	950
345	guess	937
346	coming	932
347	said	888
348	bring	860
349	tune	853
350	thrill	849
351	ask	844
352	reason	821
353	fool	814
354	broke	774

No.	Word	Frequency
	$w_{\mathcal{R}}$	f_{tags}
355	hear	764
356	lay	764
357	fade	742
358	misery	726
359	minute	698
360	want	669
361	breath	647
362	spoke	647
363	somebody	604
364	spell	575
365	roaring	575
366	answer	574
367	rattle	563
368	put	521
369	died	516
370	grown	496
371	trying	441
372	till	433
373	gave	413
374	meaning	408
375	stopped	375
376	doubt	373
377	beneath	371
378	heard	357
379	knock	355
380	someone	347
381	came	332
382	send	311
383	vow	303
384	everybody	299
385	pulled	290
386	turned	271
387	called	262
388	changed	240
389	claim	239
390	vain	228
391	went	220
392	opened	219
393	took	218
394	understand	137
395	tried	132
396	lived	131
397	though	117
398	told	106
399	kept	105
400	kissed	78
401	knew	72
402	stood	71
403	seem	69
404	carried	65
405	grew	37
406	followed	20
407	began	18
408	sweeter	8

B Story Illustrations

This appendix presents a variety of example results produced by our pipeline from Section 6.3, namely illustrations of different creative texts in different styles. Thereby, parts of succeeding text lines are gathered from poems (Section B.1) and song lyrics (Section B.2) and associated with one or more rows illustrating one visual style along a row.

B.1 Poems

Illustrations of poems in different visual styles will be presented on the following pages in landscape.

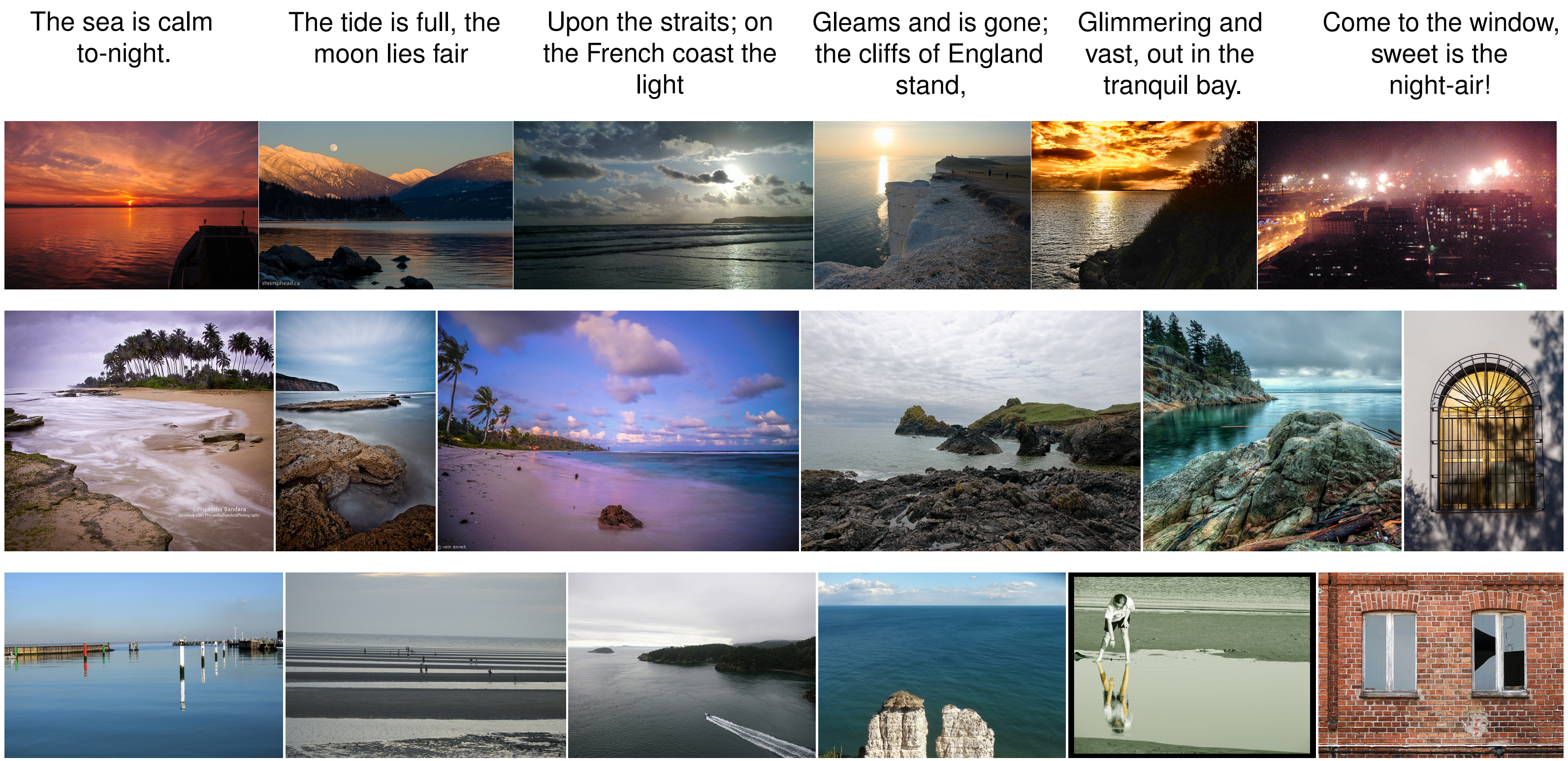
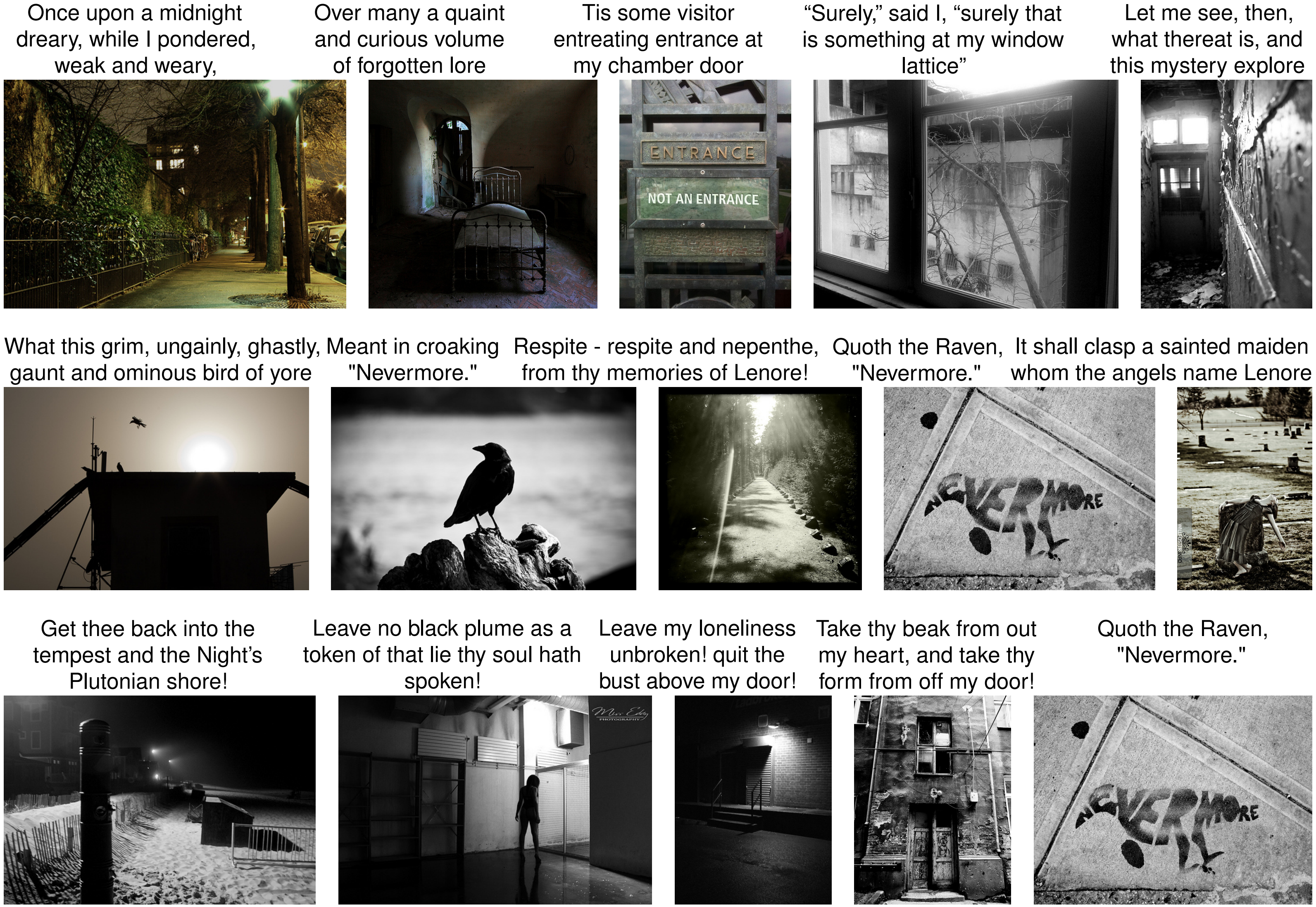

The sea is calm to-night.
The tide is full, the moon lies fair
Upon the straits; on the French coast the light
Gleams and is gone; the cliffs of England stand,
Glimmering and vast, out in the tranquil bay.
Come to the window, sweet is the night-air!

Figure B.1: *Poem: "Mathew Arnold - DOVER BEACH". Styles from top to bottom: "sunny", "long exposure", "minimal".*



Once upon a midnight dreary, while I pondered, weak and weary,

Over many a quaint and curious volume of forgotten lore

Tis some visitor entreating entrance at my chamber door

“Surely,” said I, “surely that is something at my window lattice”

Let me see, then, what thereat is, and this mystery explore

What this grim, ungainly, ghastly, gaunt and ominous bird of yore

Meant in croaking "Nevermore."

Respite - respite and nepenthe, from thy memories of Lenore!

Quoth the Raven, "Nevermore."

It shall clasp a sainted maiden whom the angels name Lenore

Get thee back into the tempest and the Night’s Plutonian shore!

