Zusammenfassung

Digitale Fotos sind für viele Menschen das bevorzugte Mittel, um Erinnerungen an persönliche Ereignisse festzuhalten, sei es die eigene Hochzeit, die ersten Schritte des eigenen Kindes oder eine Urlaubsreise. Das Hauptproblem, das sich für viele ergibt, ist das der sinnvollen Auswahl und Präsentation von Fotos. Eine typische Anwendung ist etwa die Erstellung von Fotobüchern. Eine geeignete Auswahl ist hierbei durch diverse Kriterien bestimmt: Zum einen sollen Bilder qualitativ gut und etwa unscharfe, falsch belichtete oder verwackelte Bilder aussortiert werden. Zum anderen soll ein bestimmtes Ereignis möglichst gut repräsentiert sein, so dass möglichst alle Aspekte auch in der Auswahl der Bilder repräsentiert sein sollen. Schlussendlich sollen die Fotos semantisch sinnvoll strukturiert und visuell ansprechend in einem Fotobuch angeordnet werden.

Diese Dissertation verfolgt einen ganzheitlichen Ansatz, der alle relevanten Prozesse zur Gestaltung eines Fotobuches abdeckt. Dies wird erreicht durch die Entwicklung eines intelligenten Systems zur automatisierten Fotobucherstellung. Der Prezess dabei in die Schritte Analyse, Auswahl, Anreicherung und Layout aufgeteilt, angelehnt an einen manuellen Fotobucherstellungsprozess. Die Entwiclkung dieses Systems ist getrieben durch den Nutzer und sein Bedürfnisse. Hierzu wird untersucht, wie Nutzer Fotos Fotos selektieren und wie sie diese im Fotobuch anordnen. Datenbasis hierzu sind mehrere tausende von von real gestalteten und bestellten Fotobüchern. Ausserdem wird der Zusammenhang zwischen Informationen über verschiedene Formen der Fotonutzung und deren Aussagekraft für die geeignete Auswahl für ein Fotobuch untersucht. Dieses Wissen fliesst in die Entwicklung des automatisierten Fotobucherstellungprozesses ein mit dem Ziel zu Fotobüchern zu gelangen, die sowohl die den Fotos zugrunde liegende Geschichte bestmöglich vermitteln als auch visuell ansprechend gestaltet sind.

Abstract

Digital photos are the primary means for many people to capture the memory to personal events like owns own wedding, the steps of ones own child or a nice holiday trip. The main problems these people are faced with the the meaningful selection and presentation of photos in a time where thousands of photos per person and year are shot due to practically no costs for the single photo. A typical application for these tasks are photobooks. A meaningful selection of photos for a photobook is influenced by many criteria: The photos should be aesthetically pleasing and thus unsharp, taken with the wrong exposure or on other ways bad shots should be avoided. On the other side, the story behind the photos should be reflected as good as possible by the selection. And finally, in the resulting layout of the photobook, photos should be placed in a photobook both structured semantically reasonable and visually pleasing.

In this thesis we follow a holistic approach covering all relevant processes in the creation of digital photobooks and develop a system for the automatic retrieval for and layout of photobooks. We divide the process into the steps of photo analysis, selection, augmentation and layout analog to the manual human photobook design process. The driving factor is the user and the user's needs. This thesis thoroughly analyzes how users select photos and how they arrange these photo in photobooks. This usage analysis by several thousands of real-world photo regarding their structure and semantic and by looking deeper into the relation of typical photo usages and the use of photos in photobooks. We use this knowledge for the development of methods which help to automatically create photobooks from a set of photos which both best convey the underlying stories and are also aesthetically pleasing.

Contents

1 Introduction

Since the broad availability of consumer photo cameras, photos have been a means to capture and preserve important personal moments and to share them with friends and family. Have it been conventional prints in the analogue days, its their digital counterparts on the users' hard-drives today. The digitalization of photos came along with a significant increase of photos shot every day due to the abolition of per print costs and negligible storage costs: People return from a 2 weeks holiday with often over 1000 of photos where it have been only 2 or 3 film rolls in the analogue days. This convenience comes with a price: People are often overwhelmed by the sheer masses of photos and have difficulty to manage their personal media. Thousands of printed photos rest in the darkness and isolation of shoe boxes. With the digital photography it is now pictures dsc2345.jpg to dsc2399.jpg residing in our digital shoebox, e. g., a folder called birthdayParty05. Currently, we are facing a market in which about 20 bn digital photos are taken per year for example in Europe [CC10]. At the same time, we can observe that many digital photos are never viewed nor used again. It is estimated that from all digital images only about 20% are actually printed [CC10]. This is not because we forget about these digital souvenirs. One result of a study [CC10] of CEWE Color[1] is that most users of digital cameras would like to have their photos printed. Why are not more photos (re)used and printed even though it seems to be the customer's wish? One answer is that the way we find and select photos from a large set of photos to print them needs far too much time and effort. This is even more true if the user not only wants to print the photos, but aims at more sophisticated photo products such as photo calendars or collages or even more popular: photobooks.

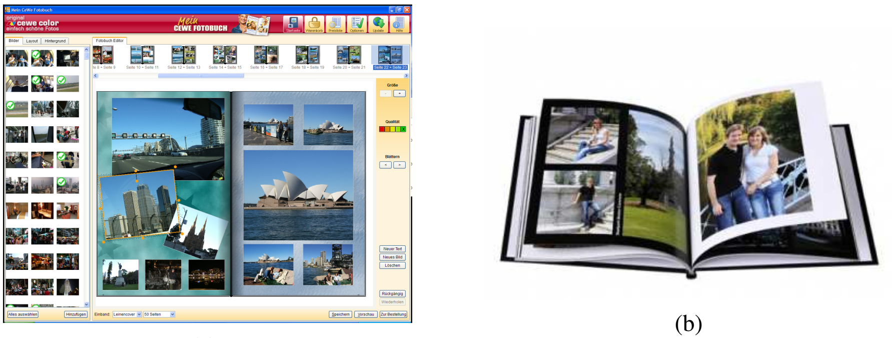
By arranging photos in albums we attempt to preserve for ourselves and convey to others the way we experienced essential events of our lives. For a long time people have been creating such photo albums using prints from analog photos and arranging them carefully in a nice book with scissors and glue. Today these tasks have become digital and service providers like CEWE Color enable users to design photobooks on a home PC. Although basic support for selection and layout is given by these tools, the problems of selection and layout basically remain the same as in the analogue days. They have even become more severe due to the increase in photos and the complex layout options of todays's photobook design tools. The constant increase in ordered digital photobooks every day over the last years [COL12] shows that photobooks are still very much appreciated. But on the other side many users state that they would design a photobook much more often if it would be easier for them. What is really needed are methods to support the user in this task and to both meet the needs of conventional photo users and the photo industry.

Let us take a closer look at the central problems we face when compiling a photobook. These problems can basically divided into the problem of determining *what* is placed into the photobook and *how* this is, aka how the content laid out onto the pages. The first

[1] CEWE color is Europe's leading photo finisher

problem is a retrieval problem: Based on a set of different criteria (features) the photos have to selected which best meet a specific query. For compilation of a photobook such a query is often on a very high semantic level, e.g. *all photos which best present the events and which are meaningful to me*. Such a query demands for features which are both highly subjective and semantically rich and there a good semantic understanding of the photos is very crucial. The research community has put much effort in the development of sophisticated methods to derive semantics from the single photo by content analysis or in combination with contextual information. Therefore, at least to a certain degree, requests can be answered such as *Get me all pictures which showing my daughter* or *Show me all pictures I have taken and Paris which show a building*. However, it is difficult to approach questions like *Get me the 20 pictures out of my collection I really like!* or *Show me pictures of my trip to Paris which are interesting to show them to people not being part of my family*. The central insufficiency we face here today is the fact that digital photos are just a poor reflection of the actual event captured. Digital cameras leave us with a pixel-based copy and some context information of what we experienced. Anything else is gone with the camera releaser, at least it is decoupled from the digital copy of the moment. Even though there is research in content-based image analysis for quite some years [SWS+00] as well as nice photo management tools [Goo06, App06], neither an automatic labeling nor a manual annotation of photos has become a success model. Research in the last years show that one possible key to narrow this so called *semantic gap* in image retrieval comes from outside the image itself[HSL+06, LBB06], by capturing and employing the context it was taken and how it was used. This context seems to be a promising source to be able to answer more semantically complex queries as needed for the selection of photos for the compilation of photobooks.

But being able to retrieve the right photo is often not enough to convey, e.g., the experience of a specific event. It is also important *how* these photos are presented. Rather than just flipping through a series of photos, people tend to carefully arrange important images in various forms such as collages, multimedia presentations, and, one of the most common form, photobooks. On top of the creative options of conventional photobooks, the digital way of of designing provides the user with additional means: photos can be resized, cropped and rotated and the pages can be decorated with textual annotations, backgrounds and additional content like a geographical map. A photobook designed on a home PC can be ordered at a photo service provider as a physical representation of the digital book. Despite, or perhaps because of these additional options in contrast to analogue photobooks, designing a photobook is generally not an easy task. The photo selection problem, as described above, is also true for the compilation of photobooks and is additionally also influenced by the photobook itself: Photobooks, e.g., have a limited number of pages and only a limited number of photos look visually pleasing on a single photobook page. Additionally, some photos might be more qualified to be placed on the front page, others might act as a nice background. These circumstances also directly influence the photo selection problem. Many people do not have the skills or time to carefully arrange photos in a photobook. The photo industry tries to overcome this problem by providing professionally designed layouts. But these layouts can never fully take into account the individual characteristics of the photos. Other people prefer to design their photobooks with the help of such templates and either get frustrated and overwhelmed by the complexity of the design options of the design tool or end up in spending endless hours to finally result in a nice design. What is needed are ways to reduce this complexity and to automate the process of photobook design. The research community has done a couple of steps in this direction [GC04, Atk04, Atk08, KS05, WQS$^+$06]. However, most approaches do either not take into account the specific demands of photobooks or do not base their algorithms on well-established rules and principles for visual layout.



(a)

(b)

Figure 1.1: Commercial tool for the digital design of photobooks and a printed photobook.

Currently there is no holistic solution available in the research community to meet the specific needs for retrieval and layout for photobooks. This thesis aims to fill this gap. The driving factor is the user and the user's needs. This thesis will thoroughly analyze how users select photos and how they arrange these photo in photobooks. The goal is to use this knowledge for the development of methods which help to automatically create photobooks from a set of photos which both best convey the underlying stories and are also aesthetically pleasing.

1.1 Scenarios

Photobooks are one of *the* means for people to archive and preserve their image-based representations of important events in their lifes. The wish to actually compile such a photobook can stem from many different situations and demands. To get a better idea of such different demands and also to potential difficulties and problems, in the following three different and typical scenarios are described in which photobooks are involved. These scenarios are analyzed thoroughly in the next Section to derive the central challenges when targeting at helping the user to design a photobook.

Holiday photobook

Peter comes back from a 3-weeks holiday to Mexico. He has visited many and has made about 3,000 photos during his trip. Back at home he decides keep the memory *to this holiday alive by compiling a photobook documenting the holiday. For this he first selects the photos which are most important to him, which are beautiful shots and which best represent the holiday. He carefully distributes the selected photos over the pages. For this he tries to keep photos together, which belong to the same event and which visually complement other photos on the pages. He has also made a photo from the Great pyramid in Chichén Itzá, but it was a rainy day and so he tries to find a better picture on the internet. He carefully lays out each individual page, highlights more important photos by placing them at more prominent positions and ensures to not occlude important parts of a photo put into the background. Finally he enriches individual events and and photos with textual annotations. The whole process took him several hours. Overall Peter is satisfied with the result. However, he wished some pages would look more balanced in terms of color composition and layout, but he is unsure how to accomplish this. It also would have helped a lot if he did not have to manually select the most important photos.*

Weblog

18 year old Alice is spending a year abroad as an Au-Pair in the USA after having finished school. She meets many new friends, gets a lot of interesting impressions and overall has a great time. As a personal diary for herself and as a way to share her visit with her friends and family at home she documents her experiences and impressions in a weblog. To better illustrate her time she augments her posts with several photos: The trip to the Niagara Falls, photos of her guest family and her guest brother's birthday party are memorable events she wants to share with her beloved at home. Back at home she still has a nice documentation of her year abroad in the form of a web blog. To also have a more tangible memory and also as a gift to her grandparents she decides to compile a photobook from this blog. With the help of a web application she is able to use the already present structure in the blog to automatically create such a photobook which structures the photos and texts in the book, e.g. posts are represented as separate chapters and also indicated in the table of contents. Also, related content, such as photos and maps showing the visited places, is automatically retrieved from the web. The result is a nice online photo album which she can send her friends and family by e-mail and which can also be printed as a physical photobook.

Social Networks

Peter is a very active member of an online social network community. He meets friends there, stays in touch with old friends and shares his personal life online by through photos and posts. Many important events in his life are documented, not only by him, but also by his friends having participated in the same activities. Thus, valuable documents of his life are spread all over his personal social network. This is nice, but he somehow feels the need, because of the constantly changing structure of his network, to preserve these documentary snippets in a more static and safe way by creating a snapshot of his life. With the help of a an application in his social network platform he is able to *automatically create an album of his events. The content for this album is retrieved from his own posts and photos and the ones from his friends. These contents are merged and grouped to the according events and represented as chapters in the resulting album. Not all photos are used for this album , but only the best and most important ones. The retrieved contents are automatically laid out in a pleasing manner in the resulting photo album. Peter is pretty satisfied with the result and shares the album with his friends on the social network. Additionally he orders a physical copy as a photobook at a photo finisher.*

1.2 Challenges

These exemplary scenarios show the need and potential for assistance in the design of photobooks. A closer into these scenarios reveals the central challenges which arise. The overall goal is to enable users to summarize their important moments in the form of a photobook. For this it is important to know what photos are important for a person to be included in a photobook, how he would like them to be organized, and on a more general level: What make a photobook a *good* photobook in the eyes of a user. Thus, we have to first *understand the user*. Additionally, we have to *understand the content* which is placed into a photobook. This not only means the individual photos, which can stem from various sources with varying amount and quality of metadata, but also additional material besides the photos, like text descriptions or other multimedia contents. The main challenge however is to use this knowledge for assisting the user in the design of photobooks.

In the following these three main challenges are described more detailed.

1.2.1 Understanding the user

What a *good* photobook is, and a pleasing design and a reasonable selection of photos, is a highly subjective. Nevertheless, we believe that many patterns regarding selection and design exist which can be applied to many types of events, photobooks and people. These patterns or rules must be the basis to be able to design system for the automatic layout of photobooks. Thus, one challenge is to understand, what people perceive as a good photobook.

One aspect is to understand *what* people like to have included in their photobooks. In the first place this means what photos they select, but also which additional content they, like e.g. text descriptions or other media like geographic maps. Looking into the scenarios it becomes clear that the photos which are the basis for a photobook can origin from a couple of different sources, e.g. from a shared album on a social network or the personal hard-drive. One challenge is to understand, what photos people prefer to have included in a photobook documenting the underlying event and what aspects of the photos and the photos' contexts are important for this. The challenge is to identify the factors that make one photo seem more important than another photo. We believe that these factors can not be read from the content of the photo alone, but are also hidden in the photo's context. This can be potentially highly dependent on the kind of photobook the user is aiming to design, but also highly dependent on the personal needs and preferences of the user.

A second point is to understand *how* a user prefers such content to be placed in a photobook which is both aesthetically pleasing and best conveys the underlying story of the documented event. This involves both understanding the principles and ways people design their photobooks, e.g. how they place the photos in relation to each other and how text is handled, and also what common or *universal* rules exist for visual layout which are specific to photobook layout or can be adapted to it.

1.2.2 Understanding the content

From the scenarios it becomes clear that the content which is placed into photobooks can be very different depending on the context. E.g. a photo taken from the album of social network web site is usually bound to a lot of contextual information like comments, ratings and relations to other photos. On the other hand these photos often have a lack of camera metadata, like the Exif header and sometimes have a poor resolution. This is different to photos on ones personal hard-drive which are usually taken directly from the camera but have a lack of additional contextual information. In addition to that one might argue, that also the content of photos hosted online might be different, as they are often already selected from a larger photo set. To be able to automatically select photos for a photobook it is important to take these different factors into account and to understand the potential differences between these different sources of photos.

Having a good intuitive understanding of the potential content of photobooks another important challenge is to have ways to extract semantic information from the photos which helps to decide if a photo should be candidate for a photobook or not. Results from the research community have shown, that this information cannot come from the content of the photo alone [LBB06]. Thus, the challenge is to determine the factors and information which decide what photo should be selected for a photobook and where and how these can be retrieved from the photos and the photos' contexts.

1.2.3 Assisting the user

The main challenge is to use the gained knowledge about the selection for and the design of a good photobook for the development of a system which assists the user and automates the relevant processes as far as possible. Thus, this challenge can also be divided into the selection and design:

When *selecting* content for a photobook, various aspects have to be taken into account. This involves the design of methods to automatically decide about the perceived quality and importance of a photo which can be affected by many aspects, but also modeling personal preferences of a person and factors that stem from the characteristics of photobooks. E.g. also the number of pages and the amount of free space on a page can affect if a specific is selected or not. The goal is to model the human rating and selection process as good as possible and to be able to automatically select the same photos from a larger set of photos a human would choose for his personal photobook.

Besides this, as outlined in the scenarios, there is strong evidence that people usually not only place photos in their photobooks, but *augment* these photos by text, maps or other material. Supporting and automating this is another challenge we are approaching in this thesis. We aim to investigate which kinds of additional content can contribute to a better photobook and how these contents can automatically retrieved based on information gathered from the user or the photos and the photos' context.

Finally, a big challenge is to automate the *layout* of determined content for a photobook. Based on findings from the analysis of user needs and design rules, we aim to build a system which both takes these aspects into account and supports to convey the underlying presented in a set of photos. For this, we also want to take into account the content and meaning of photos as a person would do when manually designing a photobook.

The goal is to develop a system which automates these processes as far as possible. However, we are aware that this is possible or wanted in all cases. In many cases it might not be possible to automatically determine an optimal solution. Thus, an additional challenge is to develop a system a system which assists the user as far as possible but also to leave enough room to control the processes and to allow to additionally manually alter the results.

1.3 Contributions to the scientific community

The contributions of this thesis can be summarized as follows:

- **Method for user-driven photo selection for digital photobooks** Photo selection or summarization is an actively researched field (a Summarization is given in Sections 3.3.3 and 3.4.1), but different applications have different demands for the selected photos. For photobooks these are different from, e.g. a summarization for a photo website. Besides other factors the selection for a photobook depend on additional aspects like the layout and size of the photobook. To our knowledge there are no works which specifically address the problem of photo selection for photobooks.

- **Thorough analysis of a large set of real-world photobooks** To our knowledge an analysis of how people deal with and design digital photobooks has not been done until now, at least not on a large scale as done in this thesis. This is surprising as photobooks usually afford a a significant amount of effort and manual interaction with the user and thus bare a lot of potential for understanding how the *average user* deals with his or her digital photographs.