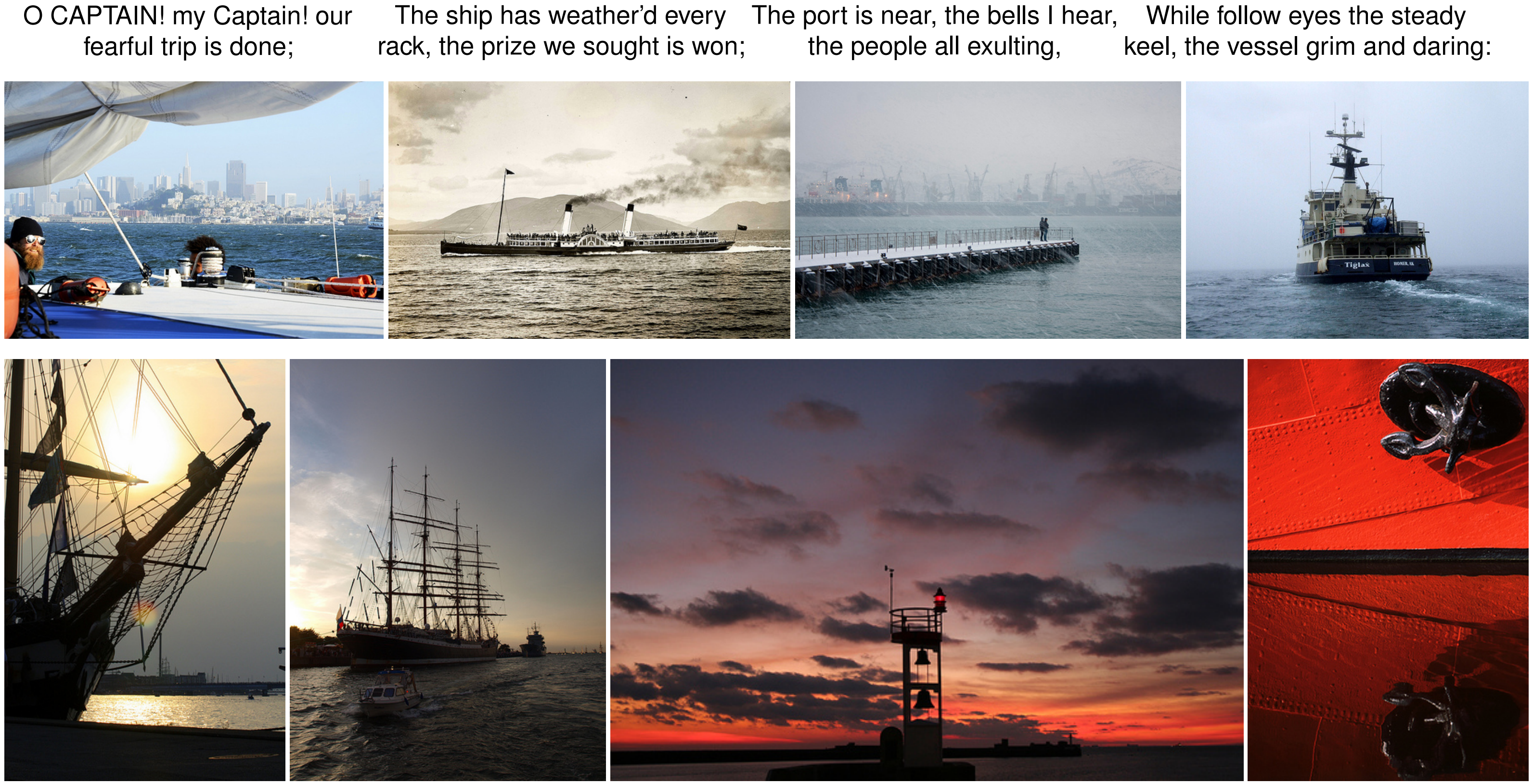
Leave no black plume as a token of that lie thy soul hath spoken!

Leave my loneliness unbroken! quit the bust above my door!

Take thy beak from out my heart, and take thy form from off my door!

Quoth the Raven, "Nevermore."

Figure B.2: *Poem: “Edgar Allan Poe - THE RAVEN”. Style: “noir”.*



O CAPTAIN! my Captain! our fearful trip is done;

The ship has weather'd every rack, the prize we sought is won;

The port is near, the bells I hear, the people all exulting,

While follow eyes the steady keel, the vessel grim and daring:

Figure B.3: *Poem: "Walt Whitman: O CAPTAIN MY CAPTAIN". Styles: "hazy" (top row), "sunny" (bottom row).*

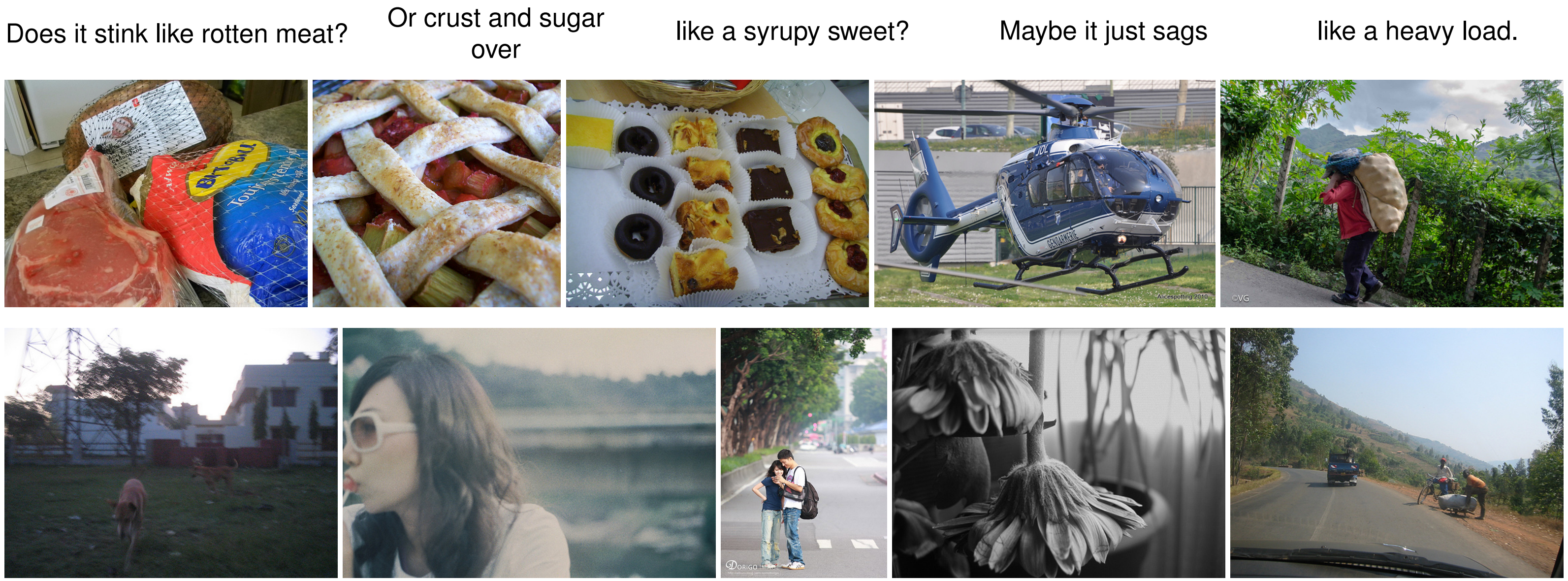
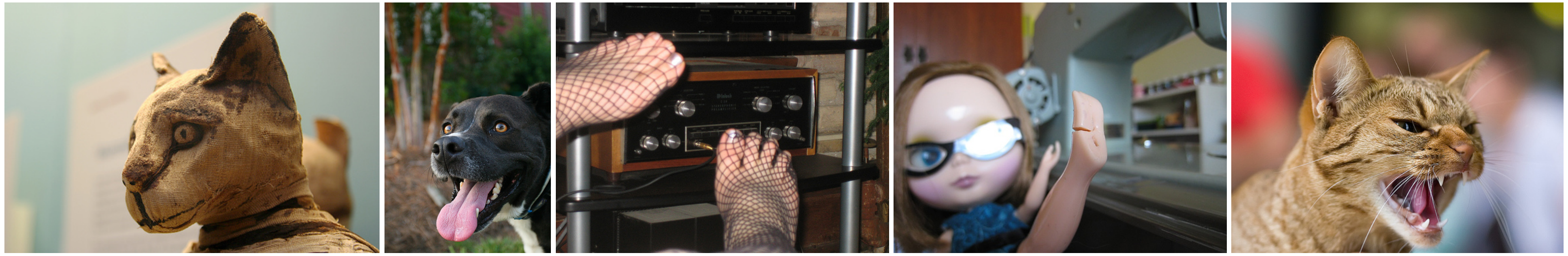


Figure B.4: *Poem: "Langston Hughes - DREAM DEFERRED". Style: "detailed" (top row), "melancholy" (bottom row).*



Be glad your nose is on your face

Imagine if your precious nose

were sandwiched in between your toes, that clearly would not be a treat,

for you'd be forced to smell your feet.

Your nose would be a source of dread

Figure B.5: *Poem: "Jack Prelutsky - BE GLAD YOUR NOSE YOUR FACE". Style: "depth of field".*

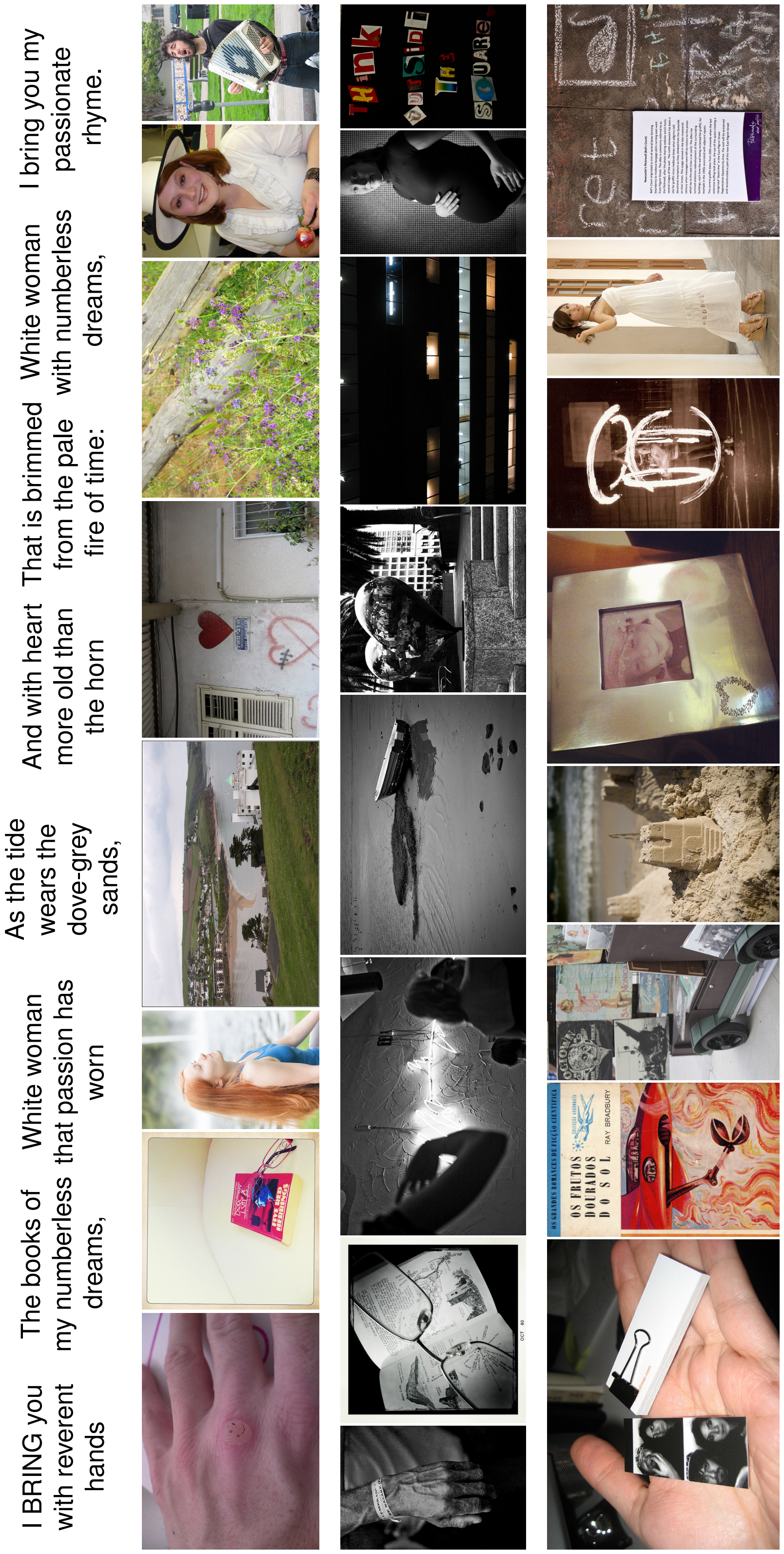

I BRING you with reverent hands
The books of my numberless dreams,
White woman that passion has worn
As the tide wears the dove-grey sands,
And with heart more old than the horn
That is brimmed from the pale fire of time:
White woman with numberless dreams,
I bring you my passionate rhyme.
Think outside the square
OS FRUTOS DOURADOS DO SOL
RAY BRADBURY

Figure B.6: *Poem: "William Butler Yeats: POET HIS BELOVED". Styles row wise: "pastel", "noir", "vintage".*

B.2 Song Lyrics

Illustrations of song lyrics in different visual styles will be presented on the following pages in landscape.