- **Method for user-driven dynamic photobook layout** In the last years, several approaches have been done to automatically design photo collages in general and also photobooks in particular. Some of these approaches focus on the some kind of optimization scheme to, e.g. reduce the amount of white space [WQS$^+$06] but do not really take the visual impression into account. Others try to take aesthetic principles into account, but still do base on a fixed visual template schema [BHW09]. The related work is more thoroughly analyzed in Section 3.5, but the gist is that up to now the user has not really been taken into account when designing a photobook page. In thesis we will therefore analyze appreciated designs of real world photobooks, extract principles and combine this with common aesthetic principles for design and layout and integrate this into the overall process of photobook creation. The goal is to develop a method which better suits the user needs than existing methods today can.

1.4 Thesis Organization

The remainder of this thesis is organized as follows. We start by presenting the overall approach followed in this thesis in Chapter 2. In Chapter 3, we give thorough analysis of related works in the field. In Chapters 5 and 4 we provide an analysis of a large set of existing photobooks and the usage of photos which forms the knowledge base for the development of our photobook creation system. The four parts of this system are described in Chapters 6 to 9, covering the analysis, retrieval, augmentation and layout of digital photobooks. The thesis is concluded by a summary and conclusion and an outlook to potential future work in Chapter 10.

Publications

Excerpts of this thesis have been published in scientific conferences, journals, books, and workshops: [ASB11, BSAT06, BSST07, BSSW07, FS09, RSB10, RSB11, San05, STB08, SBF08, SB09, SBMB10, SRB10, SREB11, SB11b, SB11a, SNN$^+$08].

2 Approach

In the first chapter several challenges have been identified which yield from the overall goal of this thesis: The support for the compilation of digitally authored photobooks. These challenges can be divided into challenges stemming from understanding the users' needs and challenges arising when employing such knowledge to support the compilation of photobooks. This distinction is also reflected in the approach of this thesis.

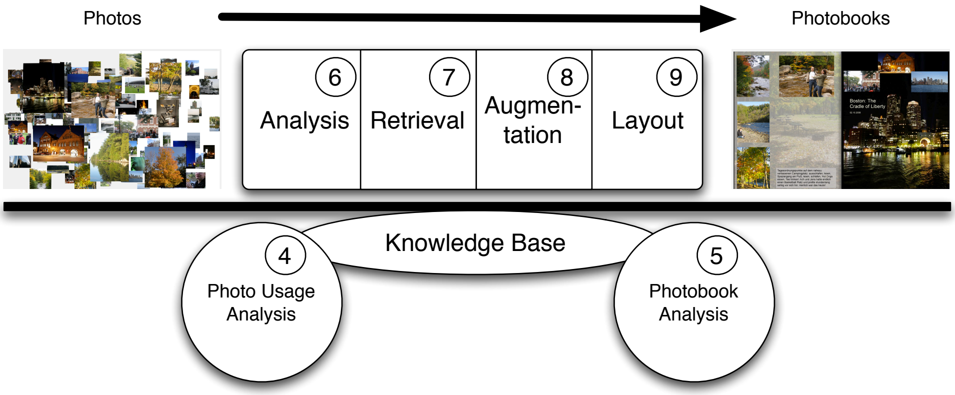


Figure 2.1: System Overview - Numbers designate the respective chapters in this thesis

An overview is depicted in Figure 2.1. From a technical perspective we want to provide a way to transform a collection of photos into a photobook. For this we have identified four more or less consecutive steps: Analysis, Retrieval, Augmentation, and Layout. This system is partly based on a knowledge base fed by the analysis of photo usage and large scale analysis of real-world photobooks. Each of these parts correspond to one chapter in this thesis which is indicated by the respective chapter number in the figure. In the following we will further elaborate on the different parts of our approach.

Photo Usage ④

We believe that an important and rich source for semantically understanding photos is to have a close look at the way they are used by people. How photos are watched, shared and combined reveals much about their meaning to the user. We aim at developing methods to exploit this usage information for the creation of digital photobooks. Thus, we aim to find out, which aspects of photo usage influence or are indicators for the way they are used in photobooks. For this we will review existing research on photo usage and provide a general photo usage model which is able to be the base for a formal metadata usage model. As a proof on concept we will thoroughly analyze a common form of photo usage, the collaborative watching of a slideshow, and its relation to the later selection of photos for a photobook.

Large-scale photobook analysis ⑤

The beauty of photobooks is that raw photo data is combined with a lot of additional contextual information. As usually much effort by the users is put into designing photobooks, it is obvious that one can derive a lot of semantics from the photos by analyzing how they are used in a photobook. Because of out collaboration with Europe's biggest photo finisher, CEWE Color, we have access to a huge amount of photobooks ordered at CEWE Color. These photobooks are anonymized and stored in a structured representation. Therefore, the contained images, textual annotations, and the structure of the photobooks can be analyzed.

The overall goal of the analysis is to derive useful semantics for photobooks as well as for the single photo which can help to design a system for the automatic retrieval of photos and the design of photobooks. With the analysis of photobooks we therefore approach two challenges mentioned in Chapter 1:

1. **Photobook Design:** One goal is to learn more about the general characteristics of photobooks regarding layout and content. The main outcome of this analysis is expected to be verification and also the derivation of specific rules for the layout. Due to the large set of analyzed photobooks we expect to be able to find certain patterns which are perceived as appealing for the majority of users who have designed photobooks. Additionally, we are interested in significant differences in the design for different types of photobooks or differences when comparing photobooks made at different times of the year (for example summer vs. christmas time).

2. **Development of Semantic Derivation Methods:** Besides getting to know more about peoples habits to design photobooks, the analysis is also expected to be helpful in developing methods for the semantic annotation of photos: How and if photos are placed in a photobook can reveal a lot about their meaning to the user. E.g. a photo which is placed more prominent than other photos can be expected as more important to the user than other photos. Thus, photos books and their content can also be used as training date to develop semantic analysis methods with the help of machine learning techniques. In addition to that semantic classifiers for photobooks can also be developed from sufficiently annotated photobooks which can then used to perform more detailed semantic analysis of other photobooks.

The analysis of a large set of photobooks regarding these aspects is done in Chapter 5.

Photo Analysis ⑥

The central goal of this thesis is to support all relevant steps for the compilation of personal photobooks. An important prerequisite for this is a good semantic understanding of the content of the photobook, both for the selection and the layout. In this thesis a hybrid approach is employed which both takes into account the raw content of the photo as well as contextual information which is bundled with the photo. This contextual information can either be directly attached to the photo, for example in its Exif header, or be part of its *social life*, for example how it was used before or where it was retrieved from. This thesis will combine these different kinds of information with the aim to generate more meaningful semantic cues then possible with content or context alone.

The approach for fulfilling this goal is the introduction of a component-based photo analysis system. In this system several photo analysis components are combined. Every component derives one specific aspect of from the single photo or combines information from one or several components to derive higher-level semantics. The derived semantics will be used for retrieval tasks in different sample applications the semi-automatic creation of personal photobooks being the main application. Details of this component-based photo analysis system MetaXa are described in Chapter 6.

Photo Retrieval ⑦⑧

A sufficiently good semantic understanding of photos is the prerequisite to be able to determine the photos from a potentially larger set, which are the most important or preferred ones for a photobook. The second contribution is a system for the automatic determination of content for photos which employs rich semantic information of analyzed photos. This system is driven by the idea that valuable cues for the decision to either select or not select a photo for a photobook are available from multiple sources, for example the photo itself, related photos, the context of photo or what kind of photobook is being authored. Such different aspects are explored and combined appropriately for an overall decision of a specific is chosen for a photobook, or not. The overall goal is to determine a collection of photos, which best represents the underlying event(s). The system is trained and evaluated by real world selections for photobooks.

As a second contribution a method is developed to enrich a photobook with additional content from the web, such as matching photos, text descriptions or maps. Observations of real world photobooks (see Chapter 5) show that many people prefer to enrich their photobooks with additional material which supports to convey the story behind the documented events. Thus, the proposed method will enrich the photo with additional assets from the web based on semantics extracted from the photos. This method is described in Chapter 8.

Photobook Layout ⑨

Besides determining *what* is placed in the photobook it is equally important *how* it is placed. The layout of a photo album on the one side should reflect the underlying structure of the event the photos are documenting and on the other side should be pleasantly designed to attract the viewer of the book. One challenge therefore is to develop methods to appropriately translate the structure of an event into a visual representation. The other challenge is to ensure that the visual attractiveness of the content. For this methods will have to be developed which are based on two important sources: The analysis of real-world photo albums and the adoption of common layout and design principles from the literature.

The main challenge is to represent such layout principles in an automatic layout process which respects the characteristics of each individual photo set. Our approach relies on the use of genetic algorithms which subsequently in several iterations which are rated, combined and and adjusted according to a an extensible set of layout rules. This method is described in Chapter 9.

3 Related Work

Image Retrieval has been a prominent topic in the literature for a long time now and first works in content-based image retrieval range back over 30 years from today [CF80]. This has witnessed the development of several survey papers [SWS$^+$00, AZP96, LSDJ06, DWGW07, LZLM07, Ino09]. Datta [DWGW07] also incorporates works with aspects and ideas going beyond pure content-based image analysis and retrieval alone. A survey specifically targeting at the context-based aspects of image retrieval is given in [LBB06]. The key problem which is approached in most of these works is often referred to as the *semantic gap* in image retrieval. The central goal is to narrow this gap between the raw image data and the human perception of the scenery shown in an image as far as possible. The early works which are summarized in the mentioned survey papers mainly try to derive semantically rich information from the image and it's context directly and thus stem from semantically poor, low-level information to high level semantic information (bottom-up). This high-level semantic information, e.g. events, persons, or emotions, ideally reflect the way images are perceived by humans.

In recent years we can observe a new trend which approaches this key problem from a different perspective and analyzes how photos are perceived and dealt with to better understand their semantics for humans (top-down). These works also go beyond pure image analysis and also specifically consider aspects relevant for photographs. Some of these works even analyze the correlation between human perception and raw image image data. One of the most exciting new fields is the analysis about the social aspects of photos, that means how do people interact with and use photos. These studies are a starting point for many interesting attempts to consider the usage of photos as an additional, valuable source for semantic image annotation and thus enable for semantic image retrieval. One reason for this new interest in this social aspect of photos is perhaps the growing availability of relevant data due to the growing interest of people to share their photos online. To give just one example, in 2007 over 60 million photos were added to facebook each week[1]. This leads to a growing amount of photo interaction which takes place in the virtual world and therefore is comparably easy accessible and traceable computationally. Figure 3.1 illustrates these top-down vs. bottom-up approaches to narrow the semantic gap in image retrieval. A good discussion of these contrary views regarding image annotation is given by Hare et. al. [HSL$^+$06]. However, the authors have a slightly different definition of top-down and bottom-up. As bottom-up for semantic image understanding approaches they see approaches that automatically provide semantic labels for photos and parts of photos based on content analysis including their own concept of Semantic Spaces [HSLN08]. As top-down the authors see approaches which are model-driven, mostly based on ontologies and thus by basing on semantic models for images. We extend this model-driven perspective to a user- and usage-driven one.

Photos are a special means for many persons to capture a snapshot of important situ-

[1] `http://blog.facebook.com/blog.php?post=2406207130`

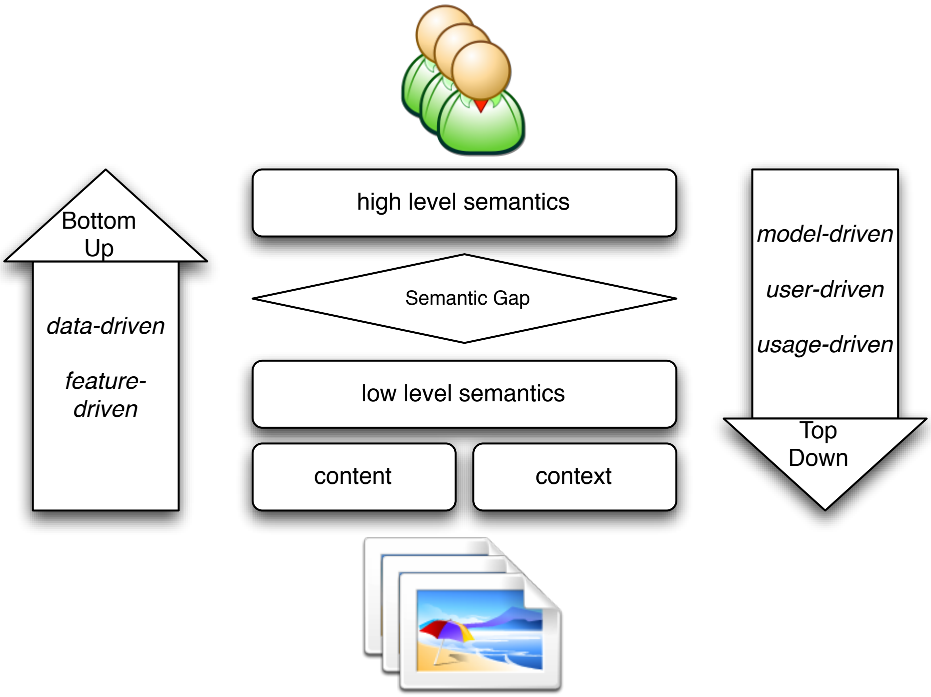


Figure 3.1: Top-Down vs. Bottom-Up approaches to bridge the semantic gap

ations, people and places and thus are both a way to preserve this snapshot for personal use and a vehicle convey to others. However, these subjective connections to real life situations are usually not present in the photos' content and for a long time now it has become clear the "One way to resolve the semantic gap comes from sources outside the image ..." [SWS$^+$00]. On the one side this can be context information which, e.g. camera capture parameters stored inside the image, but also a good understanding of why and how people take photographs and how they deal with them. This user- and usage-driven view on semantic photo understanding is the main interest of this article.

Despite the large number of surveys in image retrieval, to our knowledge there is so far no survey summarizing the works specifically dealing with semantic retrieval for (personal) digital photographs. With this article we are aiming at filling this gap and specifically target at works focussing on the above mentioned top-down aspects of photo analysis and retrieval. A special focus will be placed on the social aspects of photo retrieval. This interest has emerged especially in the last years beginning with first studies done by Frohlich et al. [FKP$^+$02].

We begin with a brief analysis of the specific aspects of photos compared to conventional images and summarize scientific works following our so-called top-down approach in Section 3.1. We then analyze works which concentrate on semantically analyzing and structuring photos regarding different aspects in Section 3.2. Section 3.3

focusses on semantic photo analysis and retrieval on a large scale and Section 3.4 summarizes applications of semantic photo retrieval. The chapter closes with a disucssion about works which relate to the layout of multimedia applications and digital photobooks in particualr ins Section 3.5.

3.1 The Nature of Digital Photos

In this section we take a closer look at the field of semantic image retrieval from a user's perspective which we referred to as top-down approaches. The most direct source for image semantics is the interaction of users with their photos. In the last years one can observe a growing interest in examining this, what can be called, social life of photos.

Looking at photographic images compared to images in general one can observe a number of special characteristics. First of all, a photo is usually a natural image which is a 2D representation of a natural scene. Natural images differ in their visual characteristics compared to artificial images regarding the number of colors, kind of edges, presence of faces, etc. which has to be taken into account e.g. for a good image retrieval system. Photos are also not just defined by their visual content, but are embedded in a rich context of other information, e.g. photographic metadata captured by the camera, photos which have been taken at the same event or the incentives of the photographer to actually press the releaser of the photo camera. Photos are usually taken with a specific purpose in mind, e.g. to capture the memory to a specific event, to share a moment with others or to prepare a gift for others. Studies have shown that the memory to events and thus also to associated photos are primarily remembered by (is decreasing order) the event itself, who was involved and where and when it happened [Wag86]. Conventional Image Retrieval Systems usually rely solely on image content analysis and cannot suffice these demands sufficiently.

In the following we distinguish between works incorporating studies analyzing the way people use and interact with photos in Section 3.1.1 and works which build on the outcomes of such studies to formally describe and automatically capture these usages in Section 3.1.2. We also survey different definitions of semantics in the context of digital photos in Section 3.1.3, review relevant metadata standards for photo annotation in Section 3.1.4 and provide an overview over the relation of photos to the semantic web and semantic web technologies in Section 3.1.5.

3.1.1 Studying Photo Usage

Recently, one can observe a growing interest of researchers in understanding the way people deal with their digital photographs. A special interest hereby lies on the activities regarding sharing of photos with others. An early work from an anthropologist's perspective analyzing the use of home videos and photos was done in [Cha87]. The author provides a series of field studies and interviews on home photo users exploring what people do with their photos as well as what their personal photos mean to them and aims to reveal the culturally structured behavior of underlying seemingly spontaneous photographic activities. A more practical study was done by Frohlich et al. [FKP$^+$02] with the goal to find requirements for photoware, which the authors define as technologies enabling photo sharing. Different kinds of photo sharing activities are categorized along the two dimensions time (same time, or different time) and space (same place, different place) leading to the categories co-present and remote sharing, archiving, and sending. These different categories were analyzed by interviewing and observing eleven PC-owning families regarding their use of printed and digital photos. With a similar goal in mind, but specifically targeting at mobile photoware, [AEN$^+$09] has studied how people use mobile devices to capture and share digital photos. For this, 26 participants were equipped with cameraphones and tools and accounts to upload and share photos. The main conclusion of the study was, that users generally appreciated a consistent integration of relevant tools enabling for photo managing and sharing, e.g. preferred an easy integration of photo sharing websites into their mobile phone. Another empirical study regarding the sharing of cameraphone photos was done in [VHDA$^+$05]. The authors equipped people with advanced technologies for mobile image usage and found out that users quickly adapted their photo taking and sharing activities to these new technologies. The study was compared to an an earlier study of the same authors regarding uses of digital photos in general [VHDT$^+$04].

One of the most popular means to share photos nowadays are online communities. We will do a thorough analysis of these communities in Section 3.3. However, some works have explicitly focussed on studying the way people share photos in such communities. A recent study analyzes different factors which impact the sharing of photos in online communities [NNY09]. The authors distinguish between motivational factors to contribute photos and structural factors which stem from the design of the photo community platforms. Additionally they analyze how the length of the membership in a platform does influence the sharing behavior.

As activities before sharing, but after capturing Kirk et al. have established the terminus photowork [KSRW06]. These activities are, e.g. reviewing, downloading, organizing, editing, sorting and filing of photos. The authors have categorized these activities according to different stages in the photo life cycle into the pre-download phase (before downloading photos from the camera to a computer), the download phase itself, and a pre-share stage. The authors have interviewed several subjects regarding their habits in activities in these different stages.

An older, but still relevant study aiming at finding out how people manage their digital photographs was carried out by Rodden [RW03] by observing and analyzing the management habits of 13 participants over a period of 6 months. The people were equipped with a special photo management tool which recorded their image usages. The result of the study was that the most important features of a photo management software are the possibility to chronologically sort a collection for easier browsing and to present overviews in thumbnail form. Crabtree [CRM04] has analyzed the way people naturally collaborate around photos and share collections of photographs. Bentley [BMH06] has studied the similarities between consumer photo usage and music usage.

The interest in understanding the human usage of photos has also led to dedicated workshops clustering different aspects of photo usage. In [FWK08] the boundaries between activities regarding preparation of photo for sharing (photowork) and actually using these photos to communicate with others (phototalk) are challenged by analyzing means of collaborative photowork. Another workshop[LDKT08] has focussed on the analysis of co-located photo sharing activities. These are sharing activities which take place at the same place and at the same time, e.g. watching a photo slide show together with friends.

3.1.2 Describing and Capturing Photo Usage

Besides works which try to better understand, how people use and interact with photos, there are some first attempts to capture and formally describe these usages. In [HNO$^+$05] a general model is proposed to capture the processes in media production. Based on this model several instances been developed for different domains, e.g. for the production of digital photobooks [STB08].

The representation of photo usage is so far not sufficiently addressed in current metadata standards. However, MPEG-7 [MSS02] provides a very high-level model to describe different usage events related to photos. In XMP [Ado05] a more detailed schema is provided which, however, mainly focusses on the usage in professional environments. The authors of [SBF08] has reviewed different forms of usage in the personal domain and has defined a model to capture such events which is capable to be implemented as an extension of XMP.

Despite these efforts in understanding and also modeling and capturing the usage of photos, so far little attention has been spent to actually employ such information to better semantically understand these photos. A method to derive information about a photo's importance from its usage in photobooks has been presented in [SB09].

3.1.3 Photo Semantics

The notion of semantics in the context of Image Retrieval was for a long time limited to the linkage of simple text annotations to images or image parts with no additional structure. In [HLES07] the authors argue that this simple image to word translation often does not suit the needs of a real world semantic retrieval system and propose a faceted model of image semantics rather than simple tags. These *Semantic Facets* of images are divided into object, spatial, temporal, activity/event, abstract and related concept,context and topic facets. These facets are either hierarchically, discrete or continuously organized.

Also other works have discovered that the semantics of multimedia are not static or unique but are composed of multiple semantics depending on the context [al.05]. Additionally, there not only multiple semantics, but semantics do change over time by interacting with them and are thus emergent [SBC06, SGJ01]. Also, as pointed out before, a photo seldom comes alone and has an intrinsic meaning. It rather changes or gets its meaning by placing it in context to other photos or by interacting with them [SGJ01].

The authors of [SJ07] have recognized this emerging character of multimedia semantics and have extended the view by defining five types of different semantics. These are Natural, Analytical, User, Expressive, and Emergent Semantics. Based on these different kinds of semantic a Semantic Ecosystem is proposed in the form of a framework for multimedia semantics.

3.1.4 Photo Metadata Standards

One way to capture semantics of photos in a way which is understandable by a wide range of tools and therefore makes these semantics accessible easily are dedicated standards for photo metadata. Due to the diversity of different tools and systems producing this metadata, a number of different, partly competing photo metadata standards have evolved over the years. In the following we briefly review different photo metadata standards, summarize their strengths and weaknesses and conclude, what is missing so far in our opinion.

One of the most used metadata standards for digital photos which used by literally every digital camera is EXIF[ISSE02]. The main purpose is to store camera related information together with the photo, typical ones being the time the photo was taken, aperture, white balance setting, shooting program, flash setting, etc. Most camera makers add proprietary metadata which is specific to their cameras and which is not defined in the Exif-Standard, e.g. information about the used lens-type. Some recent cameras also store information about the GPS-Position or even detected or recognized faces. Metadata in Exif is represented as simple key-value pairs. Even though the end user might use only a few of the key-value pairs they are relevant at least for photo editing and archiving tools which read this kind of metadata and visualize it.

PhotoRDF [LB02] is a project for describing and retrieving (digitized) photos with (RDF) metadata. It describes the RDF schemas, a data-entry program for quickly entering metadata for large numbers of photos, a way to serve the photos and the metadata over HTTP, and some suggestions for search methods to retrieve photos based on their descriptions. The standard is separated into three different schemas; Dublin Core, a Technical Schema, which comprises more or less entries about author, camera and short description, and a Content Schema, which provides a set of 10 keywords. With PhotoRDF, the type and number of attributes is limited, does not even comprise the full EXIF scheme and is also limited with regard to the content description of a photo. Thus, PhotoRDF is not widely used in practice so far by current applications.

The DIG Initiative Group of the International Imaging Industry Association is a group consisting of 80 leading companies in the imaging industry. The DIG35 [dig01] standard is meant as a more structured replacement for Exif and aims to define a standard set of metadata tags for digital images that can be widely implemented across multiple image file formats. A reference implementation is provided in XML. In addition to Exif DIG35 also provides means to capture basic information about the photo's history, e.g. if it was cropped oder combined with another photo. Despite these advantages DIG35 is not used on a large scale in practice.

MPEG-7 [MSS02] is a quite sophisticated ISO-Standard to describe Multimedia-Content. Besides algorithms for low-level media features also a schema to describe multimedia semantics is defined with the help of XML Schema. These schemas can be simplified with the help of profiles to accommodate for different applications where not every aspect of MPEG-7 is needed. Aspects relevant for photos in MPEG-7 are, e.g. the description and representation of content-based image features, limited capturing of the usage and interaction history of a photo, and semantic content description. MPEG-7 itself does not, besides the definition of the XML representation, define, how the metadata is stored beside or in the photo. The authors of [NH02] review the capabilities of MPEG-7 regarding its ability to hold semantic multimedia information. They also present a model and a syntax for multimedia semantics in the form of an extension of one part of MPEG-7.

MPEG 21 [BPWK06] is a standard primarily targeting at handling the production, transmission and distribution of multimedia content. It therefore does not focus on representing semantics of image data. The main aspects of the standard are the modeling and identification of digital items and Intellectual Property Management and Protection.

The Extensible Metadata Platform (XMP)[Ado05] and the IPTC-NAA-Standard (IPTC)[Com99] have been introduced to define how metadata (not only) of a photo can be stored with the media element itself. XMP borrows from several metadata standards, e.g. IPTC, Exif and Dublin-Core and defines different XMP-Schemas to store these kinds of data besides the photo or in the photo header with the help of RDF. XMP is extensible by defining additional XMP Schemas. As it is developed and maintained by Adobe, it is mainly used by Adobe products to store metadata information. XMP itself is distributed under the open source BSD license.

One problem with all metadata standards is, that the same information can be expressed in different standards and formats and thus can be stored multiple times and potentially ambiguously in the same photo. E.g., the information when a photo was shot can be both stored in the Exif and the XMP header of a photo. The Metadata Working Group (MWG) has tried to solve this problem by defining guidelines [Met09] of how to cope with these multiple definitions.

Metadata and the end user typically get in touch in the form of descriptive metadata that stem from the context of the photo. At the same time, in more than a decade many results in multimedia analysis have been achieved to extract many different valuable features from multimedia content. With MPEG-7 a very complex standard has been developed that allows to describe these features in a metadata standard and exchange content and metadata with other applications. However, both the complexity of MPEG-7 and the many optional attributes in the standard have lead to a situation in which MPEG-7 is used only in very specific applications and has not get a world wide accepted standard for adding metadata to a media item. Especially in the area of personal media, in the same fashion as in the tagging scenario, a small but comprehensive shareable and exchangeable description scheme for personal media is missing.

3.1.5 Photos and the Semantic Web

With the rise of the semantic web [BLHL$^+$01, SLH06] a few attempts have been made to also employ related concepts and technologies for the annotation and management of digital photos. An early attempt of bringing multimedia to the Semantic Web was presented by [Hun01] by building an ontology for parts of MPEG-7 represented in RDF Schema. In [LB02] an attempt was made to actually use Semantic Web Technologies on the Web for photo retrieval by the definition of a photo-metadata Ontology (PhotoRDF) and a system and prototypes to retrieve photos over http on the basis of this ontology. What however is often missing is a meaningful way for end users to retrieve photo content. With M-OntoMat-Annotizer [BPS$^+$05] the aceMedia project has developed a system to link MPEG-7 low-level features to higher-level semantics expressed in ontologies and other Semantic Web Technologies. Another tool for semantic photo annotation on the Semantic Web has been developed with PhotoStuff [HwSG$^+$05]. In contrast to M-OntoMat-Annotizer also the annotation of parts of an images is supported. The authors of [PPT07] have extended the NF^2 [CPSS04] Image Database model to an image ontology and have proposed a general architecture for supporting creation and management of multimedia objects. In [TOPS07] the authors describe advantages of using Semantic Web languages and technologies for the creation, storage, manipulation, interchange and processing of image metadata. Along with potential use cases relevant RDF and OWL vocabularies are described and an overview over existing tools is given.

As true for the Semantic Web in general, Semantic Web Technologies seem to be practical in the first place for clearly defined use cases and domains. One such use case is described in [ZKS08] where the authors describe a Semantic Web image repository for biological research images. In [HSWW03] the authors present a system for the management of art images on the basis of Semantic Web Technologies. For this, they employ multiple existing related ontologies like IconClass [WCTV85], Wordnet [Mil95], ULAN [Tru00] and the Art and Architecture Thesaurus [Pet94].

We can conclude that Semantic Web Technologies have been proven useful for the management of digital photos and are able to enable users for more semantic queries. However, in practice such systems are often not feasible: Semantic models are often either to broad to be meaningful in practice or are tailored to only a very specific domain (e.g. art images). However, the Semantic Web and semantic web technologies in the context of digital photos could gain more attention with the current trend to share photos and social context information online. Also the growing connectivity of mobile devices could help to push the Semantic Web. E.g. Semantic Web Technologies have successfully been employed for the semantic annotation of photos on mobile devices [MO06, VFG$^+$07]. Here the semantics act as an additional source to retrieve information about a specific semantic concept or event.

3.2 Semantic Photo Analysis

Comprehensive Surveys have been written on the general field of image retrieval, mostly focussing on works considering the content of the images [DLW07, SWS$^+$00, LSDJ06, AZP96]. In this section, however, we will focus on reviewing works specifically targeting at semantic photo analysis. By semantic analysis we mean methods which aim at semantically annotating photos in some way, usually by labeling them according to specific semantic classes. One application for this is to make photos accessible for text-based queries in retrieval systems. A special kind of semantic is the perceived quality of a photo. This is relevant for many applications, e.g. when selecting the *best* photos of an event or when summarizing large photo collections. Thus, we review relevant works for image assessment in a dedicated section (Section 3.2.3). One special characteristic of photos compared to images in general is their linkage to a rich set of contextual information. Researchers have realized the potential of this contextual information for semantic analysis and thus we review respective works in Section 3.2.2 after starting with works solely considering the photo's content in Section 3.2.1.

3.2.1 Content-based Analysis

In the large and established research field of content-based image retrieval (CBIR), we find approaches that specifically address the domain of personal photo collections and digital photo albums. In these approaches, content-based analysis, partly in combination with user relevance feedback, is used to annotate and organize personal photo albums. As mentioned before, a couple of excellent surveys already exist reviewing works in CBIR. Thus, we only give a very brief overview here and highlight some works specifically targeting at semantically analyzing personal photos.

An important aspect when speaking of photo semantics are the presence and identity of persons shown in photos. An excellent survey over face recognition is given in [ZCPR03] and another article summarizing works regarding face detection was given in [YKA02]. Both survey papers are quite old but still relevant. A popular face detection algorithm, which is used by many applications, is the one proposed by Viola et. al. [VJ01]. A similar popular technique for face recognition are the ones basing on Eigenfaces [TP91].

For a more detailed review of works reading automatic semantic labeling of pictures we refer to the before mentioned surveys, particularly the one from Datta [DWGW07]. Probably the first work trying to link textual description with image data was done in [MTO99] with a co-occurrence model to keywords and low-level features of rectangular image regions. Later approaches for automatic annotation can generally be divided into ones that try to first segment pictures and annotate the different parts separately and the ones that consider pictures as a whole and thus take a more scene-oriented approach. An example for a segmentation-based approach is the work of Duygulu et. al. [DBDFF06]. The authors propose a machine translation model which translates keywords to a discrete set of clustered image regions or blobs. This method and the chosen vocabularies was employed and extended later by Cross-Media-Relevance-Model (CMRM) [JLM03] and Continious-space Relevance Model (CRM) [LMJ04]. Cusano [CCS03] has employed multiclass Support Vector Machines to categorize image regions into basic classes, e.g. sky and ground.

Regarding scene-oriented approaches one example is the system built by Oliva and Torralba [OT02] which apply basic scene basic scene annotations, such as 'buildings' and 'landscape' using relevant low-level global features. In [VFJZ99] a method is proposed to hierarchically cluster vacation images according to different semantic classes based on low-level image features. These classes (e.g. indoor / outdoor, landscape, ...) are modeled with a bayesian approach. Yavlinskiy et. al. [YSR05] have used simple global features together with robust non-parametric density estimation using kernel smoothing to perform automatic scene annotation on the Corel Data Set. Another system which was optimized for high performance is the real-time image annotation system built by by Li et. al. [LW08] which is also used in the Alipr[2] web image retrieval system.

3.2.2 Context-based Analysis

In the beginning of the last century it became clear that content based image analysis is limited when aiming at semantic photo analysis and that "One way to resolve the semantic gap comes from sources outside the image ..." [SWS$^+$00]. Thus researchers started to explore, which information in the context of the image could enhance the the quality of semantic image annotation. With the availability of time and location from digital cameras, we find works that aim to use this contextual information, sometimes in combination with content-based features, for organizing and accessing digital photo collections. Stating that "Pictures are not taken in a Vacuum" but are rather embedded in a context which gives hints to high-level image annotation [LBB06, BL04a, BL04b, BL05, BBL06] employ camera metadata like time, aperture and focal length to assist content-based semantic labeling and also derive semantic labels like indoor/outdoor from the context alone.

Many consumer cameras provide templates for camera settings for different types of sceneries, like night shots or portraits. [KKL07] employs the camera settings templates written into the photos' Exif headers to semantically classify photos. The context parameters for time and space are not only used to organize the photos but also to form clusters that represent higher-level semantic concepts such as the collaborative detection of events in photo collections in [NRD05].

With PhotoCopain [TGO$^+$06] the authors have established a semi-automatic and multimodal photo annotation system which fuses information from the photos' context with information from the web. Sinha [SJ08a] discusses various kinds of optical context data

[2] `http://www.alipr.com`

for their ability to predict various semantic photo classes. Based on a home and a public photo data set from the web a classifier is trained based on this context data. Another work of the same author combines this optical context data with high-level content-based features [SJ08b].

3.2.3 Photo Assessment

Of particular interest in many applications is the knowledge about the quality of an image, e.g. for being able to decide which are *best* 70 out of a series of 500 or to automatically delete *bad* photos from an event. But even when letting humans decide about the quality of a photo, a wide variety of different ratings can be examined [DLW07]. This shows, that the special semantic *quality* oder *aesthetic value* is highly subjective as each image is perceived differently by every person. Despite of this highly complex topic of image assessment, some works have done the attempt to provide means to, at least in part, objectively measure the quality of an image. In the following we give an overview over this field. We start by works examining the content of photos and take a closer look at works rating the aesthetic of photographic images.

Content-based image assessment has been a relevant topic in the research community for a long time, primarily to provide means to evaluate the quality of compression algorithms, e.g. determine the level of distortion or the similarity to the original image. A good, but quite outdated survey for image quality measurement is given in [Esk00]. Algorithms depending on the presence of an original (uncompressed) image are commonly referred to as bivariate assessment approaches. An early example for bivariate assessment is presented in [EF95]. Based on the assumption, that the human eye primarily detects structural changes between photos [WZ04] provides a framework which focusses on detecting structural differences between a photo a its compressed counterpart. An early evaluation of content-based image features for univariate image assessment targeting at the application to image compression formats was presented in [ASS02]. A similar approach was proposed by Li et al. [L^{+}02]. The authors propose three aspects by which humans rate the quality of an image – edge sharpness level, random noise and structural noise level – and provide methods to extract these measures from images and estimate the respective levels. Another approach to derive a measure for image quality, mainly for the evaluation of parameters for the JPEG2000 algorithm, is presented in [SBC05]. In [DVKG^{+}04] a model-based approach for image assessment is chosen, modeling image distortion as a combination of frequency distortion and additive noise. The authors derive a distortion measure (DM) and noise quality measure (NQM).

In recent years we see more and more works which take our before mentioned top-down approach for building methods for image assessment and thus aim to stem from a human perspective. One of the first works asking the question “why image quality assessment is so difficult” was presented by Wang [WBL02]. The author presents a system which was inspired by the fact that “the main function of the human eye is to extract structural information from the viewing field ... Therefore, a measurement of structural distortion should be a good approximation of perceived image distortion". Based on this philosophy, a simple image quality metric is proposed.

Similar to the semantic gap in image retrieval, Datta [DLW08] introduces the notion of an *aesthetics gap* to describe "lack of coincidence between the information that one can extract from low-level visual data (i.e., pixels in digital images) and the interpretation of emotions that the visual data may arouse in a particular user in a given situation." This aesthetic gap has been tried to overcome only by a few works so far. An interesting approach in presented by Datta in [DJLW06, DLW07]. These works try to design low-level features which are able to describe the presence of simple photographic rules like sharpness or the rule of thirds in photos. Then the correlation between these low-level features and ratings from a photo community website for a set of photos is determined and an aesthetics classifier is built. While being an interesting initial approach to learn from the user how to derive photo semantics like aesthetic, the experiments only showed limited success. One problem might have been the expected output for the system to provide a rating on a continuous scale. Other works approaching a similar problem rather try to rate an image on a more coarse scale: Ke [KTJ06] provides a method for a two-level distinction between professional and snapshot photos with high-level features. While not directly being an indicator for the quality of an image, the authors' method leads to quite good results on general web images.

[TLZ$^+$04] follows a similar goal and approach but bases on stock-photos from the Corel Database for professional photos and from a private photo set for snapshot photos, which might arguably lead to less realistic results than the work of [KTJ06]. Haider [MHMK09] proposes a Hybrid Image Quality (HIQ) measure by combining the most promising quality measures from [ASS02] and evaluating their performance based on extensive user studies.

Recently, a few works have extended photo assessment methods with visual attention analysis: Sun [SYJL09] first constructs a face sensitive saliency map of a photo and combines it with the presence of photographic rules in the respective salient regions to generate a photo quality score. You [YPHG09] also employs visual attention for image assessment but additionally focusses on the temporal aspect for video assessment. In [BE04] the authors provide an interesting solution to measure the quality of a photo before it is shot. They propose a system to measure the current scene on the-fly to the conformance of basic photographic rules like the rule of thirds and provide feedback to the photographer. The system ismeant to be implemented on a digital camera. In [OAOO09] the authors combine tags and content-based image aesthetic assessment to aid image search in social networks.

Besides these methods for assessing photos, some researchers try to measure the quality of specific parts of a photo. A common domain are faces, or facial attractiveness. E.g., Eisenthal [EDR06] aims at finding low-level features which correlate to the human perception of facial attractiveness based on Eigenfaces.

In conclusion we can observe a strong trend from simple image quality measures, mainly used for evaluating image compression algorithms to methods which stem from a human perspective. Results from psychology about the human perception of images are more and more influencing research on image assessment.

3.3 Large Photo Collections

In this section we give an overview over works focussing on the semantic exploration of large photo collections. We start with arguably the largest source for photos, the web in Section 3.3.1 before taking a look at social photo repositories in Section 3.3.2 and personal photo collections in Section 3.3.3. We conclude with a review of algorithms dedicated for analyzing large image collections in Section 3.3.4 and the evaluation of photo retrieval methods in Section 3.3.5.