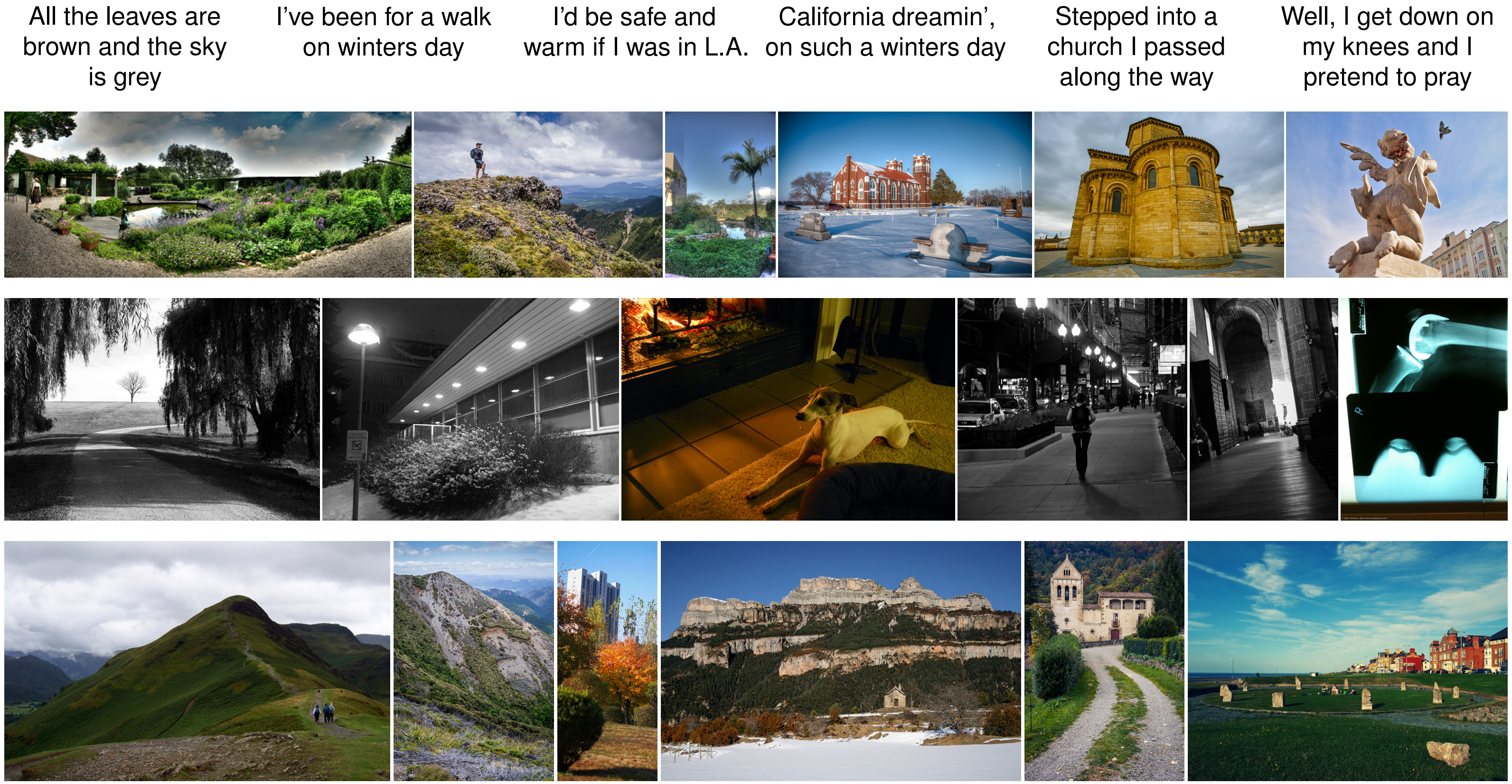
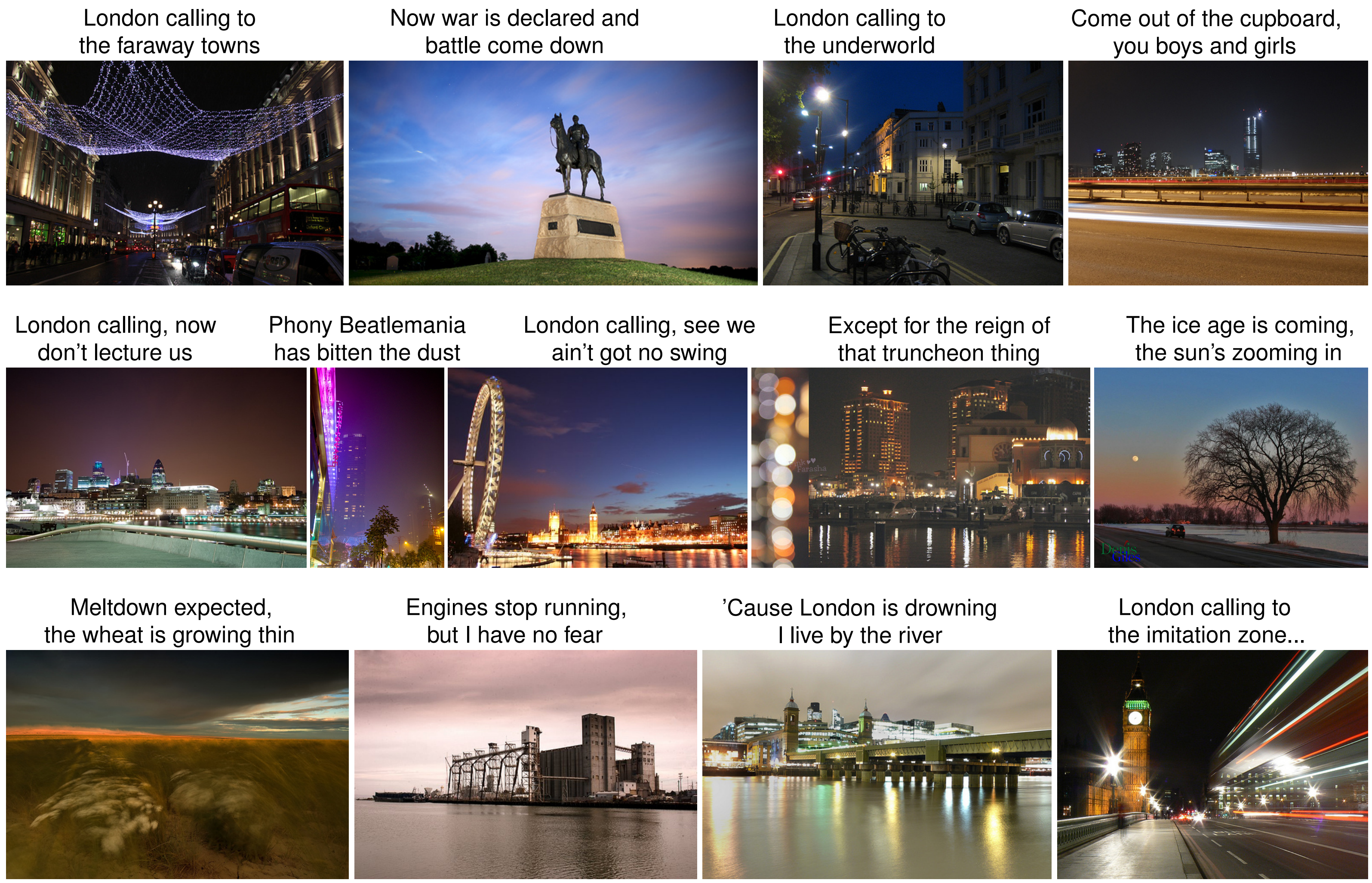

All the leaves are brown and the sky is grey
I've been for a walk on winters day
I'd be safe and warm if I was in L.A.
California dreamin', on such a winters day
Stepped into a church I passed along the way
Well, I get down on my knees and I pretend to pray

Figure B.7: *Song: "The Mamas The Papas - CALIFORNIA DREAMIN". Styles row wise: "hdr", "noir", "serene".*



London calling to
the faraway towns

Now war is declared and
battle come down

London calling to
the underworld

Come out of the cupboard,
you boys and girls

London calling, now
don't lecture us

Phony Beatlemania
has bitten the dust

London calling, see we
ain't got no swing

Except for the reign of
that truncheon thing

The ice age is coming,
the sun's zooming in

Meltdown expected,
the wheat is growing thin

Engines stop running,
but I have no fear

'Cause London is drowning
I live by the river

London calling to
the imitation zone...

Figure B.8: *Song: "The Clash - LONDON CALLING". Style: "long exposure".*

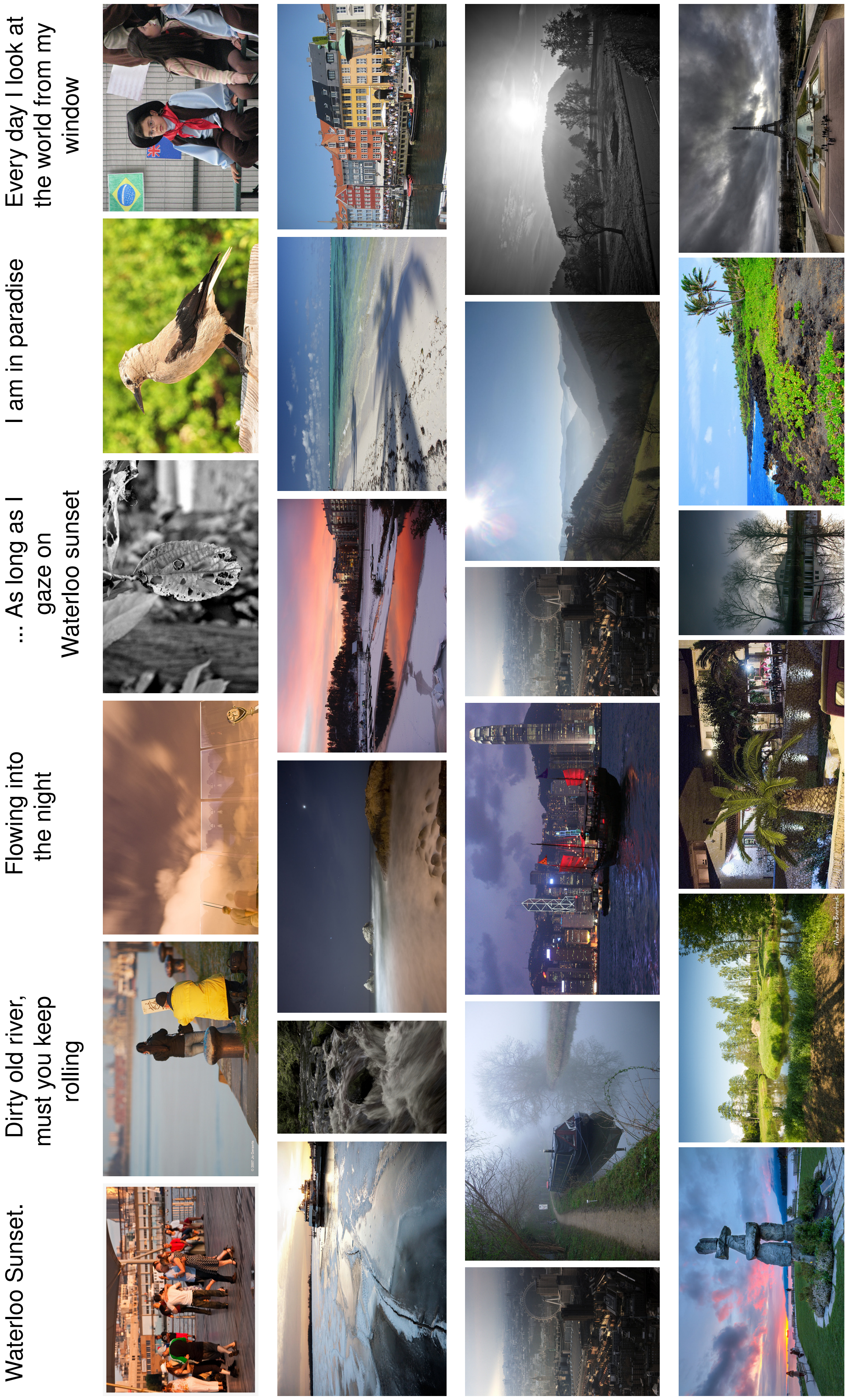
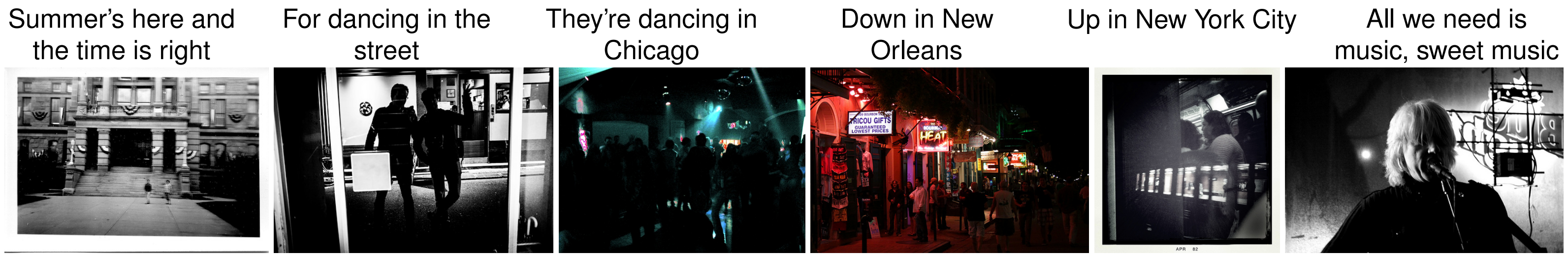

Waterloo Sunset.
Dirty old river, must you keep rolling
Flowing into the night
... As long as I gaze on Waterloo sunset
I am in paradise
Every day I look at the world from my window

Figure B.9: *Song: "Kinks - WATERLOO SUNSET". Styles row wise: "bokeh", "long exposure", "hazy", "hdr".*



Summer's here and the time is right

For dancing in the street

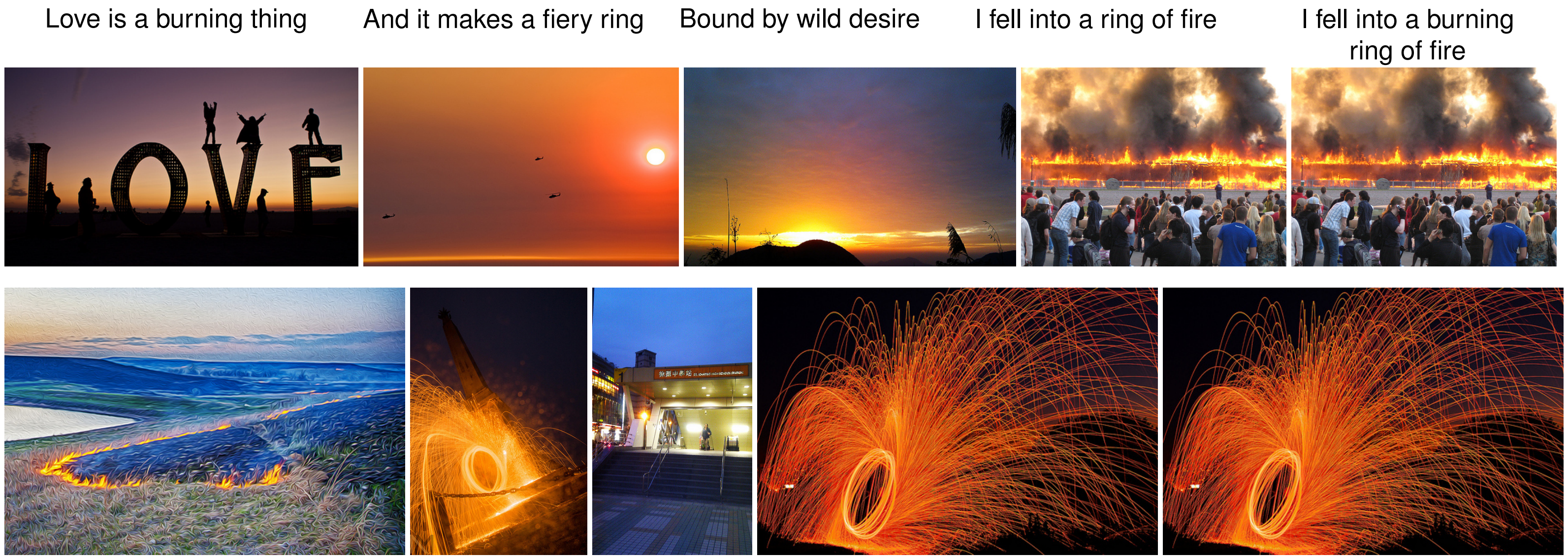
They're dancing in Chicago

Down in New Orleans

Up in New York City

All we need is music, sweet music

__Figure B.10:__ Song: "Martha and the Vandellas - DANCING IN THE STREETS". Style: "noir".



Love is a burning thing

And it makes a fiery ring

Bound by wild desire

I fell into a ring of fire

I fell into a burning ring of fire

__Figure B.11:__ Song: "Johnny Cash - RING OF FIRE". Styles: "sunny" (top row), "long exposure" (bottom row).

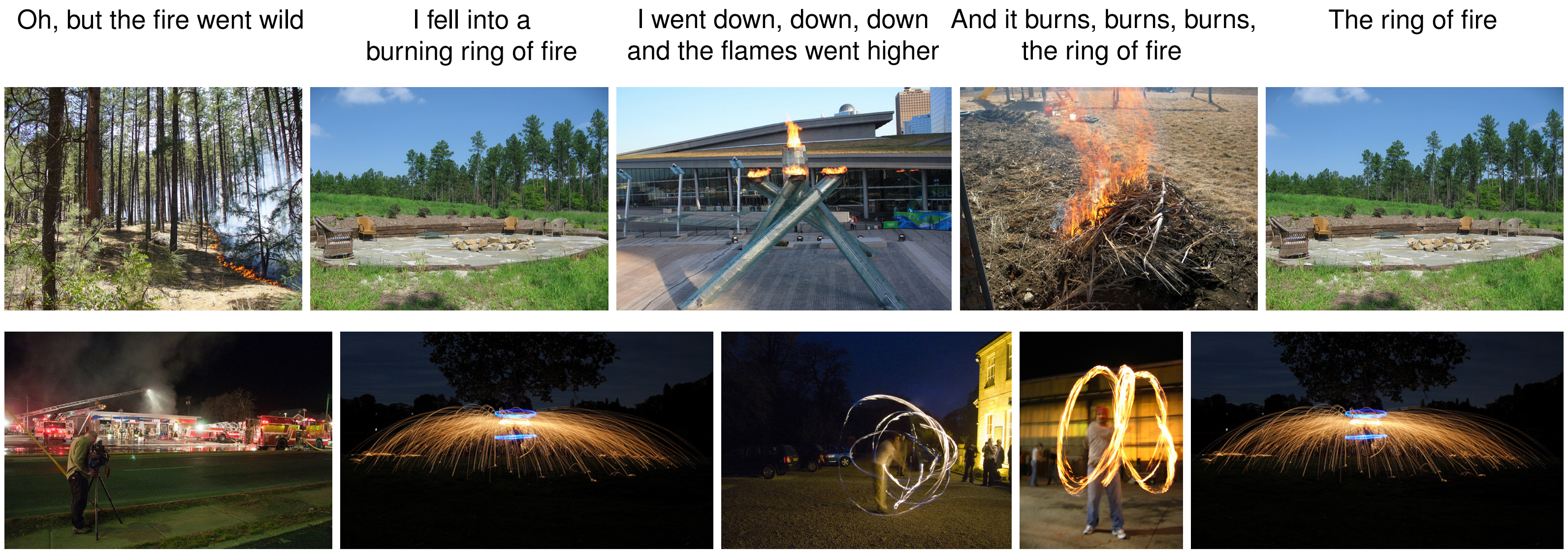

Oh, but the fire went wild
I fell into a burning ring of fire
I went down, down, down and the flames went higher
And it burns, burns, burns, the ring of fire
The ring of fire

Figure B.12: *Song: "Johnny Cash - RING OF FIRE". Styles: "hdr" (top row), "long exposure" (bottom row).*

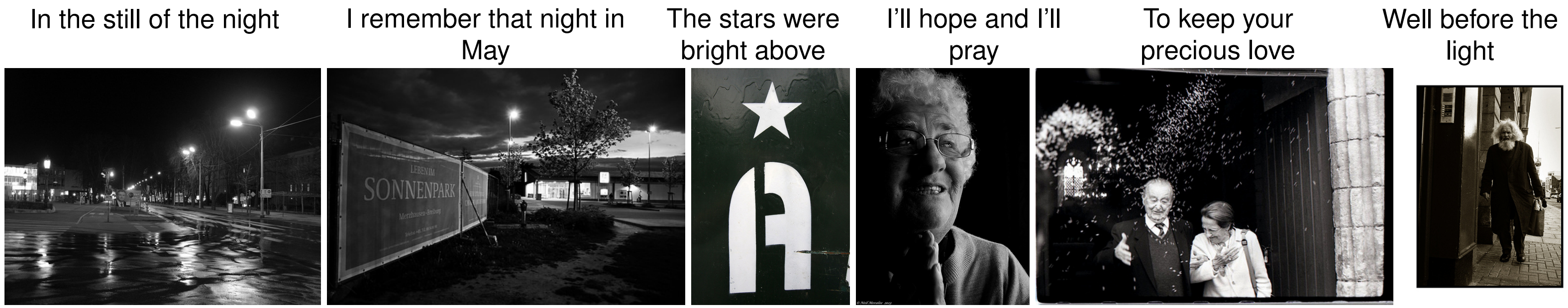

In the still of the night
I remember that night in May
The stars were bright above
I'll hope and I'll pray
To keep your precious love
Well before the light
LEBENDIG
SONNENPARK