3.3.1 Web Image Retrieval

Web image retrieval and classification aims to open up images on the Web for multimedia retrieval, often by textual queries that rather work on the surrounding content of the Web page than image content. Since the HTML code of a page provides rich context information for embedded images, most related approaches use text-based image retrieval as a basis by examining textual content such as surrounding text, image metadata, or take hints from the HTML code and especially its structure. Only secondarily, if at all, the actual Web image content is used.

Similarly, current keyword-based Web image search engines such as Google Image Search (images.google.com) or Yahoo! Search (images.search.yahoo.com) employ mostly surrounding text features and the image name and path which allows for keyword-based queries. As an addition several approaches exist to combine traditional text-based queries with content analysis of images.

A survey and comprehensive discussion of existing technologies in Web image retrieval and how they address key issues in the field can be found in [KZB04] along with application scenarios and a comparison of different existing systems. An early example for a system also considering the image content is ImageRover [SLCS99], combining text-based queries to provide initial example images following a content-based query-by-example approach to refine the result set. Other approaches combining text-based queries and content-based image similarity are [Lew00, OBMCP00, LW99]. Besides these works for general web image search more specialized approaches exist concentrating on specific domains (e.g. celebrity search by face recognition [AY00]).

Purely text-based Web image retrieval approaches are solely taking the HTML code into account to derive annotations for the images contained in the page. As one example, [TM01] uses an approach for image retrieval examining the attributes of the image tag itself, surrounding paragraph and title of the page. However, further examination of the the effectiveness of features [MT01] leads to the conclusion that the HTML source contains the most valuable initial clues, but that results improve if images themselves also are examined. Aiming to understand the role of an image on a Web page, [MHN06] uses a range of features to assign roles to Web images. The features are taken from structural document data and image metadata but not from actual image content. Based on these features, various roles of Web images are derived and used to preprocess images for display on a mobile device. By means of structural analysis of HTML documents, [GUC05] segments Web pages into semantic blocks. The authors propose that embedded images inherit the semantics from their surrounding semantic block. Other approaches consider both content and HTML context of an image for semantic annotation and clustering. An early work [FSA96] uses the HTML content and a few selected image content features for Web image search. The iFind system [HCW$^+$07] considers context and content of images and examines their layout on a page for semantic annotation. The spatial layout of images on Web pages is used to identify nearness for groups of images. Closely arranged images are then considered to also be semantically related.

3.3.2 Social Photo Repositories

Social Networks have gained much popularity in the last years and for many people they have become an important part of their social life [BE07]. Although often much less suited than dedicated photo sharing communities, social networks have motivated many people to share their photos with others. Currently, over 850 million photos are uploaded to face book per month[3]. This massive increase has motivated many scientific works in the recent years which aim to employ the collective knowledge hidden inside the network for semantic photo analysis.

Recently Nov [NNY09] has studied the incentives which drive users to share their photo online and also analyzes which other tenure and structural factors of online communities affect photo sharing. [ME07] even argues, that social communities effect the way we shoot photos. In [NGP08] the author specifically analyzes the concept of photo groups in photo community platforms and extracts relevant patterns of photo-to-group sharing practices. One result of the study is that many users are engaged in photo groups and that most of these users show high loyalty for these groups with a very strong engagement in annotating and sharing. Kennedy [KNA$^+$07] has analyzed the potential of content and context in Flickr photos to derive knowledge about the world. Specifically the authors derive representative tags for different locations in the world and aim at deriving a photo's location from its tags.

Additional knowledge in social media platforms like annotations, ratings and social relationships have also inspired researchers to seek for new ways to employ this knowledge for photo retrieval. A recent work [CI10] considers social peer relationships extracted from a social network as an additional source for photo collection clustering besides Exif-context and image content. Becker [BNG10] also tries to employ knowledge in social media platforms for event clustering.

[3] http://www.facebook.com/press/info.php?statistics

3.3.3 Personal Photo Collections

While being usually substantially smaller than the before mentioned images sources, personal photo collections, which usually reside on the single person's hard drives of their computers, can even grow to multitudes of hundreds or even thousands of single images. Thus exploring these large collections can reveal much about the single photos, but also about the single person's life and her or his social network.

In [ALCV08] a clustering algorithm is presented which clusters personal photo sets on the basis of face detection and global low-level features on the photos' backgrounds. In [ZCWT10, ZTL$^+$06] the authors try to derive a person's social network by clustering the faces shown in the photos based on appearance in the same photo, spatial relationships and similarity of the surrounding context. In [PW09] similar features are employed to derive social clusters from personal photo collections.

[MO07] employs social networks as a source for additional information to semi-automatically annotate personal photos by retrieving information about the owner's social network to provide initial photo annotations. Using the fact that photos in personal photo collections are generally highly semantically connected depending on their similarity regarding event or time Cao [CLH08] has proposed an approach to automatically annotate personal photo collections by label propagation according to multiple similarity cues. An interesting method to automatically annotate personal photo collection by web mining is presented in [JYH08].

3.3.4 Algorithms and Frameworks for Large Scale Analysis

As the amount of data increases it becomes especially important to optimize the analysis algorithms to speed up the process. Considering image retrieval, high computational cost is usually needed for feature extraction and machine learning algorithms, especially kernel-based methods like SVMs. This problem has lead to some works with the goal to provide methods to reduce the computational workload or to provide means to parallelize algorithms to be able to use large clusters of computers.

A relatively new computing paradigm called Data-Intensive Scalable Computing (DISC) [Bry07] has emerged which was specifically designed for the analysis of large data and mainly focusses on the data rather than computation and specifically considers constantly growing and changing data collections, like large image collections. A popular framework, built on top of this paradigm, developed by Google is MapReduce [DG08]. Users of the framework have to split up their specific problem into a map and a reduce function and the system automatically parallelizes the computation over several clusters of machines or processors. MapReduce is actively used at Google to process large amounts of data and has also applied by researchers to develop highly scalable machine learning algorithms [CKL$^+$07].

A system targeting at efficient semantic concept learning for large multimedia collections is presented in [YFM$^+$09]. The authors propose robust subspace bagging (RS-

BAG) for this purpose which aims at a fast learning process while minimizing overfitting which is a common problem when dealing with large collections of training samples. The authors claim to have achieved a ten-fold speedup compared to conventional SVM-based learning while maintaining the same recognition performance.

3.3.5 Evaluation and Test Sets

In the early days of image retrieval results of research activities were often highly subjective due to the difference in data sets. Usually researchers established their own image collections to evaluate their algorithms which made it difficult to compare different approaches for other researchers. Over the time a few standardized image collections have been build exactly to tackle this problem. One of the first and most used was the Corel data set, a collection of photos selected from proprietary stock-photo CDs. The Corel CDs consisted of more than 800 CDs, each containing pictures of roughly the same topic. Besides the fact, that these CDs are no longer commercially available, comparing research results on the basis of this data set is questionable. As pointed out in [MMMP02] the usage of the set is not standardized and so different evaluations usually use different subsets of the data set which lead to not comparable results. The main problem however is, that the Corel data set is not public and therefore not easily available to interested researchers.

A quite outdated overview over efforts in the evaluation of CBIR systems is given in [MMS$^+$01]. A brief overview over the field of benchmarking multimedia information systems with a strong focus on image and video retrieval up to the year 2006 was given in [MMW06]. An early example of a publicly available image data collection has been established by the benchathlon network [MMB05]. The goals of the benchathlon network are to provide tools and data sets to make CBIR systems comparable [MMMM$^+$03]. A more recent activity regarding evaluation is ImageCLEF which is a part of the Cross-Language Information Retrieval Campaign (CLEF), which attracts attention from both the industry and academics. ImageCLEF is organized in several tasks on an annual basis which tackle different aspects of image and analysis and retrieval. At least one task every year specifically deals with photographs. ImageCLEF defines tasks to be solved and additionally provides data sets on which these tasks should be carried out. In [ATSC08] the authors provide an evaluation criterion for measuring the diversity in results sets for image retrieval and adapt a test collection from ImageCLEF for this.

Another, not yet as matured as ImageCLEF, attempt to cluster research activities for commercially relevant problems and make the results comparable is the Multimedia Grand Challenge [ACM] which is held since 2008 in conjunction with the ACM Multimedia Conference. Researchers are encouraged to approach important problems for different companies in the multimedia domain and the companies are encouraged to provide meaningful data sets for evaluation purposes.

Looking at these different attempts to provide means to compare different algorithms in image retrieval, we can observe a couple of problems. First of all, the most useful data sets are usually designed for one specific research problems, as the data sets provided for the different ImageCLEF tasks. This makes these test sets perfect for evaluating this one specific task but makes them usually much less suitable for different problems. One challenge therefore is to find the right trade-off between applicability to as much research problems as possible and optimal tailoring for a specific problem. Another challenge is, that systems incorporating the user or relying on evaluations involving the user are highly dependable on how the user perceives a specific data set. E.g., different users might link different emotions or semantics to the same photo. For many tasks, not only the photo but also the story behind influences evaluation results and thus many tasks are only meaningful when performed on the users' own photos. For these cases it is not possible to define a unified test set. However, despite these potential problems the goal should still be to define meaningful test data sets which are freely available and should suit the needs for most of the research problems in the domain of personal media retrieval. A first step in this direction has been done by Shirahatti [SB05] who maps of various retrieval algorithm scores to human assessment of similarity and thus provides a way to automatically evaluate CBIR systems driven from a human perspective.

3.4 Applications

Having analyzed works dealing with the semantic annotation and retrieval of photos, in this section we focus on applications employing such semantics. We start by reviewing works dealing with the management of personal photos and focus on on mobile photo applications in a dedicated section. We conclude with a review of commercial and public end user systems.

3.4.1 Photo Management

One of the main applications of semantic photo annotation and retrieval is to aid the management of photo libraries. This becomes even more important when thinking of the massive increase of photos in personal libraries in the last years. Thus, there is a growing need to be able to easily find photos in a large collection based on information like where or when the photo was taken, at which event or which persons or objects are shown in the photo.

Rodden et al. [RW03] have found out that the main means to explore a collection of photos or to find a specific photo is based on the time information and the underlying event. They also found out that manual annotation of photo collections, which might significantly ease photo management, is usually not done and can also not be expected from the user. Following this fact Graham [GGMPW02] sees "Time as Essence" for photo browsing and proposes a system to summarize photos based on their time information. The author built a photo browser that follows this event based schema and compares it with a conventional photo browsing tool. In [GAC^{+}03] a photo management tool for large collections is presented which detects events based on their time information and presents these events along with the associated photos in a calendar like view. Other early examples of photo management systems which purely rely on content-based indexing are MiAlbum [WSZ00] and AutoAlbum [Pla00]. Girgensohn has presented another content-based photo management system which enables users to browse photos based on the people shown in the photographs [GAW04]. Specifically targeting at the application aiding users in quality screening their photos in [Lou00, LS03] an approach is presented which provides a multi-stage time-clustering algorithm to determine events and sub-events in photo collections and determines candidates for low-quality photos based on content-similarity and edge-detection. With FreeEye [RC09] an interface for large scale photo browsing was recently presented which clusters photos purely on their content-based similarity and arranges visually similar photos around a query photo.

Besides these works which either employ the photo's content or context for photo management, a couple of hybrid approaches exist. Stating that "Time matters" [ML03] introduced another system to cluster photos based on their time information to enhance photo browsing, but additionally provides a hierarchical event model consider content- and context-based features for retrieval. With SmartAlbum [TCMK02] a multi-modal system was proposed which indexes photos with a combination of speech and content-annotation. PhotoTOC [PCF02] is a successor of the AutoAlbum system [Pla00] where large collections of photos are automatically clustered based on their time information and content-based similarity. An unsupervised system with the same goal and features was later presented by FXPAL [CFGW05]. Mei [MWH$^+$06] employs the Expectation Maximization algorithm and sees events as a latent semantic topic to cluster photos based on time, content and camera settings. Gargi [GDT02] has proposed a distributed photo management system which provides search facilities for multiple users based on low-level color, texture and edge features combined with Exif-features. A more recent paper [CI10] combines content- and time-based similarity with social peer relationships for photos hosted in social networks. The goal was to provide a photo management tool driven from a more social perspective.

Besides time, location has gained more and more importance in the management of digital photographs. This is probably due to the increasing availability of location information either by directly storing GPS-Information in the photos' Exif-headers by mobile devices or GPS-Cameras or simplified tagging possibilities in social photo sharing sites on the web. In PhotoCompas [NSPGM04] time and location are used to determine event and location clusters to enhance photo browsing. In context of the MediAssist project in [OGL$^+$05] a photo management tool is presented which enables the user to perform searches by location and time to access personal photo collections. Chen [COT06] sees events or episodes in ones life as the primary means to browse photos and detects these events on the basis of time and location information. PhotoGeo [LFSBS08] follows a similar goal but aim s at providing a hierarchical view on detected events.

In Multi-User Environments a different browsing and surfing behavior of users can be observed: Users rather prefer to browse photos according to social interactions, e.g. specifically look for photos from friends or look for photos of personal interest and thus the temporal aspect is less important than in their own photo collections [KN08]. Jaffe

[JNTD06] aims at generating summaries for large photo collections of geo-referenced photos by an adaption of the Hungarian Clustering algorithm. The images are classified hierarchically and then ranked based on an empiric heuristic, emphasizing their originality. With Photo LOI [NRD05] another photo browsing tool specifically targeting at multi-user photo collections has been developed which combines photo sets sharing the same temporal, spatial and/or social context.

3.4.2 Mobile Photo Applications

Besides general photo management applications, some works specifically target at mobile photo management. In the context of the MediAssist project a mobile interface [GJL$^+$05] to personal photo archives has been developed. The user can browse a photo archive by means of time, place and other contextual data. Photo-to-Search [FXL$^+$05] is an interface for mobile web image search for mobile devices. For this, the user can search for images by providing a query picture with the integrated camera and optionally add a textual description. Combining text-based retrieval and CBIR matching relevant images are retrieved from the web. In the context of the ATLAS Project a system has been developed [PG05] that organizes photos on a mobile phone with gaussian mixture models based on time and place. An extended version and an integration into a mobile photo management tool is presented in [Pig10]. With MAMI [AXO08] a multimodal photo annotation and retrieval tool on a mobile phone has been proposed which lets users annotate, index and search photos based on speech and image input. For this photos and speech are analyzed directly on the phone without the need for an external service. With iScope [ZLL$^+$09] a system for personal image management and sharing on mobile devices has been developed. Photos are clustered according to content and context and online learning techniques are employed for the prediction of photos the user might be interested in. Monagham [MO06] presented an approach to automatically detect the identity of a person in a photo employing information from social networks similar to the PhotoCompas [NHWP04] system. Additionally the presence of bluetooth devices at the time of photo shooting are recorded and used to determine the presence of persons at the same time and place. A similar idea idea has driven the development of PhotoMap [VFG$^+$07] which also employs information about nearby bluetooth devices to connect them to FOAF (Friend-of-A-Friend) profiles. Such social information is integrated in a ContextPhoto Ontology incorporating spatial, temporal and computational context to support photo annotation on mobile devices employing the Semantic Web as a knowledge base

3.4.3 End User Applications

Besides the mentioned scientific applications for semantic photo retrieval, also a couple of commercial and open source, photo related systems exist which are meant to be used by end users and software developers to aid the analysis, retrieval and management of photos. In this section we give a short overview over the most interesting ones in our opinion.

The first commercially available image and video retrieval system which employs query-by-example was probably QBIC [FSN$^+$95]. Besides this, several others have been established, often as an addition to existing IR systems, such as Oracles *inter*Media extension to their Data Base Management System. A lot of research CBIR systems have been made available to the public, often under an open-source license. Caliph & Emir [LC08, Lux09] are java-based libraries for content-based indexing and retrieval employing MPEG-7 content-features. Other content-based systems are BRISC[4], Anaktisi [ZCPB09], GIFT [5], and imgSeek [6].

Besides these systems for managing and indexing personal photo collections, several commercial web image search engines are available on the web. All major web search engine providers (Google, Bing, Yahoo, ...) do provide means to search for visually similar images giving a query image.

Semantic photo metadata besides time has also arrived at end user photo management systems. E.g. Apple iPhoto[7] and Google Picasa[8] offer automatic event detection based on time, automatic person recognition, and allow for browsing photo collections based on on time, place, event and person. We also observe a common trend to bring personal photo management to the web and to provide means for sharing photos via social networks and photo communities.

Sophisticated semantic image analysis techniques are also used in other commercial end user systems besides photo management. E.g. with SmileBooks[9] users can semiautomatically design their own personal photobooks and let them be printed. The authoring software offers to automatically select photos based on their importance derived from the images context and content.

3.5 Photobook Layout

Most of the approaches for automatic layout and design of multimedia presentations aim at optimizing the page layout, e.g., based on minimal white space or maximization of the number of photos or regions of interest. The creation of photo collages presented by Girgensohn et al. [GC04] analyses photos for the region of interest and constructs stained-glass like photo collages from photos with faces. Other approaches aim to infer a tree-like structure on the photos and base their layout on these structure. For example, the approaches presented in [Atk04, Atk08, KS05] map the result of hierarchical clustering of photos directly to the spatial layout of the page. Others employ optimization techniques to minimize or maximize parameters such as the whitespace on the page or

[4] `http://brisc.sourceforge.net/`

[5] `http://www.gnu.org/software/gift/`

[6] `http://www.imgseek.net/`

[7] `http://www.apple.com/de/ilife/iphoto/`

[8] `http://picasa.google.com`

[9] `http://www.smilebooks.com`

the occlusion of salient regions [WQS$^+$06].

We also find approaches that derive clusters and importance of photos by content and context analysis and employ this information to select appropriate, pre-defined layout templates and placing the photos based on these [CCK$^+$06]. A recent approach by Xiao et al. [XZC$^+$08] presents a web-based software prototype for designing photo collages and also incorporates an algorithm for automated page layout. AutoCollage [RBHB06] assembles a set of images on a canvas by alpha masks to hide joins between the different images and employs energy maps and region of interest detection for distributing the images over the page. In a template-based approach by Diakopoulos et al. [DE05], pre-designed templates are used which consist of cells for photos and annotations applied to these cells. The layout is filled by matching the metadata of photos to the annotations in the cells using an optimization algorithm. An automatic color design for a hypermedia presentation for a generated spatial layout of a page is presented in [NMH03]. The color scheme selection, however, is not driven by and targeted at digital photos and photo compositions.

One of the few approaches that integrate design principles into the automatic presentation generation is presented by Lok et al. [LFN04]. In their system, the authors introduce the concept of a WeightMap to solve the problem of visual balance for presentations. A work with a motivation similar to ours for automatic photo collage layout is the one proposed by Geigel et al. [GL03]. The authors try to mimic the artistic nature of the album layout process by employing genetic algorithms. A more recent work [Gao09] presents an interesting authoring system for the selection, of photos for photobooks and theme-based grouping, automatic background selection and automatic cropping. Following a similar goal, this approach however is based on a limited set of basic templates and does not consider images as backgrounds which, however, are very common in professionally designed photobooks. Another recent work of the same group [BHW09] presents an aesthetically-driven layout engine for the automatic generation of page-based presentations with pre-defined static content where pre-assigned areas can dynamically be filled with content whilst following aesthetic principles. In [DJLW06] a very interesting and extensive study on analyzing aesthetics of photos and deriving an aesthetic rating for these is presented. While not targeting at photo presentations, the concepts and ideas in this work have been one of the inspirations for our work.

In conclusion, the approaches we find in the field provide methods which are algorithmically elegant and have been providing good contributions with regard to automating photo compositions. However, they do not sufficiently address the aesthetics of the generated results. To bring the end user in the position of becoming a great designer of his or her photo compositions, we advance the work in the field in a number of ways: (1) We do not only consider photos as input to the system but also headings and text blocks, and we reflect their specific characteristics in the layout, taking into account that in the digital age photos often come with additional labels and descriptions. (2) We implement a page design that is strongly based on aesthetic principles stemming from common design rules and on the analysis of existing, professional designs. (3) We understand the background and foreground as a coherent presentation and automate the layout of both.

(4) We aim at creating compositions that are unique and reflect the features of just the individual media set.

4 Photo Usage for Semantic Image Understanding

Features extracted from the content and context of digital photobooks usually form the basis to semantically annotate images, e.g. if they are more important or more appealing than others. However, it is difficult to decide how exactly these semantics can be derived from the features or which features are discriminant for this. This is problem is also well-known as the semantic gap in multimedia information retrieval.

Looking at the history of multimedia information retrieval we generally see two different approaches for semantic media analysis [HSL$^+$06]: One taking an bottom-up approach by inferring semantics from the signal level such as providing detectors for different scenes or concepts. Top-down approaches leverage additional context information such as ontologies or context associated to the media to improve the semantic understanding. These approaches for enriching media and therefore building the basis for retrieval are all based on the underlying data.

On the other hand there is an emerging trend for human-centered multimedia retrieval and [LSDJ06] states that “the foundational areas of MIR were often in computing-centric fields. However, since the primary goal is to provide effective browsing and search tools for the user, it is clear that the design of the systems should be human-centric.”. This implies that the means for interacting with a retrieval system are driven by the user taking the personal needs and habits into account. [ES03] for example identifies different user groups and types of images based on the usage of the photos. Affective computing tries to consider the user’s emotional state and takes this into account, e.g., [SLC$^+$02] detects emotions by analyzing the users’ faces and considers these emotions as an additional input for a multimedia retrieval system. This approach of taking the user in the loop when designing a retrieval system seems natural as in the end the user decides if a result is satisfactory for his needs or not. However, it seems that in a way these human-centered approaches stop halfway: They still rely on semantics and metadata which are often extracted without having the application of the retrieval system in mind. It seems that many methods for extracting semantics for media are not developed to be useful for the user but because *it can be done*. One rare example where a method to derive an aesthetics rating for images is developed in [DJLW06]. Here a top-down approach has been chosen to select the content-based image features relevant for the decision process, based on photographic and aesthetic rules, such as the rule of thirds or the golden ratio. Semantic analysis of digital media is usually driven by the methods available for content- and context-based media analysis and thus the means of accessing and retrieving media are driven and often limited by the kind of metadata and semantics extracted. However, it is the human giving things semantics, stating that a particular photo is more interesting than others or captures that specific moment more accurately than others. Thus, it seems obvious to consider the human and human’s interaction with media as the preferred source for semantic media annotation. Some research has been done on understanding the ways people interact with their photos, how they organize and access them, but the findings of these investigations have had little impact on media analysis and retrieval so far.

It is difficult to derive semantics from photos such as interestingness or applicability for a specific task from only analyzing the photo itself. However, if one would know, that a photo has been viewed very often and long in slideshows compared to other photos or has been ordered as a print, one might derive that this photo is more important to the owner than other photos. Similarly, if a photo was published on a web site with public access one might derive that this is a photo which is appropriate to show it to other, unknown people. These events in a photo's life cycle can be seen as the *history* of the photo. In contrast to information derived from the photo's context or content, usage is a highly dynamic and subjective source of information: Photos have different semantics for different users and usually photos can not be seen individual but are embedded in an environment which is living and evolving continuously. The usage history of a single photo can therefore not only change the semantic of a the photo itself, but also the semantics of related photos. Despite this high potential, the history of a photo has so far, contrast to other media such as audio, surprisingly seldom been used for annotation and retrieval of personal photos.

The goal of this chapter is to partly close this gap and to investigate, how observing the usage of photos can potentially help in semantically analyzing them. One way to accomplish this is to analyze the usage of photos in scenarios where it is possible to more or less directly *observe* certain semantics and try to determine the features which show a correlation with this semantic. We want to analyze if and how knowledge about the image usage can contribute to a better image understanding. For this it is important to know what photo usage is exactly. We understand photo usage as the life cycle photo which consists of a series of different *usage events*. We further define our understanding of these usage events in the following by building up a photo usage model. We also analyze which usage events and which parameters can potentially be used to derive image semantics. To prove our assumption we conduct two experiments capturing photo usage in a slideshow situation and show how captures measurements correlate with the photo importance.

4.1 Forms of photo usage

Only some works have analyzed and categorized the uses of photos. One of these works is [FKP$^+$02] where the author categorizes and analyzes one of the main forms of photo usage, sharing, which involves at least two persons. Frohlich categorizes these usages along a temporal and a spatial dimension. While sharing being the main form of photo usage, it is by far not the only one, e.g. many photos are not shared but only shot for personal purposes. We therefore take a different approach define photo usage as a life cycle consisting of different types of usage events. Our proposed model is shown in Figure 4.1. We distinguish between four types of photo usage: *View*, *Compose*, *Share* and *Archive*. Usages in these categories can either be analog or digital, e.g. a photo can

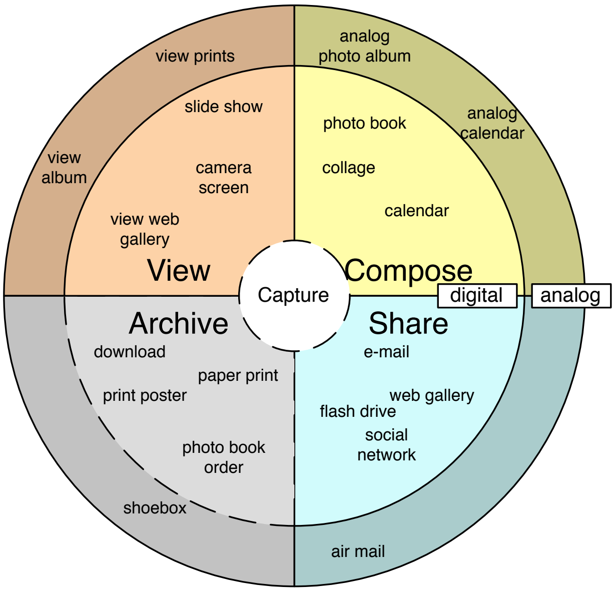


Figure 4.1: Photo Usage event types and exemplary life cycle of a photo

be viewed on a computer or on a print. At the beginning of each life cycle stands the *Capturing* of a photo which is either be done by camera or a scanner. This leads to three kinds of stages for a photo which is visualized by three rings in Figure 4.1. A photo can always change from an inner to an outer ring, but not back. E.g. by printing a photo it is transferred into the analog circle, but a photo can not be transferred back into to the digital circle (we assume that by scanning a photo, a new photo is captured). The life cycle of a photo is then defined by a series of usage events in this model. An example is illustrated in Figure 4.1: A photo is first captured, then downloaded to a computer's hard disk, then used for assembling a digital photo album which is then ordered at a photo printing company. This photo album is then viewed by a group of people.

In the following we take a more thorough look into the four usage categories:

View

Viewing a photo can be done in various ways. According to our model we distinguish between a digital and an analog form. Viewing images digitally is usually done on a PC or some kind of digital device, like a cell phone, a digital photo frame or even the camera itself. Viewing a photo on a PC can be done through many different channels. The simplest form is viewing a photo or a series of photos with some kind of photo viewer software, like e.g. flipping through a set of photos in a shoebox of photo prints. These kind kind of viewing events can be distinguished in the number of people involved (e.g. alone or with friends) and the purpose (e.g. quickly scanning a set of photos to find a specific one or doing a slide show documenting a holiday for friends). A special case of viewing is when viewing a photo with a group of people. By viewing a photo you also share it with them. In our model we distinguish between sharing and viewing. However, when, e.g., watching a photo slideshow with friends, we can model this as two separate events: The first event is a share event *slideshow* which involves none or multiple view events for each photo (The slideshow could have been stopped before the last photo or a photo could be viewed more than one time).

Viewing a photo is a type of event which can tell us much about its semantics. Especially setting information about *if, how* and *why* a photo was viewed into relation to other photos can bear al lot of interesting information. We will specifically prove this assumption in the next Section.

Compose

By composition we mean the usage of the photo as part of a multimedia composition. Examples for this are slideshows, web albums, or photobooks. All of these compilations employ a set of photos to produce a new multimedia asset. Besides the information that the photo was used in a composition it is also interesting to know *how* this photo was used. As shown in [SBC06], the placement of a photo in a multimedia presentation tells us much about its semantics and importance for the author of the presentation, e.g. if a photo is bigger than other photos on a page than one may conclude that it is more important. A composition can be made both in the analog and in the digital world, e.g. a photo can be done virtually on a PC or by placing prints in a conventional photo album. A Slideshow can be compiled by placing image files in a folder or by organizing analog slides for a diascope.

It is hard to present a general model to capture all aspects of different uses in a composition. We therefore opted to define the more general aspects in the `Composition` class and provide the possibility to define subclasses of this class for a more detailed description. We assume that in a composition some kind of importance ordering exist on the photos embedded, e.g. based on the size of a photo in a photobook or the time duration of a photo shown in slide show. The position in this ranked list is designated by the attribute `rank`. Together with the attribute `nrOfAssets` the position in the rank can be determined. The attribute `type` gives a textual representation of the composition type. In Figure 5.3 we have given an example subclass for photobooks which defines if the photo was used as a cover photo, which percentage of the page is covered by the photos, and the overall number of photos on the page.

Share

We consider sharing of photos as all activities where photos or photo compositions are handed over or made available to another person or a group of persons. This can, e.g., be done by sending photos as an attachment of an email, loading them up to an online community such as Locr, or Flickr, or physically handing them over to another person by handing over a burned CD or downloading them from a friend's camera. Most of these activities are hard to track, especially when no dedicated software application is involved in the relevant processes. For example when copying some photos on a friend's flash drive it is hard to infer that this is for sharing the photo with others rather than for making a personal backup. However in our model we try to cover the aspects that seem to be most interesting subsequent usages. Again we provide an attribute `type` to specify the kind of sharing activity, e.g. e-mail or upload to a web site. What is particular interesting is how the photo was shared, that means if it was made publicly available to an unknown audience, to a small group of users, or only to a single person. For this we have defined three levels: private (1), restricted (2), and public (3). With the `URL` attribute a location for the shared photo can be given,e.g. the URL of a photo community site where the photo was uploaded.

Archive

As Archiving a photo we see all activities which preserve a photo in any form. A special form of archiving is printing a photo which is the only way in our our model to transfer a photo into the analog world (indicated by a dashed curve in Figure 4.1). This can be the printing of photos on a printer attached to a home PC or ordering prints or other photo products from a professional photo services provider. The process is heavily intertwined with composition processes, e.g., authoring a photobook on a PC (composition) and ordering the result at a photo finisher company (printing).

As with composition activities it is on the one side interesting *if* a photo was printed and how often, but on the other side also *how* this was done. When ordering photos from a photo finisher interesting information are the size of the print, the kind of order, e.g., a cup, a t-shirt, or an ordinary paper print, and how often it was ordered. By setting this information in relation the other items of the same order, one can infer additional information, e.g., if a specific photo was ordered in a larger size than the others or if more copies were ordered than from other photos in the same order. On the other hand photos can be printed on a home printer. Basically the same attributes apply here.

The usage history of a photo can be modeled as a series of usage events according this model. In Section we will present a formal and practical definition of this model. We have extended the categorization of [FKP$^+$02] to reflect the additional dimensions photowork and end use. Additionally we have illustrated the life cycle of a photo which potentially ends in a final end usage such as a print or a photobook. This categorization is shown in Figure 4.1. Additionally, as we are not only interested in sharing activities of photos but in general photo usages, we also consider additional usages not directly related or targeted at sharing such as editing.

4.2 Photo Annotation from Photo Usage

Having identified potential interesting metadata about a photo's life cycle we now want to give some examples of the potential uses of this additional information. We first introduce a framework to define custom ratings for photos based on our introduced metadata usage model. After that we give some concrete examples to employ the usage history of for the automatic creation of photo slide shows and the automatic creation of a yearbook.

Generally a photo can be rated by evaluating the different usage information attached to the photo. Suppose there are n different types of usage events $\mathcal{E}_1 \ldots \mathcal{E}_n$ and let $p \in P$ be a photo out of a set photos P and $e_i \in ev_i(p)$, where $ev_i(p)$ denotes the set of events of type i associated with a photo p. If the event $e_k \in \mathcal{E}_k$ is defined by l attributes then let $a_k : \mathcal{E}_k \rightarrow \mathbb{R}^l$ be a mapping to an equivalent real valued vector. A rating for such an event e_k is then defined as $R_{vk}(e) = v_{0k} + \sum_{i=1}^{n} v_{ik} a_k(e_k)_i$, where $v_k \in \mathbb{V}_k = \mathbb{R}^{l+1}$ is a vector of parameters to customize the rating function. An overall rating for a photo p for a parameter vector $v \in \mathbb{V} = \mathbb{R}^k \times \mathbb{V}_1 \times \ldots \times \mathbb{V}_k$ is then defined as:

$$R_v(p) = \sum_{i=1}^{n} (\sum_{e_i \in ev_i(p)} (R_{vi}(e_i)) |ev_i(p)|^{-1} + v_{i0} |ev_i(p)|)$$

The parameters v_{i0} denote, how the number of events of type i a photo was involved does affect the overall rating. A simple application for this is to set all components of v_i to 0 except the v_{i0}. Then the corresponding rating function would not take into account what the parameters of a specific event were, but only how often this events occurred during the life cycle of the corresponding photo.

This general model is the basis for the derivation of concrete semantic annotations from image usage. In the next section we will apply this model to our use case of viewing a photo slide show.

4.3 Examining photo usages

Having introduced our general model for image usage in the following we want to analyze a specific walkthrough through this model for its potential to hold semantic information about the photos. As shown by [FWK08] the main usages of photos are end uses involving more than one person. Thus, we chose to analyze a typical scenario involving photos, the presentation of a photo slideshow. We will analyze several measurements of this slideshow and relate them to their importance for the owner. The goal of this experiment first to understand if and how these measurements relate to semantic information. From this we aim to derive which kind of information is useful to be captured and how this information can be combined with additional context or content information to derive meaningful semantics.

4.3.1 Photo Slide Show

One typical form of co-present photo sharing are slideshows where one or more persons present photos to an audience, usually as a mean to present a specific event such as a holiday trip or a family party [FWK08]. In the analogue days this have usually been conventional slides presented with an diascope. Today digital images are usually shown on a home PC, a laptop computer which can be connected to a TV, or via a projector. We

	Lanzarote	Crete
Nr of Photos	578	447
Selected for Album	272	140
Guests	29m,28w,25m	56m, 54w, 25m
Presenter	29w	29m

Table 4.1: Data and Participants of slide show user study

	mean viewing time (clustered)	
	selected	unselected
Lanzarote	6,1 (7,3)	5,8 (5,4)
Crete	7.0 (8.0)	6,1 (5.6)

Table 4.2: Mean viewing durations in seconds for the two test sets

assume, that the way people deal with a specific photo at the slide show correlates with its importance for the user.

Setup

We conducted a user study with two distinct user groups and two photo sets. The details are shown in Table 4.1. All photos were more or less unselected photo sets, only some obviously bad photos had been deleted directly on the camera. We wanted to observe a common situation were photos are shown to friends and family to document a holiday trip. In both experiments a host invited several people to his/her home to show them photos on his/her laptop in the living room. The hosts were operating the laptop and were skipping through the photo set while talking about the holiday. The photos were viewed with a modified version of an open source photo viewer. With this we recorded how long every single photo was viewed on the screen. Our hypothesis was that the longer a photo was viewed the more important it is to the person. To prove this hypothesis we compared the view durations times with photos selected from the same data set for compiling a photobook by the host. We assumed that photos selected for the photobook were more important to the user than photos that have not been selected. Thus we took the information if a photo was selected or not as the ground-truth of the semantic annotation important or not important.

Figure 4.2a shows the setting where the experiment took place and an example of a page of a photobook containing a subset of the photos show in the slide show.

Results

Figure 4.3 shows the results of the photo viewing times and an indication if a photo was selected for a photobook or not. The slideshow of the first experiment lasted about one

(a) Setting (b) Photobook

Figure 4.2: The setting of the experiment and a photobook containing a subset of the photos shown in the slide show

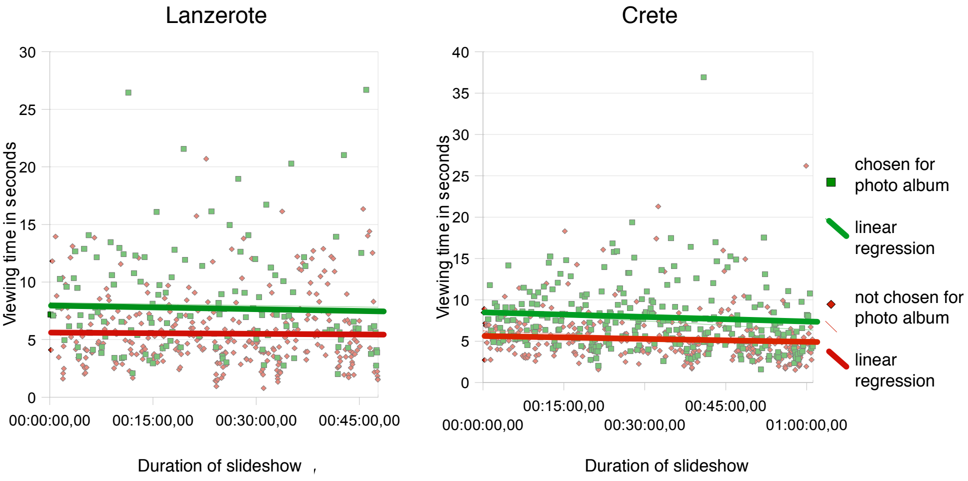


Figure 4.3: Viewing durations of photos chosen and not chosen for photobook

hour and the slideshow of the second about 45 minutes. For both data sets two regression curves indicate the mean viewing time for every photo. In Table 4.2 the mean viewing times of the photos distinguished by the fact if they have been selected for the album or not is shown.

Analysis

Looking at the regression curves in Figure 4.3 we can observe that it is decreasing for both classes in both test sets. This effect seems reasonable as we might expect, that viewers tend to get bored to the end of a slide show and skip faster through the images. The second obvious result is, that the regression curves are more or less parallel, but are separated by an offset of several seconds. This is a hint for a possible correlation between the viewing duration for a photo and the choice for a selection for a photobook. However, it seems impossible to decide if a photo should be selected for a photobook based on the viewing time alone. This assumption is backed by the result of a Fisher test which results in a p-value of > 0.05.

A way to overcome this problem is motivated by an interesting observation during the experiments: A specific event, such as a landmark or situation was usually documented by several photos. When conducting the slide show, the host often stopped at the first or one of the first photos of this event and took it as a handle to tell an interesting story or provide further details about the event or place. When skipping further through the slide show often better or more appealing photos of this event were shown, but the host did not stop at these photos as he had already talked about the event. This lead us to the idea, rather than assuming a correlation between viewing time and the choice if a photo was selected, to assume a correlation between the viewing times of an event and the choice if this event was documented in the photobook or not.