Figure B.13: *Song: "The Five Satins - IN THE STILL OF THE NIGHT". Style: "noir".*

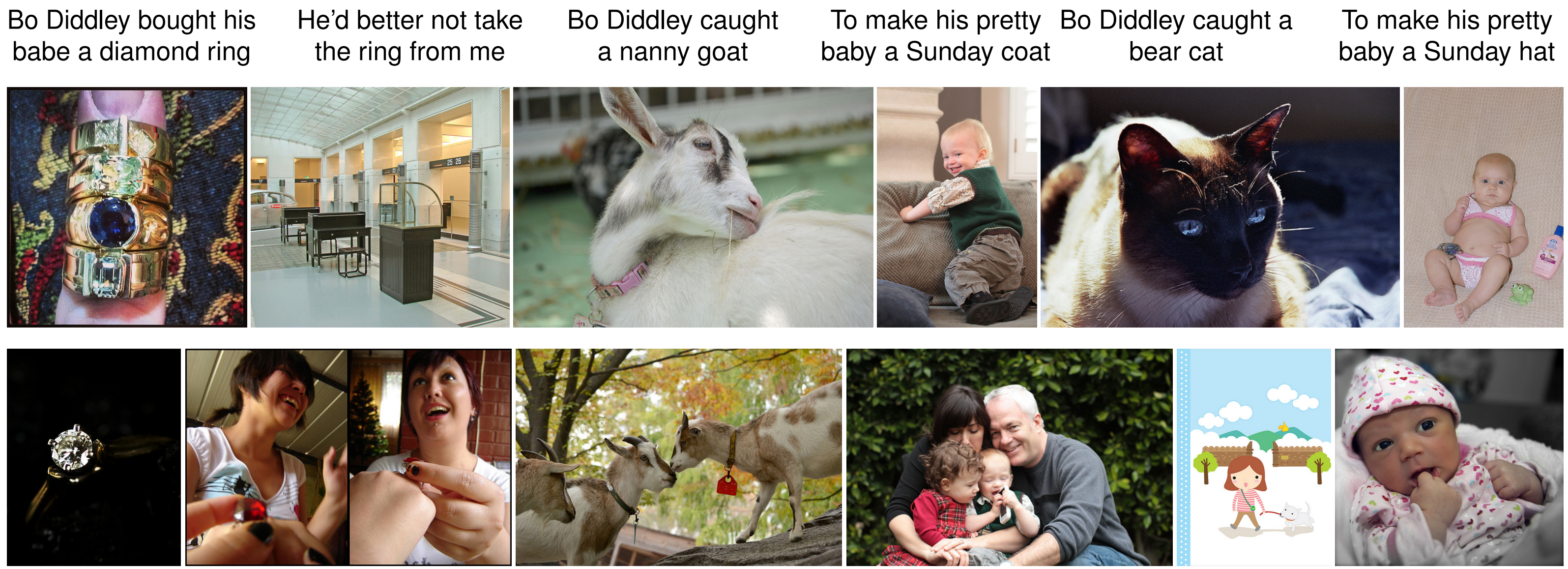

Bo Diddley bought his babe a diamond ring
He'd better not take the ring from me
Bo Diddley caught a nanny goat
To make his pretty baby a Sunday coat
Bo Diddley caught a bear cat
To make his pretty baby a Sunday hat

Figure B.14: *Song: "Bo Diddley: BO DIDDLEY". Styles: "pastel" (top row),"romantic" (bottom row).*

Bibliography

[AAL$^+$15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[ADA$^+$04] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. Interactive Digital Photomontage. *ACM Trans. Graph.*, 23(3):294–302, 2004.

[ADMF83] Givoanni Adorni, Mauro Di Manzo, and Giacomo Ferrari. Natural Language Input for Scene Generation,. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 175–182. Association for Computational Linguistics, 1983.

[AECOK16] Hadar Averbuch-Elor, Daniel Cohen-Or, and Johannes Kopf. Smooth Image Sequences for Data-driven Morphing. *Comput. Graph. Forum*, 35(2):203–213, 2016.

[Aesnd] Aesthetics. Oxford Dictionaries [Online]. `https://en.oxforddictionaries.com/definition/aesthetics`, (n.d.). Accessed: May 3, 2017.

[AEWQ$^+$15] Hadar Averbuch-Elor, Yunhai Wang, Yiming Qian, Minglun Gong, Johannes Kopf, Hao Zhang, and Daniel Cohen-Or. Distilled Collections from Textual Image Queries. *Computer Graphics Forum*, pages 131–142, 2015.

[AMP11] Mohamed Aly, Mario Munich, and Pietro Perona. Indexing in Large Scale Image Collections: Scaling Properties and Benchmark. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 418–425, 2011.

[AVT16] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems (NIPS)*, pages 892–900. Curran Associates, Inc., 2016.

[AZ12] R. Arandjelović and A. Zisserman. Multiple queries for large scale specific object retrieval. In *British Machine Vision Conference (BMVC)*, 2012.

[BBGO11] Alexander M. Bronstein, Michael M. Bronstein, Leonidas J. Guibas, and Maks Ovsjanikov. Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval. *ACM Trans. Graph.*, 30(1):1:1–1:20, 2011.

[BCP00] M. Bréal, N. Cust, and J.P. Postgate. *Semantics: Studies in the Science of Meaning*. W. Heinemann, 1900.

[BDF$^+$03] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching Words and Pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[BF01] Kobus Barnard and David Forsyth. Learning the Semantics of Words and Pictures. *Proc. of ICCV*, 2:408–415, 2001.

[BG04] I.A. Bolshakov and A. Gelbukh. *Computational linguistics: models, resources, applications*. Fondo de Cultura Económica, 2004.

[BGBG62] William S. Baring-Gould and Ceil Baring-Gould. *The Annotated Mother Goose: Nursery Rhymes Old and New*, page 156. Bramhall House, 1962.

[BGL+94] Jane Bromley, Isabelle Guyon, Yann Lecun, Eduard Säckinger, and Roopak Shah. Signature Verification using a "Siamese" Time Delay Neural Network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 737–744, 1994.

[BI07] Oren Boiman and Michal Irani. Detecting Irregularities in Images and in Video. *International Journal of Computer Vision*, 74:17–31, 2007.

[BJ05] Kobus Barnard and Matthew Johnson. Word Sense Disambiguation with Pictures. *Artificial Intelligence*, 167(1-2):13–30, 2005.

[BJJ+10] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the Symposium on User Interface Software and Technology (UIST)*, pages 333–342. ACM, 2010.

[BKG11] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[BL94] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy & Experimental Psychiatry*, 25(1):49–59, 1994.

[BNL+13] Subhabrata Bhattacharya, Behnaz Nojavanasghari, Dong Liu, Tao Chen, Shih-Fu Chang, and Mubarak Shah. Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos. In *Proc. of MM*, pages 361–364. ACM, 2013.

[BPD06] Soonmin Bae, Sylvain Paris, and Frédo Durand. Two-scale Tone Management for Photographic Look. *ACM Trans. Graph.*, 25(3):637–645, 2006.

[Bré97] Michel Bréal. *Essai de sémantique*. Science des significations. Hachette, 1897.

[Bri95] Eric Brill. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[BSGL15] Dennis R. Bukenberger, Katharina Schwarz, Fabian Groh, and Hendrik P. A. Lensch. Rotoscoping on Stereoscopic Images and Videos. In *Vision, Modeling & Visualization (VMV)*, pages 111–118. Eurographics Association, 2015.

[BSL17] Dennis R. Bukenberger, Katharina Schwarz, and Hendrik P. A. Lensch. Stereo-consistent Contours in Object Space. *Computer Graphics Forum (CGF)*, 2017.

[BSPP13] Nicolas Bonneel, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister. Example-Based Video Color Grading. *ACM Trans. Graph.*, 32(4):39:1–39:12, 2013.

[BTMC04] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10(3/4):349—373, 2004.

[CCQ15] Allan Campbell, Vic Ciesielksi, and A. K. Qin. Feature Discovery by Deep Learning for Aesthetic Analysis of Evolved Abstract Images. In *Proc. of EvoMUSART*, pages 27–38. Springer, 2015.

[CCT+09] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2Photo: Internet Image Montage. *ACM Trans. Graph.*, 28(5):124:1–124:10, 2009.

[CFL+15] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. Palette-based Photo Recoloring. *ACM Trans. Graph.*, 34(4):139:1–139:11, 2015.

[CHL05] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *Proc. of CVPR*, pages 539–546, 2005.

[CHLMT07] Luis Carretié, José A Hinojosa, Sara López-Martín, and Manuel Tapia. An electrophysiological study on the interaction between emotional content and spatial frequency of visual stimuli. *Neuropsychologia*, 45(6):1187–1195, 2007.

[CL96] Jim Cowie and Wendy Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80–91, 1996.

[CMBea10] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and et al. *Delevoping Language Processing Components with GATE Version 6 (a User Guide)*. The University of Sheffield, version 6 edition, 2010.

[CMBT02] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, pages 168–175, 2002.

[COSG+06] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color Harmonization. *ACM Trans. Graph.*, 25(3):624–630, 2006.

[CPZ09] K. Chatfield, J. Philbin, and A. Zisserman. Efficient Retrieval of Deformable Shape Classes using Local Self-Similarities. In *Workshop on Non-rigid Shape Analysis and Deformable Image Alignment, ICCV*, pages 264–271, 2009.

[Crond] Crowdsourcing. Oxford Dictionaries [Online]. `https://en.oxforddictionaries.com/definition/crowdsourcing`, (n.d.). Accessed: October 11, 2017.

[Cry90] David Crystal. *Linguistics*. Penguin language & linguistics. Penguin UK, 1990.

[Cry01] D. Crystal. Linguistics: Overview. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 8948–8954. Pergamon, 2001.

[CS01] Bob Coyne and Richard Sproat. WordsEye: An Automatic Text-to-Scene Conversion System. In *Proceedings of SIGGRAPH*, Annual Conference Series, pages 487–496. ACM, 2001.

[CSN05] Youngha Chang, Suguru Saito, and Masayuki Nakajima. Example-based Color Transformation for Image and Video. In *Proc. of the International Conference on Computer Graphics and Interactive Techniques (GRAPHITE)*, pages 347–353, 2005.

[CUF16] Hang Chu, Raquel Urtasun, and Sanja Fidler. Song From PI: A Musically Plausible Network for Pop Music Generation. *CoRR*, abs/1611.03477, 2016.

[Cun05] Hamish Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics*, pages 665–677, 2005.

[CUS04] Youngha Chang, Keiji Uchikawa, and Suguru Saito. Example-based Color Stylization Based on Categorical Perception. In *Proc. of the Symposium on Applied Perception in Graphics and Visualization (APGV)*, pages 91–98. ACM, 2004.

[CW96] Sharon Rose Clay and Jane Wilhelms. Put: Language-Based Interactive Manipulation of Objects. *IEEE Computer Graphics and Applications*, 16(2):31–39, 1996.