To test this hypothesis we first performed time clustering on the two photo sets according the algorithm presented in [Pla00] to determine separate events in the data sets. This resulted in 58 events in the Lanzarote data set and 71 events in the Crete data set. We then assigned to each cluster the viewing time of the longest viewed photo in the respective cluster. Based on this viewing time, each cluster was classified into belonging to one of three groups:

- *long viewed:* Viewing time is above or equal to the third Quartile ($t >= Q_3$)
- *medium viewed:* Viewing time is between first and third Quartile ($Q_1 > t > Q_3$)
- *short viewed:* Viewing time is below or equal to the first Quartile ($t <= Q_1$)

The clusters were additionally classified into three classes regarding their representation in the photobook:

- *strongly represented:* The percentage of selected photos in the cluster is above the mean percentage over all clusters.

		Crete	Lanzarote
Nr. of clusters		71	58
Photos per cluster Ø		6,3	9,97
selected photos per cluster Ø		1,97	4,69
long viewed cluster	strongly represented	3 (37,5%)	7 (63,6%)
	weakly represented	5 (62,5%)	4 (36,4%)
	not represented	0 (0,0%)	0 (0,0%)
medium viewed cluster	strongly represented	18 (66,6%)	8 (44,4%)
	weakly represented	5 (18,5%)	9 (50,0%)
	not represented	4 (14,8%)	1 (5,6%)
short viewed cluster	strongly represented	20 (55,5%)	13 (44,8%)
	weakly represented	5 (13,9%)	4 (13,8%)
	not represented	11 (30,5%)	12 (41,4%)

Table 4.3: Results of time clustering analysis

- *weakly represented:* The percentage of selected photos in the cluster is below the mean percentage over all clusters.
- *not represented:* No photo of the cluster is included in the photobook.

The results are shown in Table 4.3. The first obvious fact is that every long viewed cluster is represented in the resulting photobook at least to some extend in contrast to the medium and short viewed clusters where some clusters are not represented at all. Thus, we can conclude for the two test sets that *if events are viewed comparably long during a slide show they will also be represented in a photobook of the same event*. However, the opposite is not true: *If clusters are viewed comparably short it does not necessarily mean that they are underrepresented in a photobook of the same event.*

To conclude our analysis, at least on the basis of our our two test sets the viewing time alone is not a valid hint for the importance of photo to a person. It can only act as an additional means in a system of combination of other context- or content-based features to support the derivation of higher-level features such as the importance of a photo to a user.

4.4 Summary

In this chapter we have given a brief overview of related works in the field of photo usage analysis and have proposed a model which categorizes different kinds of photo usage and based on this different life-cycles of photos. This model distinguishes between the categories view, compose, share, and archive on the one side and digital and analog on tho other The history of a photo can be modeled as a series of events belonging to on of these categories.

We have also given the hypotheses that this photo history can act as a means to derive higher-level metadata for a photo. This hypotheses has been proved by conducting two experiments which compared the duration a photo was viewed during a slide show with the fact, if this photo was included in a photobook of the same event. The presence in this photo book was taken as an indicator if this photo is important to the person having compiled the photobook or not.

Based on these experiments we come to two conclusions: First, the viewing time is not a sufficient indicator for the importance of a photo to a user. It is however a cue for the importance of the underlying event. Second, at least in our our experiments the viewing time alone cannot be taken as the only feature to decide about the importance of a photo or a event.

This has two important impacts for the use of this usage information for the selection of relevant photos: A framework for the selection of photos has to consider not only features of single photos but has to be aware of related photos. This might different for other scenarios and other usage information, but we can at least conclude that we should always consider the context of the single photo. Second, usage information alone seems not to be sufficient for the automatic selection of photos. However, it can be an additional means, in combination with other metadata, to enhance quality of the selection of photos for photobooks.

In this chapter we have only considered the viewing time of a photo in a slide as a form of usage, which has however been proven to be useful for the semantic annotation of photos, in our case to gather information about the importance of photos. This is a reasonable motivation to capture and preserve such data.

5 Structural and Semantic Analysis of Photobooks

Photobooks have for decades been a vehicle to present photos in a pleasant and organized manner. There is evidence, that these photobooks are not designed by chance, but that authors aim to support the underlying story with the design. In addition to conventional paper-based photobooks, the authors of digitally designed photobooks are faced with much more degrees of freedom for the design, e.g. cropping and resizing photos. It is important to understand, how people actually use these various options to be able to provide methods to support the process. Especially important is to understand, how a *good* photobook distinguishes from a *bad* photobook.

In this chapter we will approach this question by analyzing a large amount of real world photobooks. With CEWE Color as a project partner we have access to a large number of structural representations of photobooks as well as their pixel-based content. By exploiting a large number of photobooks we can strive to understand how people design their personal photobooks. Furthermore, we can learn what kind of photos actually find their way into the album.

The results of our analysis are meant to drive the development of methods for the automatic content selection and design of photobooks in the remainder of this thesis. The analysis is split into two parts: The first part focusses on the structural analysis of photobooks. Thus, we aim to gain more insight into typical characteristics of photobooks regarding the placement, number and and size of photos and texts, the amount of pages, and other structural characteristics.

In addition, we seek to get to know more about typical photobooks on a semantically higher level. For this we develop a semantic classifier for photobooks which distinguishes between different types and by this also find out, which features enable us to distinguish between these different types. A second semantic classifier developed in this chapter detects semantic entities in a photobook which typically relate to separate events. These two semantic classifiers are partly used find out semantic differences between between different types of photobooks. In particular we aim to answer the following questions:

How social are photobooks? We are interested in the question of how social aspects are represented in photobooks. Thus, how are social relations between people reflected in the photobook, how prominent are people shown in the book and are there differences for different types of photobooks. **How expressive are photobooks?** By expressive we mean, how explicit or well one or more events are documented in photobooks. A photobook can, e.g. be not much more than a collection of more or less unrelated photos, or it can be a detailed documentation which is which is supported by how the user has designed the photobook, how he has placed and structured the photos and how he has annotated them.

How vivid are photobooks? Photobooks can be either quite uniform and factual in their overall impression or more vivid and informal. One might suspect, that this is different for different purposes or classes of photobooks, e.g. a book documenting a party might be more vivid then a professionally designed photobook of a wedding.

We start by giving a description of the photobooks in general and the characteristics of our dataset and our overall analysis approach in the next section. Section 5.2 describes the results of the structural photobook analysis. In Sections 5.3 and 5.4 we introduce the development of semantic classifiers for photobook types and sub-albums which are partly used in Section 5.5 approaching the questions above. The chapter closes with a summarization of implications for the remainder of this thesis.

5.1 Digital Photobooks

Personal photobooks have always been a popular way for organizing photos in a pleasant way and to preserve memories and share them with others. For this people do not only place photos in a book, but decorate them with additional snippets like text annotations or page decorations. Have glue, scissors, and pencils been the tools in the analog days to assemble a photobook, the process has become digital nowadays. The user gets support by authoring tools like the CEWE photobook application [1] to digitally master photobooks. Such authoring software allows to arrange digital images on the pages of an album, add textual annotations, and design the book with the preferred colors and style. In the end the user can order a print from a photo finishing company. A typical example of a double page of such a photobook is shown in Figure 5.1. Such a double page typically consists of both photo and text elements which can occur in various sizes and orientations and can also overlap each other.

Creating a photobook is a form of digital story telling that reveals much about the user, the album and the different parts of the album. By authoring the photobook the user has implicitly enriched the photos, the single pages and the book as a whole with different kinds of semantics: She or he has established relations between photos, texts and pages. Photos and text become more prominent than others by their size and placement in the book, which may reflect their importance for the user. Some photos are clustered into groups or put on special pages which might allow to draw conclusions about relations between photos and their semantics. We aim to reveal these hidden semantics by analyzing authored photobooks and the contained photos. Basis for this are the CEWE photobook software and photobooks which have been digitally authored and ordered.

5.1.1 Analysis Concept

We are dividing the analysis of photobooks in this Chapter in into two steps: a *structural* and *semantic* analysis. Photobooks, as present at CEWE Color, are already available in a well-structured form: We can clearly distinguish between pages, text areas and image

[1] `http://www.smilebooks.com`

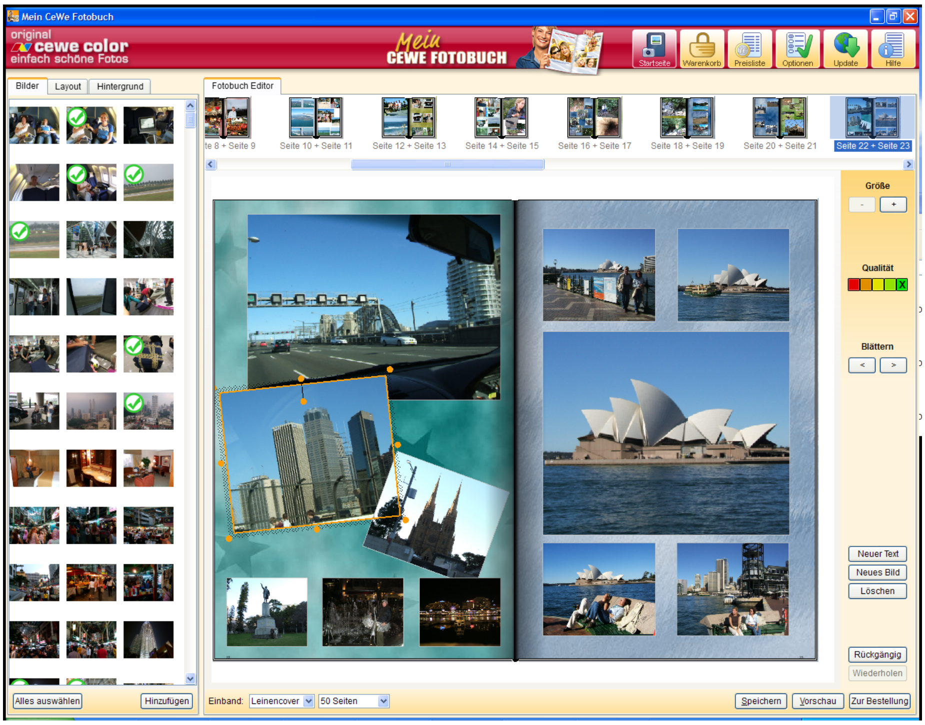


Figure 5.1: Example of a typical photobook double page with photos in different sizes and orientations

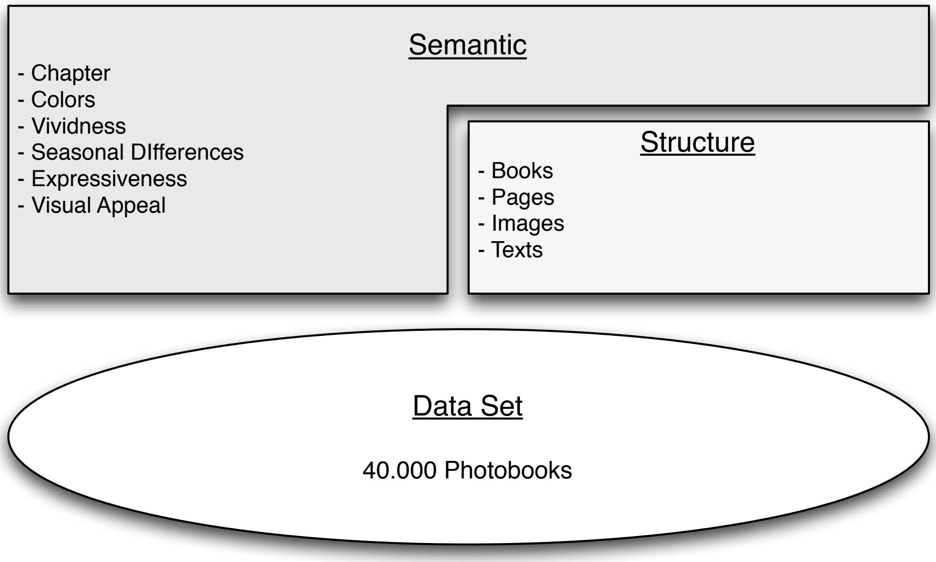


Figure 5.2: Concept for structural and semantic analysis of photobooks

areas on the pages. Details are given in Section 5.2.1. Based on this structure we are aiming to learn more about the characteristics of real world photobooks by statistically analyzing the values of these characteristics in our data set.

Such a structural analysis has the advantage of giving an overview about directly measurable characteristics of a photobook. However, many aspects are not captured this way. The given structure, e.g. does not directly tell us how and if a photobook is divided into several semantically distinct parts or chapters or how much effort the user spent in designing the photobook. Such semantic characteristics cannot fully reliably be extracted but are still very valuable cues to better understand how people design photobooks. The second part of this chapter thus aims at developing methods to extract such semantic characteristics of photobooks. This is partly done by employing results from the structural analysis. An illustration of this approach is given in Figure 5.2.

5.1.2 Test Data Set

Basis for our analysis are about 44.000 photobooks which have been ordered at CEWE Color. The photobooks have been authored and ordered by users with the help CEWE photobook application in the time period from 3/2008 to 8/2010. From all orders in this time period, a portion of the daily orders of the day are selected and added to our test data set. The photobooks originated from all over Europe but the majority of orders come from Germany.

For our analysis, the photobooks have been completely anonymized, that means that all information about the person having placed the order has been removed from the photobook. Also, to maintain the privacy of the test data, our semantic analysis runs only on photo features, the photos themselves are removed from the data set. Thus, our test data set consists of all structural photobook information and the extracted photo features.

CEWE Color offers a wide variety of different kinds of photobooks ranging from low budget soft cover albums for a couple of Euros to premium books printed on photo paper in linen or leather cover. The photobooks of our test data set are not selected according to this criteria and therefore this variety is also reflected in our test data set.

5.2 Structural Analysis

One goal of this chapter is to gain insights in the general characteristics of photobooks. For this we look at the different semantic levels of a photobook separately. This means we take a look at the photobook as a whole as well as the single entities such as a single page, photo or text description. In the following we perform a statistical analysis of features on each of these levels. Which statistics have been chosen is mainly motivated by explorative analysis of the album features.

Besides the structural photobook information our data set consists of information about the authoring process. This means we know how long the user spent authoring the album and which actions he performed during this time. This information can also help us to understand the photobook and the user authoring the album and we will thus take a look at some aspects of the authoring process.

5.2.1 Structural Photobook Model

Photobooks are represented by a structural model capturing their content and layout over the pages of the book. This model serves as the input for the semantic analysis and directly represents the structural representation of photobooks ordered at CEWE Color. Figure 5.3 illustrates this structure in a UML model. It is divided into two parts, representing the photobook as a whole, the contained photos and their associated metadata.

A photo is defined by a number of features which stem from the photo's content and context together forming the *Photo Metadata*. These features are either directly extracted from the photo or derived by analyzing and combining other features. Extracted features are, e.g., the capture time and location or the camera extracted from the photo's Exif header. Derived features are, e.g., the percentage of photos that contain faces or the brightness of the photo derived from the brightness histogram of the photo. A more thorough description of our framework for content-based and context-based metadata enhancement for photos is given in Chapter 6.

A *Photo Album* consists of one or more pages which each holds an arbitrary number of photos and text areas. A page can either be an album cover, the book spine or a regular

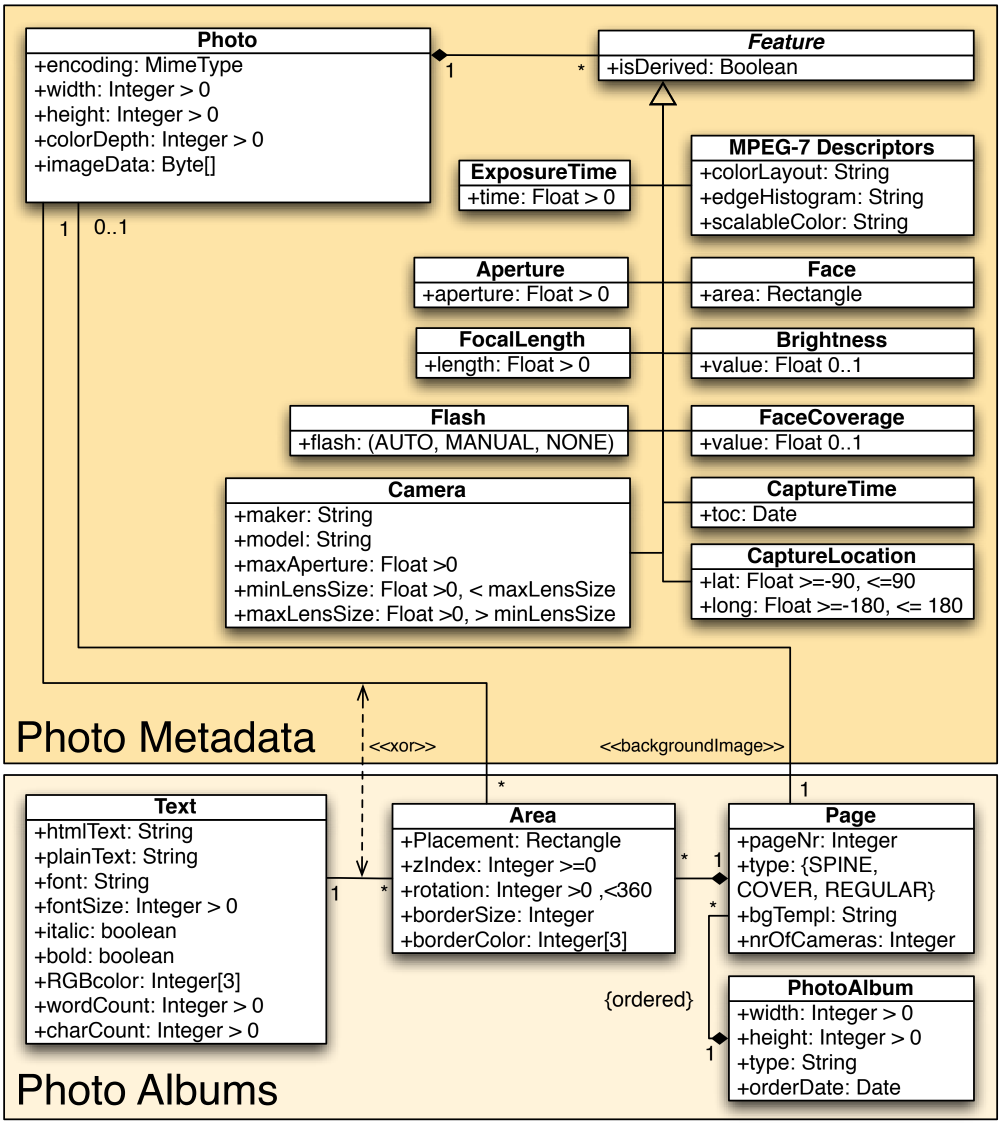


Figure 5.3: Photobook Model (UML class diagram)

page. It may hold a single image acting as the background. Photos and text areas can be of arbitrary dimension and can be rotated and overlap.

5.2.2 Album Level

Looking at the photobook as a whole helps us to understand characteristics of photobooks in general. The main aspects we looked into are the number of pages, the size and kind of the photobook, the time span covered in the album, and the usage of cameras for the photos in the album. In the following we take a deeper look into the latter two aspects as this happened to be the most interesting.

Album time span: One of the central aspects is the time period that is covered by the photobooks, that means the difference of the time stamps of the oldest and newest photo in the album. We performed a temporal analysis on all albums in our data set. A plot of the probability density function of album time spans is shown in Figure 5.4. This plot also shows the 30 quantile, the mean, and the median time span. The most obvious result is, that the time spans are not evenly distributed, but show a strong peak in shorter time spans. 30% of the albums have a time span of 26 days or less and 50% a time span of 276 days or less. Really short term events of only a couple of days are rare: Only 10% of the albums have a time span of up to 3.5 days. If we think of events such as a wedding or a birthday party than photos albums documenting such events are either very rare or these events are only documented as part of an album documenting several separate events. Interesting to see is that the peak in the density plot is at about 26 days as this is a typical time period for a long holiday trip which could be worth to be documented in a photobook. The rest of the albums is distributed over a rather broad time range. This seems reasonable as we can expect that the variance in the time spans of long term albums is much higher than the variance in short term albums.

Number of cameras: Another interesting feature is the number of cameras used for the photos in an album. To a certain degree we can derive the number of people having contributed photos to the photobook when assuming that every person only owns one camera or one camera is not shared among a group of people like a family. We know that this is a strong assumption, but from our own experience and from interviewing people in our group we found out that people rarely own more than one camera. The presence of more than one camera in an album therefore at least seems to be a good indicator if more than one person was involved contributing photos to the photobook.

To determine the number of cameras we looked at the number of distinct values for the camera information in the photos' Exif headers of one album. To minimize the error we preprocessed the data set by removing photo without an Exif header or a missing value in the camera field which, from our analysis of the data sets, is usually an indicator has been edited or scanned from a print.

The mean number of cameras was 2.8 which is a strong sign that not only one's own photos are used for an album but photos are shared among others. The photobook could be a compilation of photos from attendants of a holiday trip or the author of the photo

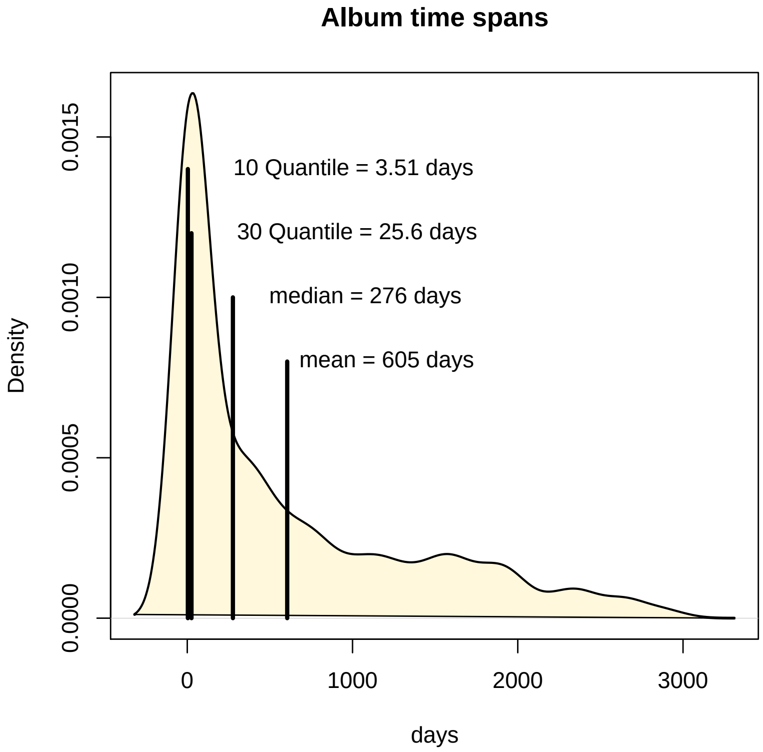


Figure 5.4: Probability density function of album time spans

could have added single photos from photo sharing sites or other sources to enrich his album.

5.2.3 Page Level

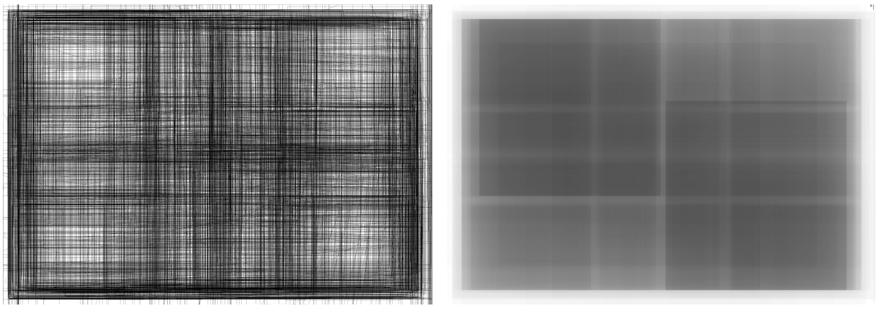
Besides looking at a photobook as whole, we are interested in characteristics of the single album pages. The most obvious characteristic of an album page is what kind of contents it holds. We analyzed our data set regarding the number of photos and text items on a page.

The number of photos and text items are a way to distinguish between different types of albums, e.g. a holiday album may hold more images on a page than a baby album.

Text items per page: The mean number of text items per page is 0.5. Only about 30% of all pages do contain text. This may be hint that only special pages are annotated with text and these pages may be a way to identify the beginning of a semantic unit in a photobook. Also, only 10% of the pages do contain more than one text item. This may be a hint that people tend to annotate pages as a whole rather than annotating single photos.

Images per page: The mean number of photos per page is 2.9 and the majority of pages does not contain more than 3-4 images. This shows us that users, despite having the possibility to place much more photos on page, prefer to have their photos been shown comparably large in a photobook.

Page Layout

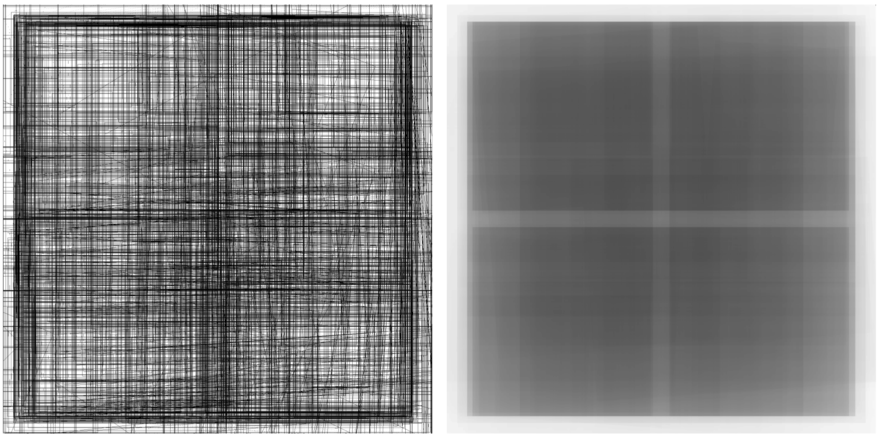


(a) Cumulated Borders of Photos

(b) Cumulated Photo Areas

Figure 5.5: Distribution of borders and areas of photos for XXL CEWE photobooks

Besides knowing how many elements are present on the single photobook pages it is important for the development of automatic photobook layout methods to know how they are placed in typical photobooks. What we are looking are typical patterns in the individual page layouts which can lead to rules for the automatic layout. However, it is hard to automatically derive such patterns by analyzing thousands of photobook pages.



(a) Cumulated Borders of Photos (b) Cumulated Photo Areas

Figure 5.6: Distribution of borders and areas of photos for Small CEWE photobooks

Thus, we opted for a manual approach by visually analyzing the layout of the photobook pages. For this we built separate border and weight maps for each type of photobook: For the *weight maps*, we first built a mask image for every (inner) photobook page in our dataset by setting pixels which correspond to a point in the photobook page to black and the remaining pixels to white. We then built the mean image from all of these mask images resulting in a weight map. We did the same for the *border maps* but for this only set the borders of the photo areas to black and leaving the inner part of the photos white.

Two results of such cumulated maps are depicted in Figures 5.5 and 5.6. These maps correspond to the two extremes in the portfolio of CEWE Color: XXL CEWE photobooks are the largest and most expensive photobooks wheres SMALL CEWE photobooks are the smallest and cheapest photobooks.

Looking at Figure 5.5b we can observe, that there is a strong accumulation of photo borders both at the borders of the photobook page and in the middle horizontal and vertical line. Additionally there also strong accumulations both in the horizontal and vertical axis which correspond to the golden ratio. This is an interesting finding as the golden ratio is one of the fundamental design principles (for details see Chapter 9) which obviously is also present in the design of digital photobooks. What we can generally observe is that users seem to prefer layout which do not incorporate tilted photos. Looking at the cumulated photo areas of the same photobook type in Figure 5.5b we can additionally see, that people seem to favor photo being places in the top left or bottom right part of the photobook page.

The same observations are generally also true for the smaller and cheaper photobook type shown Figure 5.6. However, the presence of the golden ratio is a lot less obvious. One explanation for this might be that these types of photobooks much cheaper than the XXL variant and thus less effort is done by the user to carefully arrange the individual pages. This might also be an explanation why the accumulation of photos in the top left and bottom right part of the pages is not present in the Small variant of photobooks. We found out that text is significantly more often present in higher price photobooks than in budget photobooks. The top right and bottom left parts are typical places for textual annotations for the images on the page. This can not be found in the cheaper photobooks which most do not contain any text at all.

5.2.4 Texts

Besides photos the pages of an album can contain one ore more text items which further describe one or more photos or one or more pages. We have already shown that only about 30% of the pages contain text which is a strong sign that such a text description is a way to emphasize specific parts of an album. One of the main characteristics of text descriptions is their length in words and their font size. Knowing these features may be a way to further semantically categorize text items as, e.g. a page or image description. Thus we analyzed the text items in our test data set regarding these two features. The median number of words is 5 and 30% of the text items contain up to 3 words. Thus, descriptions seem to be rather short in albums and only very briefly give additional information to the album contents. Text items with more words may designate a different kind of annotation, e.g. diary type text describing a specific day or place of a holiday but not solely describing a single photo.

In this context it may also help to take the font size of the text into account. We found out, that the median text size is 22 and that 30% of the test items have a font size of at most 16. Thus, these seem to be typical sizes for general annotations in a photobook. A text with a font size significantly exceeding these values may be a candidate to be a more important annotation such as an album title or the title for an event in an album.

5.2.5 Authoring Process

Insights into the authoring process of a photobook gives additional insights into the photobook and its meaning for the user. One important aspect is how much time the user spent to author the album as this gives us a hint for the importance of the album for the user: An album for which the user spent a couple of hours to author may be more important to him than an album only quickly put together in a couple of minutes. However, we have to be careful with assumption as some users may be more skilled than others and might be able to author a comparably good photobook much quicker than others. Thus when employing this information for deriving the importance for the user we have to take into how much time the user spent for other albums.

Thus we analyzed the authoring times of the photobooks in our test set. The probability density plot of the according time lengths is shown in Figure 5.7. For the analysis we summed up the time spans of the separate authoring sessions and removed time slot with no activity for at least 10 minutes to compensate for times where the authoring tool was actually running on the computer but not used. The median of the authoring times is 3.3 hours. This shows that people generally put a considerable amount of effort into the authoring process.

The information about the authoring time enables us to estimate how strong thoughts and intentions of the the photobook author are reflected in the photobook: Semantics derived from an album with a comparable long authoring time might therefore be more confident than from quickly authored albums.

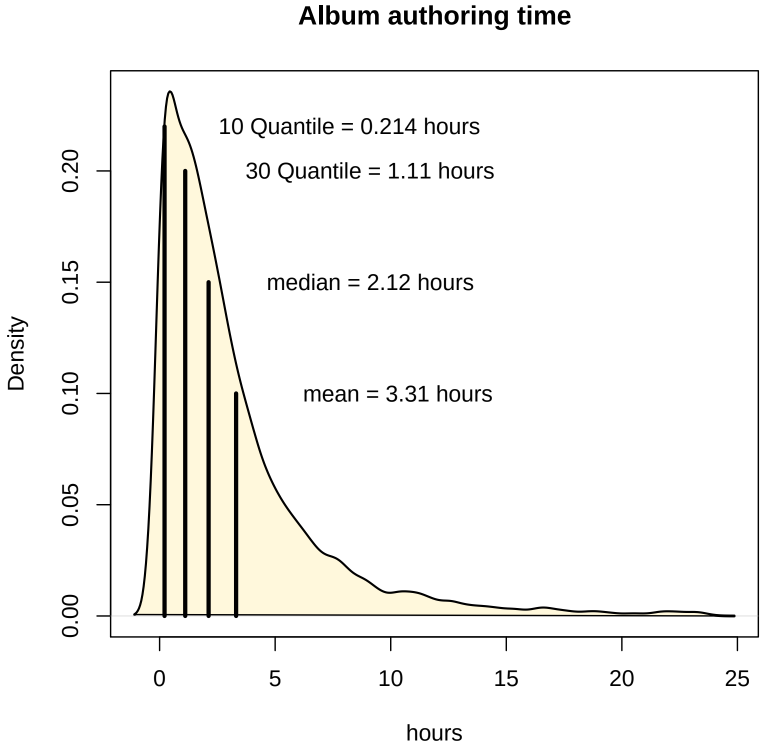


Figure 5.7: Time to author a photobook

5.3 Photobook type classification

One goal of our analyses is to detect differences between different types of photobooks. As shown in Section 5.1 these semantics are not directly present in the photobooks' structure of our test collection. Thus, we need a way to automatically extract these semantics for a large set of photobooks. In the following we describe the development of such classifiers by using a small portion of photobooks as a training data set.

The main requirement for the the reliable derivation of characteristics for a specific semantic label is the availability of a sufficient large, labeled data-set. This labeled data set is the ground truth for our analysis. Our problem is, that we are equipped with a quite large data, but this data set is not semantically labeled in any sense. E.g., we do not know if a photobook is documenting a holiday or a wedding without looking at the book manually.

To compensate for this, we opted for a quite pragmatic method: We choose those samples from our test data set for which we are quite sure about their semantics according to the values of one or more features. By this we select only a small portion of the test data set as training samples. Samples of this labeled set are manually inspected to avoid wrong labels. With the help of this labeled samples we determine additional features discriminant for the respective semantic label. By this we are able to derive rules for the semantic annotation of book parts, also for samples not fulfilling our initial characteristics. We are aware that this approach has two major drawbacks: The labeled training samples are not evenly distributed over the test data set and therefore we can not be sure that derived rules for this labeled set can be used for labeling the rest of the data set. On the the other side we can not be sure that the automatically determined does contain wrongly labeled samples. However, the manual inspection of samples of the automatically labeled data set determined by our approach has shown that that our method is feasible.

From our and CEWE COLOR's experience we know that there exist some typical types of photobooks which are often ordered. Our goal was to determine features to distinguish between these different kinds of albums. For this we determined a set of labeled albums by selecting them by characteristics typical for specific kinds of photobooks. We then analyzed these labeled photobooks for additional discriminative characteristics.

Assumptions on the Data Set

The most typical albums according to CEWE COLOR are travel albums and albums documenting a party-like event, e.g. a birthday party or a wedding. Thus we opted to analyze our data set according to these typical event. Specifically chose three event types: a wedding and a birthday party and photobooks documenting a journey. We assume that we can select a considerable large subset of all albums belonging to these classes by looking for typical keywords in the title of the photobooks.

5.3.1 Ground-Truth Determination

We chose a quite pragmatic approach to select albums documenting such an event: We looked for typical keywords of on the title pages of the photobooks. Photobooks ordered at CEWE Color origin from countries all over Europe but the majority of albums is ordered from Germany. We therefore restricted the input data set to albums ordered from Germany and selected albums that contained of typical german keywords on the title page of the album. The selected keywords were *Hochzeit* (Wedding) for the wedding class, *Urlaub* (Holiday) for the travel class, and *Geburtstag* (Birthday) for the birthday class. We only looked for one very typical keyword for every class as our goal was not to select as much albums as possible but to select the albums that quite reliably belonged to the respective class.

Analysis

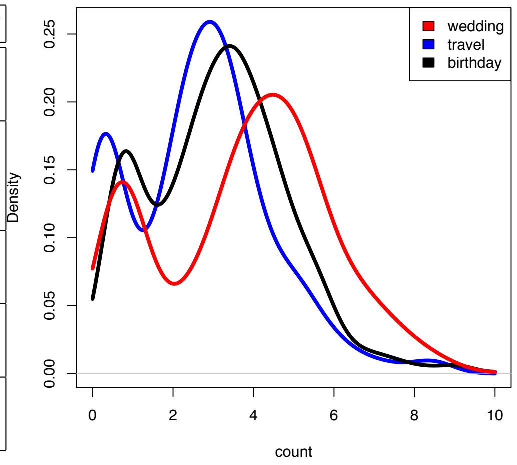
Our goal was to find additional features which significantly differ when looking different classes of albums. For the classification of albums we opted to choose the following features:

- Average **number of words** per page
- Average **number of images** per page
- **Album authoring time**: How long it took for the author to compile the album.
- **Time Span** of the album: The time span which is covered in the album, retrieved from the the time stamps in the photos' Exif headers.
- Average **Number of Faces** per photo

We chose these features as we assumed that they would significantly differ for different classes. E.g., we assumed that a wedding album would show significantly more persons than a travel album.

Figure 5.8: Results (p-values) of student's t-test for significance of different features to decide about different album classes

	Wedding	Travel	Birthday
Average #Words	0.12	$< 10^{-4}$	0.09
Average #Images	0.004	$< 10^{-4}$	0.14
Auth. time	0.94	0.002	0.007
Time Span	0.28	$< 10^{-4}$	0.9
Average #Faces	$< 10^{-4}$	$< 10^{-4}$	0.0007



(b)

(a)

Table 5.9a shows the result of our analysis for the classification of albums. As in the determination of sub albums we performed a t-test for every album class and feature. A p-value lower than 0.05 designates a significant difference in the respective feature values.

5.3.2 Discussion and Classifier Training

One can see that most of the p-value are smaller than 0.05 which means that they are suitable to differentiate between the different classes of photobooks. The only feature that is significantly different in all classes is the average number of faces shown in a photo. Figure 5.9b shows the distribution for the respective classes. The average number of faces is 4 (wedding), 2.6 (travel), and 3.1 (birthday). The album time span is only significantly different for the travel class. This seems reasonable as journeys usually cover time periods of several days or weeks while birthday and wedding events are only single day events. Another discriminative feature to distinguish between travel and other album classes is the number of words per page. We observed that a travel album on average consists of 10.3 per page while a wedding album consists of 6.3 and a birthday album of 6.5 words. This, again, seems reasonable as we observed that travel photobooks often contain rather long text passages in a diary-like manner. We used the labeled data set to train a Multiclass Naive Bayes classifier. The resulting multi-class classifier showed an accuracy of $79,46\%$.

We can conclude that, for a limited number of classes, we have identified features which quite reliably determine the type of a photobook. The results show that our approach is feasible and we can expect that the accuracy can be increased more by tuning the parameters for the training of the classifier or by considering additional features. On the other hand we can expect that the accuracy of our classifier will decrease if we add additional classes of photobooks to it.