[CZM$^+$11] Ming-Ming Cheng, Guo-Xin Zhang, N. J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global Contrast Based Salient Region Detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 409–416, Washington, DC, USA, 2011. IEEE Computer Society.

[Dar72] Charles Darwin. *The expression of the emotions in man and animals*. John Murray, London, 1872.

[DBFF02] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object Recognition As Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 97–112. Springer-Verlag, 2002.

[DBHM03] D.J. Duke, P.J. Barnard, N. Halper, and M. Mellin. Rendering and Affect. *Comp. Graph. Forum*, 22(3):359–368, 2003.

[DBLFF10] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What Does Classifying More Than 10,000 Image Categories Tell Us? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–84, Berlin, Heidelberg, 2010. Springer-Verlag.

[DCF$^+$15] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language Models for Image Captioning: The Quirks and What Works. In *ACL*, pages 100–105, 2015.

[DCsF15] Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 1486–1494. Curran Associates, Inc., 2015.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-scale Hierarchical Image Database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE Computer Society, 2009.

[DHK13] Li Deng, Geoffrey E. Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: an overview. In *ICASSP*, pages 8599–8603. IEEE, 2013.

[DJLW06] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proc. of ECCV*, pages 288–301. Springer, 2006.

[DJLW08] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Comput. Surv.*, 40(2):5:1–5:60, 2008.

[DJS+09] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of GIST Descriptors for Web-scale Image Search. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pages 19:1–19:8, 2009.

[DLW05] Ritendra Datta, Jia Li, and James Z. Wang. Content-based Image Retrieval: Approaches and Trends of the New Age. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR)*, pages 253–262. ACM, 2005.

[dMMM06] Marie-Catherine de Marneffe, Bill Maccartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, 2006.

[DNSG07] Sylvain Delplanque, Karim N'diaye, Klaus Scherer, and Didier Grandjean. Spatial frequencies or emotional effects? A systematic measure of spatial frequencies for IAPS pictures by a discrete wavelet analysis. *Journal of Neuroscience Methods*, 165(1):144–150, 2007.

[DOB11] Sagnik Dhar, Vicente Ordonez, and Tamara L. Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *Proc. of CVPR*, pages 1657–1664, 2011.

[DSG+12] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What Makes Paris Look Like Paris? *ACM Trans. Graph.*, 31(4):101:1–101:9, 2012.

[EEVG+15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[EGW+10] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[EHBA09] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. PhotoSketch: A Sketch Based Image Query and Compositing System. In *SIGGRAPH 2009: Talks*, pages 60:1–60:1. ACM, 2009.

[Ekm99] Paul Ekman. *Basic Emotions*, chapter 3. John Wiley & Sons, Ltd., 1999.

[Ekm16] Paul Ekman. What Scientists Who Study Emotion Agree About. *Perspectives on Psychological Science*, 11(1):31–34, 2016.

[Emond] Emotion. Oxford Dictionaries [Online]. `https://en.oxforddictionaries.com/definition/emotion`, (n.d.). Accessed: May 3, 2017.

[Emp57] William Empson. *7 Types of Ambiguity: A Study of Its Effect in English Verse*. Meridian Books, 1957.

[FAR07] Raanan Fattal, Maneesh Agrawala, and Szymon Rusinkiewicz. Multiscale Shape and Detail Enhancement from Multi-light Image Collections. *ACM Trans. Graph.*, 26(3), 2007.

[Fel98] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[FFFP04] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *IEEE CVPR Wksp. of Generative Model Based Vision (WGMBV)*, 12:178, 2004.

[FFGG+10] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building Rome on a Cloudless Day. In *ECCV*, pages 368–381, 2010.

[FGI+15] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From Captions to Visual Concepts and Back. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482. IEEE Computer Society, 2015.

[FHS+10] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every Picture Tells a Story: Generating Sentences from Images. In *Proc. of ECCV*, pages 15–29. Springer-Verlag, 2010.

[FHX+14] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Shaogang Gong, and Yuan Yao. Interestingness Prediction by Robust Learning to Rank. In *Proc. of ECCV*, pages 488–503, 2014.

[FK64] W. N. Francis and H. Kucera. *Brown corpus manual: Manual of information to accompany a standard corpus of present day edited American English*. Brown University, Providence, Rhode Island, 1964.

[FK82] W. N. Francis and H. Kučera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, 1982.

[FK15] Jacob John Foley and Paul Kwan. *Feature Extraction in Content-Based Image Retrieval*, pages 5897–5905. IGI Global, 3rd edition, 2015.

[FPY$^+$16] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proc. of EMNLP*, pages 457–468, 2016.

[FSU13] Sanja Fidler, Abhishek Sharma, and Raquel Urtasun. A Sentence Is Worth a Thousand Pixels. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1995–2002, Washington, DC, USA, 2013. IEEE Computer Society.

[GCSS06] Dan B Goldman, Brian Curless, David Salesin, and Steven M. Seitz. Schematic storyboarding for video visualization and editing. *ACM Trans. Graph.*, 25(3):862–871, 2006.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proc. of CVPR*, pages 580–587, 2014.

[GEB15] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture Synthesis Using Convolutional Neural Networks. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 262–270. Curran Associates, Inc., 2015.

[GEB16a] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423. IEEE, 2016.

[GEB$^+$16b] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling Perceptual Factors in Neural Style Transfer. *CoRR*, abs/1611.07865, 2016.

[GGR$^+$13] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The Interestingness of Images. In *Proc. of ICCV*, pages 1633–1640, 2013.

[GHP07] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report 7694, California Institute of Technology, 2007.

[GLGW12] Yanwen Guo, Ming Liu, Tingting Gu, and Wenping Wang. Improving Photo Composition Elegantly: Considering Image Similarity During Composition. *Comput. Graph. Forum*, 31:2193–2202, 2012.

[GPAM$^+$14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[GRCH96] R. Gaizauskas, P. Rodgers, H. Cunningham, and K. Humphreys. GATE User Guide. `http://gate.ac.uk/`, 1996.

[GSN07] David Gavilan, Suguru Saito, and Masayuki Nakajima. Sketch-to-collage. In *ACM SIGGRAPH 2007 Posters*. ACM, 2007.

[GT06] Marieke Guy and Emma Tonkin. Folksonomies, tidying up tags? *D-Lib Magazine*, 12(1), 2006.

[GVSJ04] Smuel D. Gosling, Simine Vazire, Sanjay Srivastava, and Oliver P. John. Should We Trust Web-based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, 59(2):93–104, 2004.

[HA15] Elad Hoffer and Nir Ailon. Deep Metric Learning Using Triplet Network. In *Proc. of SIMBAD*, pages 84–92, 2015.

[HB10] Jeffrey Heer and Michael Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *ACM Human Factors in Computing Systems (CHI)*, pages 203–212, 2010.

[HDL11] Johannes Hanika, Holger Dammertz, and Hendrik P. A. Lensch. Edge-Optimized À-Trous Wavelets for Local Contrast Enhancement with Robust Denoising. *Comp. Graph. Forum*, 30(7):1879–1886, 2011.

[HDY+12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[HE07] James Hays and Alexei A. Efros. Scene Completion Using Millions of Photographs. *ACM Trans. Graph.*, 26(3), 2007.

[Hei10] Andreas M. Hein. Identification and Bridging of Semantic Gaps in the Context of Multi-Domain Engineering. In *Forum on Philosophy, Engineering & Technology*, 2010.

[Hep00] Mark Hepple. Independence and commitment: assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2000.

[HFB+09] Tom Haber, Christian Fuchs, Philippe Bekaert, Hans-Peter Seidel, Michael Goesele, and Hendrik P. A. Lensch. Relighting Objects from Image Collections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–634, 2009.

[HFM+16] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual Storytelling. *CoRR*, abs/1604.03968, 2016.

[HGO+10] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. J. Guibas. Image Webs: Computing and Exploiting Connectivity in Image Collections. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3432–3439. IEEE Computer Society, 2010.

[HLK+17] Ruizhen Hu, Wenchao Li, Oliver Van Kaick, Hui Huang, Melinos Averkiou, Daniel Cohen-Or, and Hao Zhang. Co-Locating Style-Defining Elements on 3D Shapes. *ACM Trans. Graph.*, 36(3):33:1–33:15, 2017.

[HSA07] C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*, 2007.

[HSGL11] Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Non-rigid Dense Correspondence with Applications for Image Enhancement. *ACM Trans. Graph.*, 30(4):70:1–70:10, 2011.

[HSGL13] Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Optimizing Color Consistency in Photo Collections. *ACM Trans. Graph.*, 32(4):38:1–38:10, 2013.

[HXR$^+$16] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural Language Object Retrieval. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015.

[IXTO11] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Proc. of CVPR*, pages 145–152, 2011.

[Jac] Sebastian Jacobitz. http://petapixel.com/2016/08/08/understanding-basic-aesthetics-photography/. Accessed: October 27, 2016.

[JAFF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proc. of ECCV*, pages 694–711, 2016.

[JDA$^+$11] M. K. Johnson, K. Dale, S. Avidan, H. Pfister, W. T. Freeman, and W. Matusik. CG2Real: Improving the Realism of Computer Generated Images Using a Large Collection of Photographs. *IEEE Transactions on Visualization and Computer Graphics*, 17(9):1273–1285, 2011.

[JDF$^+$11] Dhiraj Joshi, Ritendra Datta, Elena A. Fedorovskaya, Quang-Tuan Luong, James Ze Wang, Jia Li, and Jiebo Luo. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.*, 28(5):94–115, 2011.

[JGRF10] Time Johnson, Pierre Georgel, Rahul Raguram, and Jan-Michael Frahm. Fast Organization of Large Photo Collections using CUDA. In *Wksp. on Comp. Vis. on GPUs, ECCV*, 2010.

[JKS$^+$15] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Image Retrieval using Scene Graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678. IEEE Computer Society, 2015.

[JM08] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall International, 2nd edition, 2008.

[JWL06] Dhiraj Joshi, James Z. Wang, and Jia Li. The Story Picturing Engine—a System for Automatic Text Illustration. *TOMCCAP*, pages 68–89, 2006.

[KDSH14] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What Makes an Image Popular? In *Proc. of WWW*, pages 867–876. ACM, 2014.

[KF67] Henry Kučera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, 1967.

[KFF15] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, pages 3128–3137, 2015.

[KHLS13] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-Scale Video Summarization Using Web-Image Priors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2698–2705, Washington, DC, USA, 2013. IEEE Computer Society.

[KLB$^+$14] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What Are You Talking About? Text-to-Image Coreference. In *CVPR*, pages 3558–3565, 2014.

[KM03] Dan Klein and Christopher D Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3–10. MIT Press, 2003.

[KMS15a] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. Joint Photo Stream and Blog Post Summarization and Exploration. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3081–3089, 2015.

[KMS15b] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. Ranking and Retrieval of Image Sequences from Multiple Paragraph Queries. In *Proceedings of the International Conference on Computer Vision (CVPR)*, pages 1993–2001, 2015.

[Kol06] Vladimir Kolmogorov. Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1568–1583, 2006.

[KPD$^+$11] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, pages 1601–1608, 2011.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of NIPS*, pages 1097–1105, 2012.

[KSL$^+$16] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *Proc. of ECCV*, pages 662–679, 2016.

[KSX14] Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction. In *CVPR*, pages 4225–4232, 2014.

[KTH$^+$14] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing Image Style. In *BMVC*, 2014.

[KTJ06] Yan Ke, Xiaoou Tang, and Feng Jing. The Design of High-Level Features for Photo Quality Assessment. In *Proc. of CVPR*, pages 419–426, 2006.

[KVMP15] Satwik Kottur, Ramakrishna Vedantam, José M. F. Moura, and Devi Parikh. Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes. *CoRR*, abs/1511.07067, 2015.

[KX13] Gunhee Kim and Eric P. Xing. Jointly Aligning and Segmenting Multiple Web Photo Streams for the Inference of Collective Photo Storylines. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 620–627. IEEE Computer Society, 2013.

[KX14] Gunhee Kim and Eric P. Xing. Reconstructing Storyline Graphs for Image Recommendation from Web Community Photos. In *CVPR*, pages 3882–3889, 2014.

[KZG$^+$17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1):32–73, 2017.

[LBC08] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, The Center for Research in Psychophysiology, University of Florida, 2008.

[LC09] C. Li and T. Chen. Aesthetic Visual Quality Assessment of Paintings. *IEEE J. Sel. Topics Signal Process*, 3(2):236–252, 2009.

[LCWCO10] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. Optimizing Photo Composition. *Comput. Graph. Forum*, 29(2):469–478, 2010.

[Lep17] Ernest Lepore. Semantics. `www.britannica.com/science/semantics`, 2017. Accessed: May 3, 2017.

[LFKU14] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2664. IEEE, 2014.

[LG13] Zheng Lu and Kristen Grauman. Story-Driven Summarization for Egocentric Video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721. IEEE Computer Society, 2013.

[LGG12] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering Important People and Objects for Egocentric Video Summarization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353, 2012.

[LGLC10] C. Li, A. Gallagher, A. C. Loui, and T. Chen. Aesthetic quality assessment of consumer photos with faces. In *Proc. of ICIP*, pages 3221–3224. IEEE, 2010.

[LHE$^+$07] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. *ACM Trans. Graph.*, 2007.

[LHLF15] Tianqiang Liu, Aaron Hertzmann, Wilmot Li, and Thomas Funkhouser. Style Compatibility for 3D Furniture Models. *ACM Trans. Graph.*, 34(4):85:1–85:9, 2015.

[Liu12] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[LJ93] Barbara Landau and Ray Jackendoff. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 1993.

[LKWS16] Zhaoliang Lun, Evangelos Kalogerakis, Rui Wang, and Alla Sheffer. Functionality Preserving Shape Style Transfer. *ACM Trans. Graph.*, 35(6):209:1–209:14, 2016.

[LLC10] Congcong Li, Alexander C. Loui, and Tsuhan Chen. Towards Aesthetics: A Photo Quality Assessment and Photo Selection System. In *Proc. of MM*, pages 827–830. ACM, 2010.

[LLC12] Kuo-Yen Lo, Keng-Hao Liu, and Chu-Song Chen. Assessment of photo aesthetics with efficiency. In *Proc. of ICPR*, pages 2186–2189, 2012.

[LLJ$^+$14] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. RAPID: Rating Pictorial Aesthetics Using Deep Learning. In *Proc. of MM*, pages 457–466. ACM, 2014.

[LLS$^+$15] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z. Wang. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *Proc. of ICCV*, pages 990–998, 2015.

[LMB$^+$14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[Low99] David G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proc. of Int. Conf. on Comp. Vis. (ICCV)*, pages 1150–1157. IEEE Computer Society, 1999.

[LRM15] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN Models for Fine-grained Visual Recognition. In *Proc. of ICCV*, pages 1449–1457, 2015.

[LRT$^+$14] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient Attributes for High-level Understanding and Editing of Outdoor Scenes. *ACM Trans. Graph.*, 33(4):149:1–149:11, 2014.

[LS15] A. Lindner and S. SÃ¼sstrunk. Semantic-Improved Color Imaging Applications: It Is All About Context. *IEEE Trans. Multimedia*, 17(5):700–710, 2015.

[LSBS12] Albrecht Lindner, Appu Shaji, Nicolas Bonnier, and Sabine Süsstrunk. Joint Statistical Analysis of Images and Keywords with Applications in Semantic Image Enhancement. In *Proc. of MM*, pages 489–498. ACM, 2012.

[LSFF09] L. J. Li, R. Socher, and L. Fei-Fei. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. In

Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pages 2036–2043. IEEE, 2009.

[LUB+16] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval. *ACM Compututing Surveys (CSUR)*, 49(1):14:1–14:39, 2016.

[LWT11] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based Photo Quality Assessment. In *Proc. of ICCV*, pages 2206–2213, 2011.

[LZT10] Shuo Li, Yu-Jin Zhang, and Huachun Tan. Discovering latent semantic factors for emotional picture categorization. In *IEEE International Conference on Image Processing (ICIP)*, pages 1065–1068, 2010.

[Mas51] Frank J. Massey. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[MBGS15] Ricardo Martin-Brualla, David Gallup, and Steven M. Seitz. Time-lapse Mining from Internet Photos. *ACM Trans. Graph.*, 34(4):62:1–62:8, 2015.

[McA93] D. P. McAdams. *The Stories We Live by: Personal Myths and the Making of the Self*. W. Morrow and Company, 1993.

[McA01] Dan P. McAdams. The Psychology of Life Stories. *Review of General Psychology*, 5(2):100–122, 2001.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013.

[MF14] Mateusz Malinowski and Mario Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 1682–1690, 2014.

[MFL+05] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. Emotional category data on images from the International Affective Picture System. *Behav. Res. Methods*, 37(4):626–630, 2005.

[MH10] Jana Machajdik and Allan Hanbury. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proc. of the ACM International Conference on Multimedia (MM)*, pages 83–92, 2010.

[Mil95] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[MJHL10] B. Thomee Mark J. Huiskes and Michael S. Lew. New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative. In *Proc. of the 2010 ACM Int. Conf. on Multimedia Information Retrieval (MIR)*, pages 527–536. ACM, 2010.

[MKM+94] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the*

Workshop on Human Language Technology (HLT), pages 114–119. Association for Computational Linguistics, 1994.

[MKNM95] B. M. Mehtre, M. S. Kankanhalli, A. D. Narasimhalu, and G. C. Man. Color matching for image retrieval. *Pattern Recogn. Lett.*, 16(3):325–331, 1995.

[MML11] Regan L. Mandryk, David Mould, and Hua Li. Evaluation of Emotional Response to Non-photorealistic Images. In *Proc. of the ACM Symposium on Non-Photorealistic Animation and Rendering (NPAR)*, pages 7–16, 2011.

[MMP12] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *Proc. of CVPR*, pages 2408–2415, 2012.

[MMS93] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330, 1993.

[MOO10] Anush K. Moorthy, Pere Obrador, and Nuria Oliver. Towards Computational Models of the Visual Aesthetic Appeal of Consumer Videos. In *Proc. of ECCV*, pages 1–14, 2010.

[MPLC11] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the Aesthetic Quality of Photographs Using Generic Image Descriptors. In *Proc. of ICCV*, pages 1784–1791, 2011.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[MS01] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2001.

[MSC$^+$13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, pages 3111–3119, 2013.

[MW47] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[NN05] L. Neumann and A. Neumann. Color Style Transfer Techniques using Hue, Lightness and Saturation Histogram Matching. In *Comp. Aesthetics in Graphics, Visualization and Imaging*, pages 111–122, 2005.

[NNRS15] Chuong H. Nguyen, Oliver Nalbach, Tobias Ritschel, and Hans-Peter Seidel. Guiding Image Manipulations using Shape-appearance Subspaces from Co-alignment of Image Collections. *Computer Graphics Forum (Proc. Eurographics 2015)*, 34(2), 2015.

[NOSS11] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic Quality Classification of Photographs Based on Color Harmony. In *Proc. of CVPR*, pages 33–40, 2011.

[OAH11] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Color Compatibility from Large Datasets. *ACM Trans. Graph.*, 30(4):63:1–63:12, 2011.

[OBLS14] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. In *Proc. of CVPR*, pages 1717–1724, 2014.

[oEB98] The Editors of Encyclopædia Britannica. Ambiguity. `www.britannica.com/topic/ambiguity`, 1998. Accessed: August 1, 2017.

[OKB11] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, pages 1143–1151, 2011.

[OST57] C.E. Osgood, G.J. Suci, and P.H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana, USA, 1957.

[OT01] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. Journal of Comp. Vis.*, 42:145–175, 2001.

[OVB$^+$15] Fumio Okura, Kenneth Vanhoey, Adrien Bousseau, Alexei A. Efros, and George Drettakis. Unifying Color and Texture Transfer for Predictive Appearance Manipulation. In *Proc. of the Eurographics Symposium on Rendering (EGSR)*, pages 53–63. Eurographics Association, 2015.

[PH94] David D. Palmer and Marti A. Hearst. Adaptive Sentence Boundary Disambiguation. In *Proceedings of the Conference on Applied Natural Language Processing (ANLC)*, pages 78–83. Association for Computational Linguistics, 1994.

[PH97] David D. Palmer and Marti A. Hearst. Adaptive Multilingual Sentence Boundary Disambiguation. *Computational Linguistics*, 23(2):241–267, 1997.

[PH12] Genevieve Patterson and James Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2751–2758. IEEE Computer Society, 2012.

[Pic00] Rosalind W. Picard. Toward Computers that Recognize and Respond to User Emotion. *IBM Systems Journal*, 39, 2000.

[PK15] Cesc Chunseong Park and Gunhee Kim. Expressing an Image Stream with a Sequence of Natural Sentences. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pages 73–81. MIT Press, 2015.

[PKD05] François Pitié, Anil C. Kokaram, and Rozenn Dahyot. N-Dimensional Problility Density Function Transfer and its Application to Colour Transfer. In *International Conference on Computer Vision (ICCV)*, pages 1434–1439. IEEE Computer Society, 2005.

[PMY$^+$16] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[Por80] Martin F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[PTSSK16] Jaesik Park, Yu-Wing Tai, Sudipta N. Sinha, and In So Kweon. Efficient and Robust Color Consistency for Community Photo Collections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 430–438. IEEE, 2016.

[QHTM14] X. Qian, X. S. Hua, Y. Y. Tang, and T. Mei. Social Image Tagging With Diverse Semantics. *IEEE Transactions on Cybernetics*, 44(12):2493–2508, 2014.

[RAGS01] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color Transfer between Images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[RAY$^+$16] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1060–1069. JMLR.org, 2016.

[RBHB06] Carsten Rother, Lucas Bordeaux, Youssef Hamadi, and Andrew Blake. AutoCollage. *ACM Trans. Graph.*, 25(3):847–852, 2006.

[RDS$^+$15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015.

[RFF12] Olga Russakovsky and Li Fei-Fei. Attribute Learning in Large-scale Datasets. In *Proceedings of the European Conference on Trends and Topics in Computer Vision (ECCV)*, pages 1–14, Berlin, Heidelberg, 2012. Springer-Verlag.

[RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[RKKB05] Carsten Rother, Sanjiv Kumar, Vladimir Kolmogorov, and Andrew Blake. Digital Tapestry. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–596. IEEE Computer Society, 2005.

[RP11] Erik Reinhard and Tania Pouli. Colour Spaces for Colour Transfer. In *Proc. of the International Conference on Computational Color Imaging (CCIW)*, pages 1–15. Springer, 2011.

[Rus80] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[RWM89] James A. Russell, Anna Weiss, and Gerald A. Mendelsohn. Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, 57(3):493–502, 1989.

[SBC06] Maria Shugrina, Margrit Betke, and John Collomosse. Empathic Painting: Interactive Stylization Through Observed Emotional State. In *Proc. of the ACM Symposium on Non-photorealistic Animation and Rendering (NPAR)*, pages 87–96, 2006.

[SBL17] Katharina Schwarz, Tamara L. Berg, and Hendrik P. A. Lensch. Auto-Illustrating Poems and Songs with Style. In *Computer Vision – ACCV 2016*, pages 87–103. Springer International Publishing, 2017.

[SBNNB11] Aixin Sun, Sourav S. Bhowmick, Khanh Tran Nam Nguyen, and Ge Bai. Tag-based Social Image Retrieval: An Empirical Evaluation. *Journal of the American Society for Information Science and Technology*, 62(12):2364–2381, 2011.

[Sch10] Katharina Schwarz. Text-to-Video: Coherent Image Aggregation from Semantic Text Analysis. Master's thesis, Ulm University, 2010.

[SDBR15] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*, 2015.

[Semnda] Semantics. Dictionary.com Unabridged [Online]. `http://www.dictionary.com/browse/semantics`, (n.d.). Accessed: May 1, 2017.

[Semndb] Semantics. Merriam-Webster Learner's Dictionary [Online]. `http://www.learnersdictionary.com/definition/semantics`, (n.d.). Accessed: May 1, 2017.

[Semndc] Semantics. Merriam-Webster [Online]. `https://www.merriam-webster.com/dictionary/semantics`, (n.d.). Accessed: May 1, 2017.

[Semndd] Semantics. Oxford Dictionaries [Online]. `https://en.oxforddictionaries.com/definition/semantics`, (n.d.). Accessed: May 1, 2017.

[Semnde] Semantics. The American Heritage Dictionary of the English Language [Online]. `https://ahdictionary.com/word/search.html?q=semantics`, (n.d.). Accessed: May 3, 2017.

[SF11] Mohammad Amin Sadeghi and Ali Farhadi. Recognition Using Visual Phrases. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1745–1752. IEEE Computer Society, 2011.

[SFFL17] Katharina Schwarz, Christian Fuchs, Manuel Finckh, and Hendrik P. A. Lensch. EmoTune - Changing Emotional Response to Images. *Color and Imaging Conference (CIC)*, 2017(25):317–323, 2017.

[SFHA14] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 810–817. European Language Resources Association (ELRA), 2014.

[SHL12] Katharina Schwarz, Tobias Häußler, and Hendrik P.A. Lensch. An Efficient Parallel Strategy for Matching Visual Self-similarities in Large Image Databases. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pages 281–290. Springer, 2012.

[SI07] Eli Shechtman and Michal Irani. Matching Local Self-Similarities across Images and Videos. In *Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. IEEE, 2007.

[Sim06] David R. Simmons. The association of colours with emotions: A systematic approach. *Journal of Vision*, 6(6), 2006.

[SKC+15] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Workshop on Vision and Language (VL)*, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Proc. of CVPR*, pages 1–9, 2015.

[SMGE11] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven Visual Similarity for Cross-domain Image Matching. *ACM Trans. Graph.*, 30(6):154:1–154:10, 2011.

[SMJ07] R. A. Stevenson, J. A. Mikels, and T. W. James. Characterization of the Affective Norms for English Words by discrete emotional categories. *Behav. Res. Methods*, 39(4):1020–1024, 2007.

[Sna09] Keith N. Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections*. PhD thesis, University of Washington, 2009.

[SPDF13] Yichang Shih, Sylvain Paris, Frédo Durand, and William T. Freeman. Data-driven Hallucination of Different Times of Day from a Single Outdoor Photo. *ACM Trans. Graph.*, 32(6):200:1–200:11, 2013.

[SPW+13] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

[SRC+10] Katharina Schwarz, Pavel Rojtberg, Joachim Caspar, Iryna Gurevych, Michael Goesele, and Hendrik P. A. Lensch. Text-to-Video: Story Illustration from Online Photo Collections. In *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)*, pages 402–409. Springer, 2010.

[SS08] Konrad Schindler and David Suter. Object Detection by Global Contour Shape. *Pattern Recogn.*, 41(12):3736–3748, 2008.

[Ss12] Kuo (Djohan) Shiang-shiang. Information about LRC. `http://www.mobile-mir.com/en/HowToLRC.php`, 2012.

[SSDL11] Christian Spika, Katharina Schwarz, Holger Dammertz, and Hendrik P. A. Lensch. AVDT - Automatic Visualization of Descriptive Texts. In *Vision, Modeling & Visualization (VMV)*, pages 129–136. Eurographics Association, 2011.

[SSS06] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo Tourism: Exploring Photo Collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006.

[Stynd] Style. Oxford Dictionaries [Online]. `https://en.oxforddictionaries.com/definition/style`, (n.d.). Accessed: May 3, 2017.

[STZ+16] Xiaoyong Shen, Xin Tao, Chao Zhou, Hongyun Gao, and Jiaya Jia. Regional Foremost Matching for Internet Scene Images. *ACM Trans. Graph.*, 35(6):178:1–178:12, 2016.

[SV04] Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: an Affective

Extension of WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

[SvZ08] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[SWL18] Katharina Schwarz, Patrick Wieschollek, and Hendrik P. A. Lensch. Will People Like Your Image? Learning the Aesthetic Space. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[SWS+00] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[SZ03] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proc. of Int. Conf. on Comp. Vis. (ICCV)*, pages 1470–1477, 2003.

[SZ08] J. Sivic and A. Zisserman. Efficient Visual Search for Objects in Videos. *Proceedings of the IEEE*, 96(4):548–566, 2008.

[SZ14] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.

[TBS15] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2Movie: Aligning Video Scenes With Book Chapters. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[TD16] Flora Ponjou Tasse and Neil Dodgson. Shape2Vec: Semantic-based Descriptors for 3D Shapes, Sketches and Images. *ACM Trans. Graph.*, 35(6):208:1–208:12, 2016.

[TFF08] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[Tho14] Bart Thomee. Yahoo! Webscope dataset YFCC-100M. `http://labs.yahoo.com/Academic_Relations`, 2014.

[TLZ+04] Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. Classification of Digital Photos Taken by Photographers or Home Users. In *Proc. of PCM*, pages 198–205, 2004.

[TSF+15] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The New Data and New Challenges in Multimedia Research. *CoRR*, abs/1503.01817, 2015.

[TV08] Johan W. Tangelder and Remco C. Veltkamp. A Survey of Content Based 3D Shape Retrieval Methods. *Multimedia Tools Appl.*, 39(3):441–471, 2008.

[TYS09] Litian Tao, Lu Yuan, and Jian Sun. SkyFinder: attribute-based sky image search. In *SIGGRAPH '09: ACM SIGGRAPH 2009 papers*, pages 1–5, New York, NY, USA, 2009. ACM.

[Vas01] Nuno Vasconcelos. Image Indexing with Mixture Hierarchies. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3–10. IEEE Computer Society, 2001.

[Vas02] Nuno Vasconcelos. Exploiting Group Structure to Improve Retrieval Accuracy and Speed in Image Databases. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 980–983. IEEE, 2002.

[vH10] Alexis van Hurkman. *Color Correction Handbook: Professional Techniques for Video and Cinema*. Peachpit Press, 2010.

[VHW16] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. *CoRR*, abs/1607.08378, 2016.

[VIN15] S. Vijayarani, M. J. Ilamathi, and M. Nithya. Preprocessing Techniques for Text Mining: An Overview. *International Journal Computer Science and Communication Network*, 5(1):7–16, 2015.

[VM94] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394–409, 1994.

[VOPT16] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. Predicting Motivations of Actions by Leveraging Text. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2997–3005. IEEE, 2016.

[VSS04] Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. Developing Affective Lexical Resources. *PsychNology Journal*, 2(1):61–83, 2004.

[VTBE15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, pages 3156–3164, 2015.

[VXD+15] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *Proceedings the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, pages 1494–1504. Association for Computational Linguistics, 2015.

[WC06] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Trans. on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.

[Wil97] Yorick Wilks. *Information Extraction as a Core Language Technology*, pages 1–9. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

[Win71] T. Winograd. Procedures as representation for data in a computer program for understanding natural language. Technical Report MAC AI-TR-84, MIT, Cambridge, 1971.

[WJC12] Xiaohui Wang, Jia Jia, and Lianhong Cai. Affective Image Adjustment with a Single Word. *The Visual Computer*, 29:1121–1133, 2012.

[Wor10] Princeton University "About WordNet". WordNet. `http://wordnet.princeton.edu`, 2010.

[Wun96] Wilhelm Wundt. *Grundriss der Psychologie*. Leipzig: Engelmann, 1896.

[WYX11] Baoyuan Wang, Yizhou Yu, and Ying-Qing Xu. Example-based Image Color and Tone Style Enhancement. *ACM Trans. Graph.*, 30(4):64:1–64:12, 2011.

[XBK$^+$15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, pages 2048–2057, 2015.

[XHE$^+$10] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*, pages 3485–3492. IEEE Computer Society, 2010.

[XKS15] Bo Xiong, Gunhee Kim, and Leonid Sigal. Storyline Representation of Egocentric Videos with an Applications to Story-Based Search. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4525–4533, Washington, DC, USA, 2015. IEEE Computer Society.

[XWL$^+$08] Xuemiao Xu, Liang Wan, Xiaopei Liu, Tien-Tsin Wong, Liansheng Wang, and Chi-Sing Leung. Animating Animal Motion from Still. *ACM Trans. Graph.*, 27(5):117:1–117:8, 2008.

[YFU12] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–709. IEEE Computer Society, 2012.

[YJW$^+$16] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image Captioning with Semantic Attention. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[YTC$^+$15] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing Videos by Exploiting Temporal Structure. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[YvGR$^+$08] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek. Emotional valence categorization using holistic image features. In *IEEE International Conference on Image Processing (ICIP)*, pages 101–104, 2008.

[YWHZ11] Kuiyuan Yang, Meng Wang, Xian-Sheng Hua, and Hong-Jiang Zhang. *Tag-Based Social Image Search: Toward Relevant and Diverse Results*, pages 25–45. Springer London, 2011.

[Zip49] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.

[ZJC11] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image Retrieval with Geometry-Preserving Visual Phrases. In *Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*, pages 809–816. IEEE, 2011.

[ZK15] Sergey Zagoruyko and Nikos Komodakis. Learning to Compare Image Patches via Convolutional Neural Networks. In *Proc. of CVPR*, pages 4353–4361, 2015.

[ZKZ+15] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *arXiv preprint arXiv:1506.06724*, 2015.

[ZP13] C. L. Zitnick and Devi Parikh. Bringing Semantics into Focus Using Visual Abstraction. In *CVPR*, pages 3009–3016, 2013.

[ZPV13] C. L. Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the Visual Interpretation of Sentences. In *ICCV*, pages 1681–1688, 2013.

[ZXL+16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. *CoRR*, abs/1612.03242, 2016.

[ZZ10] Mingtian Zhao and Song-Chun Zhu. Sisley the Abstract Painter. In *Proc. of the ACM Symposium on Non-Photorealistic Animation and Rendering (NPAR)*, pages 99–107, 2010.