5.4 Sub-Album classification

Photobooks usually document one or more events of a specific user. As the chapters in a book, these events can usually be divided into several sub-events. Thinking of a 4-week holiday trip this could be different places or days or in the case of a wedding album the ceremony and the evening party. In the remainder we will refer to this sub-events as sub-albums. We observed that the borders of such sub-albums are usually on page borders. Our goal was to find features or combinations of features in photobooks which designate such event borders on page breaks.

5.4.1 Assumptions on the Data Set

As already mentioned we assume that events in a photobook usually happen on page breaks. This assumption is backed up by our own experience and experience of our industrial partner. This assumptions enables us to model different events in an album as sub-albums, where a sub-album is defined as a consecutive number of pages in an album. The other assumption we make on our data set is that events are usually defined by time. [RW03] has found out that time is the top-most mean to retrieve from a specific event and events in a photo collection can be determined by time-based clustering.

5.4.2 Ground-Truth Determination

Following our general approach for a ground-truth determination we filter our test data set for such samples that show clear signs to designate an event border. Thus, we filter our test data set to samples that at least fulfilled the following characteristics:

- All photos should be equipped with a time stamp from their Exif header.
- The time span of the album should be between two and 12 months as we assume that in long-term photobooks time cluster can be detected more reliable.
- The album should contain at least 60 pictures, 5 text items and 30 pages to be sure that enough features are available to reliably designate sub-album borders.

From this filtered data set we determined sub-albums by the following method: We employed the time clustering algorithm proposed in [PCF02] to cluster the album images according to their time stamp. The advantages of this algorithm are that it does not expect a number of expected clusters as input, such as in the k-means algorithm, and that it does not not take a fixed threshold to determine cluster boundaries but rather adapts this threshold to the time gaps in the data set. This algorithm is parametrized by a *window size* d which specifies how many many neighbor time stamps should be taken into account for the determination of cluster border and a second parameter K which tunes the amount of generated clusters. We adapted the algorithm by changing these two parameters to better suit the characteristics of photobooks: We chose second parameter K empirically to be $log(20)$.

For the window size d we chose a value of 6 as we assume that most event breaks happen before a double page and the median number of photos on a single photobook page is 3 according to the analysis in Section 5.2. We clustered the photos of the filtered albums accordingly with leaving out the photos on the title of the photobook as we experienced that these photos are usually taken randomly (according to their time stamp) from the input photo set. This resulted on the one hand in a couple of big clusters with more than 10 photos and on the other hand in several small clusters of one or two photos. The big clusters usually had their borders on page breaks which backs our hypothesis that event borders usually take place at page breaks. The smaller clusters however were either spread over a very long page span, e.g. the two photos of a cluster were put on page 2 and 90 of the album, or a group of small clusters was put on a single page. To compensate for this we filtered out cluster spread over a page range of more than double their cluster size to remove outliers. Additionally if the boundaries of two consecutive clusters did not coincide with a page break we merged the two cluster. Interestingly on average only $7,8\%$ of the cluster were filtered out in the first step. This shows us that most photobooks are temporally ordered. The result is a set of pages labeled with being the beginning of a sub-album or not.

	sub-album break	other pages	p-value
#words	5.99	3.3	< 0.0001
#images	3.65	3.8	0.084
Color Layout	50.83	41.6	< 0.0001
Edge Histogram	211.17	187	< 0.0001
Scalable Color	208.55	163.1	< 0.0001

Table 5.1: Average values for features for sub-album determination

5.4.3 Analysis

Having determined a labeled set of pages being the begin of a sub-album or not, we determined and analyzed additional features potentially indicating such a border:

- **number of words** on the page
- **number of images** on the page
- **picture similarity** between consecutive pages based on three different content-based similarity measures

These features have been chosen based on various assumptions based on exploratory analysis of sample photobooks. First we found out that a sub-event in an album is often marked by a separate text title. Second we observed that often the start of a sub-event is marked with one or two single photos on a page while putting more photos on the other pages. Third we assume that the photos inside a sub-event are significantly more similar to the rest of the photos in the album. That is why we calculated the difference between different pages according to three different dissimilarity measures based of content-based image descriptors adopted from the MPEG-7-Standard [MSS02]. We chose to employ the *color-layout* and *scalable color* descriptor to measure changes in the color scheme and the *edge histogram* descriptor to measure changes in the texture of images. The MPEG-7 standard also makes recommendations for suitable distance measure for the descriptors which we adopted. To compensate for outliers we chose to employ centroid-based cluster similarity. Thus for every page we determined the medoid image according to the different descriptors and distance measures and determined the distance between two pages by calculating the distance between these medoids accordingly. By this we calculated three content-based similarity measures for every album page to its preceding page.

A summary of the results of our analysis is given in Table 5.1. Figure 5.9 exemplarily depicts the distribution of page similarity based on the MPEG Edge Histogram Descriptor.

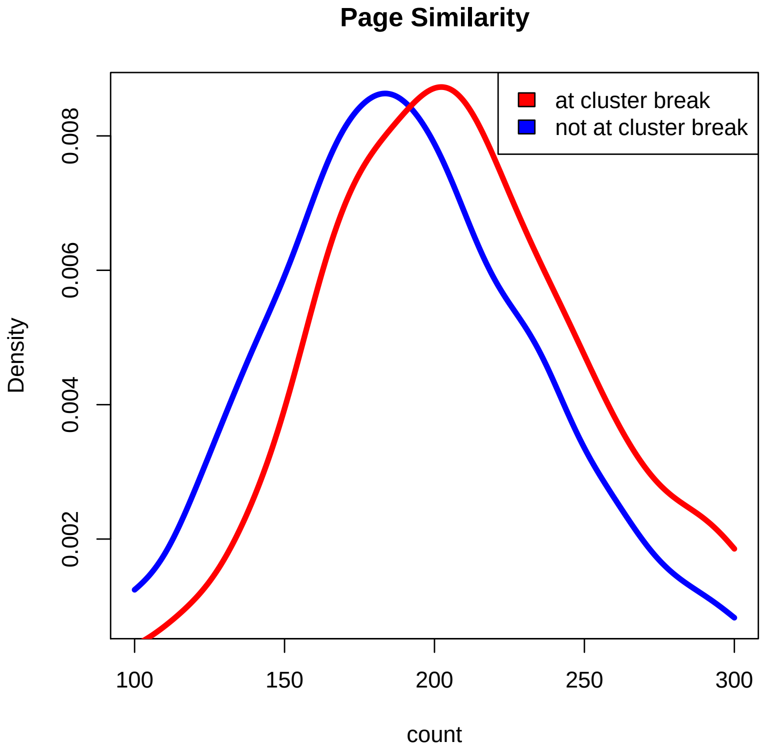


Figure 5.9: Page similarity based on MPEG-7 Edge Histogram Descriptor

5.4.4 Discussion

One can observe that all features but the number of images feature seem to be good indicators to distinguish between a page being the beginning of a sub-event or being a normal page. This is backed up by the result of a student's t-test. Usually a p-value under $0,05$ considered as showing a significant difference between to data sets. We experimentally trained a Naive Bayes Classifier based on these features which resulted in an accuracy of $82,45\%$. This is still not accurate enough to employ this classifier on our whole data set to determine boundaries of sub-albums, but it shows that event borders in a photobook can be detected even when not considering time as a feature. However we expect that the accuracy be even be significantly increased when combing our method with clustering based on the photos' time stamps. By this we can, to a certain extend, also quite reliably detect event borders when having to cope with invalid or missing photo time stamps.

5.5 Semantic Photobook Analysis

In the following we answer questions which help to understand photobooks on a semantic level. For some some of these questions we employ the semantic classifiers which were described in the last two sections.

5.5.1 Socialness

To answer the question if a photobook is also a *social* photobook or to which degree, we first have to define, what makes a photobook social or how we can measure the *socialness* of a photobook. The common understanding of social is referred to as the the kind and degree of interaction between two or more individuals. Thus, the socialness of a photobook can be described as the degree to social relations shown in the photobook. In addition, we can also differentiate these social relations on their degree of intimacy: How close are the relations of people in the photobook? Are they only strangers to each other or are they intimate friends or a couple?

We are aiming to at least some extent find answers to these questions by analyzing the contents of the photobooks of our test collection. As indicators for the degree of socialness we have identified a number of indicators which will be further explored in the following.

Number of Persons

The presence of persons in a photobook is generally a good indicator, that the photobook has a social aspect. The more people are shown, the more social the underlying event or story is. E.g. if a photobook shows many photos with people one may generally conclude that the photobook comprises of a strong social aspect. However if looking on a photo level, thus how many persons are shown in the photo, one may also be able to decide, how much the people are related to each other.

In Figure 5.10 the distribution of number of faces in a photo is shown. For this we employed the face detection algorithm proposed in [VJ01]. We found out that over 85% of all photos in our test collection contain faces. The presence of persons in a photo is usually a sign that it is a quite personal and emotional photo. Knowing how many photos in an album contain faces can in turn tell us more about the type of the album. One may assume that a rather emotional event such as a wedding would also contain a lot of photos showing persons. 5.10 shows the distribution over all photobooks for the average number of faces per photo. Obviously the majority of photobooks show an average of nearly two persons per photo.

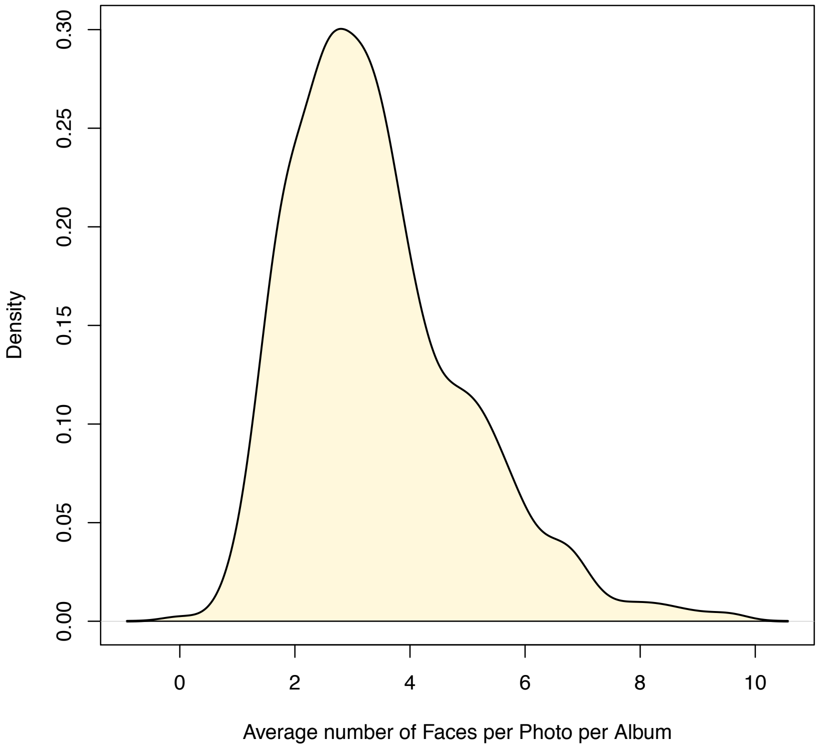
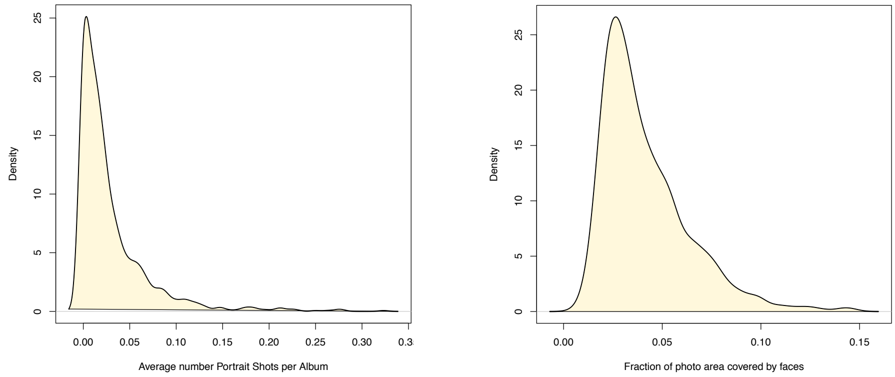


Figure 5.10: Distribution of average number of faces per photo per photobook

Looking again at the results of our initial analysis of features for the development of a classifier for photobook types, we are also able to derive semantics for different kinds of photobooks. Figure 5.9b shows that the number of faces significantly differs for different photobook types: On average, travel photobooks show less people (2.6) in the contained photos than photos in birthday (3.1) or wedding photobooks (4). This backs up the common intuition that, e.g. weddings in general incorporate much more social aspects than e.g. a holiday trip. Also our manual inspection showed, that almost all photos show people and often large groups of people e.g. in the wedding ceremony or at an evening party. On the other side, photobooks documenting a journey also show people, but to a much less extend as not only the social relations are documented, but often the main topic are, e.g. a nice landscape or famous landmarks or buildings. Thus, we can conclude that there is a significant difference between different kinds of photobooks regarding the presence of social aspects.

Person Dominance

Not only the number of persons in a photo is an important indicator for the socialness of a photobook, but, perhaps more important, how dominant they are shown in the photo or the photobook. Thus, one person, which occupies a large portion of a photo could indicate a portrait shot which would be considered as much more social than a photo which shows a large group of people, but only in the background.



(a) Fraction of Portrait Shots of albums

(b) Face Coverage

Figure 5.11: Face coverage and portrait shots

To cope with these differences we additionally determined the fraction of portrait shots taken over all photos in out test set. As portrait shots we have defined photos which show a single or two faces which occupy at least 20% of a photos' area. The result is depicted in Figure 5.11a. The majority of photobooks does not have more than 10% of photos showing portraits. Giving the average number of faces per photo (nearly 2) per photo photobook, this seems to be not much much, but backs up our visual inspection of the photobooks: Most photos showing people are showing their full body or additional objects besides the person itself. Typical examples are, e.g. photographs with one or two persons in front of landscape or a famous building. Figure 5.11b shows an overview over the area in photos showing faces which is covered with faces.

5.5.2 Intimacy

Besides at looking at the content of photobooks, one may also consider the origin of the contained photos as an indicator of the socialness of a photobook. If the photos in a photobook only originate from a single person, supposedly the author of the photobook, this can be an indicator of lesser social involvement than if the photo originate e.g. from a two or more people.

As a measure for the the diversity of origins of photos in a photo the number of different cameras can be taken. To a certain degree we can derive the number of people having contributed photos to the photobook when assuming that every person only owns one camera or one camera is not shared among a group of people like a family. We know that this is a strong assumption, but from our own experience and from interviewing people in our group we found out that people rarely own more than one camera. The presence of more than one camera in an album therefore at least seems to be a good indicator if more than one person was involved contributing photos to the photobook. To determine the number of cameras we looked at the number of distinct values for the camera information in the photos' Exif headers of one album. To minimize the error we preprocessed the data set by removing photos without an Exif header or a missing value in the camera field which, from our analysis of the data sets, is usually an indicator has been edited or scanned from a print.

The mean number of cameras over all photobooks in our test set is 2.8 which is a strong sign that not only one's own photos are used for an album but photos are shared among others. The photobook could be a compilation of photos from attendants of a holiday trip or the author of the photo could have added single photos from photo sharing sites or other sources to enrich his album.

5.5.3 Expressiveness

Often photobooks are not only a means to preserve the memory to events for only the owner and creator, but can also act as vehicle to express one's feelings and thoughts to others. This can be done done on various ways, either rather factual or more emotional. E.g. a photobook could have a very strict visual layout with none or only a few, but very precise and emotionless descriptions for the photos (like a photo of the Eiffel Tower with only the words *Eiffel Tower* as a description). A more emotional photobook would e.g. have a more casual layout (like tilted photos) and a more slang like language (*Wow, what a cool view of that old tower!*).

In this section we are aiming at analyzing this expressiveness of real world photobooks. For this we will mainly focus on the text portions of photobooks.

Image-Text Ratio

In a first step, we analyze the ratio of text and image items in the photobooks of our test set.

Text items per page: The mean number of text items per page in our test set is 0.5. Only about 30% of all pages do contain text. This may be hint that only special pages are annotated with text and these pages may be a way to identify the beginning of a semantic unit in a photobook. Also, only 10% of the pages do contain more than one text item. This may be a hint that people tend to annotate pages as a whole rather than annotating single photos.

Images per page: The mean number of photos per page is 2.9 and the majority of pages does not contain more than 3-4 images. This shows us that users, despite having the possibility to place much more photos on page, prefer to have their photos been shown comparably large in a photobook.

Text lengths and font sizes

Besides photos the pages of an album can contain one ore more text items which further describe one or more photos or one or more pages. We have already shown that only about 30% of the pages contain text which is a strong sign that such a text description is a way to emphasize special parts of a photobook. One of the main characteristics of text descriptions is their length in words and their font size. Knowing these features may be a way to further semantically categorize text items as, e.g. a page or image description. Thus we analyzed the text items in our test data set regarding these two features.

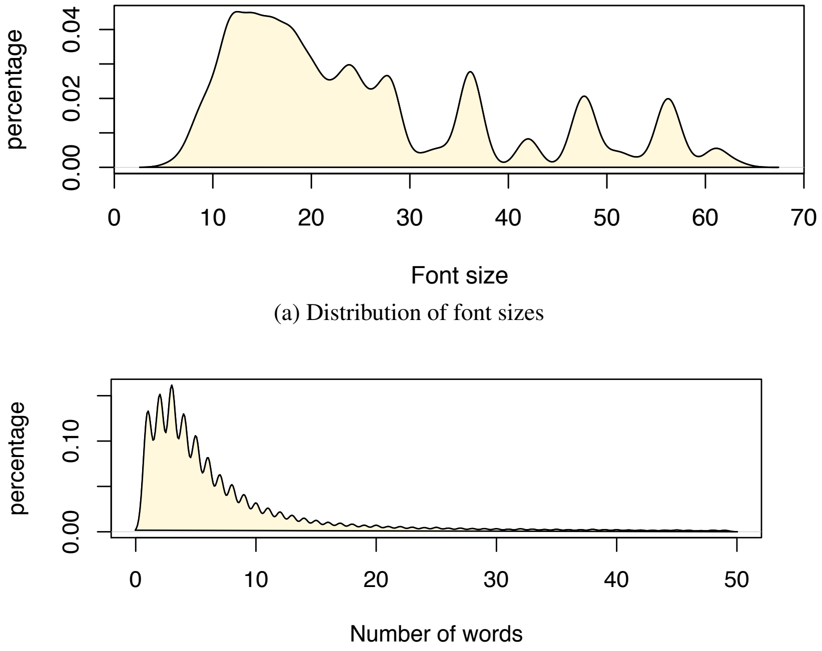
The median number of words over all text items in our test set is 5 and 30% of the text items contain up to 3 words. Thus, descriptions seem to be rather short in albums and only very briefly give additional information to the album contents. Text items with more words may designate a different kind of annotation, e.g. diary type text describing a specific day or place of a holiday but not solely describing a single photo. The distribution of text lengths is depicted in Figure 5.12b.

In this context it may also help to take the font size of the text into account. We found out, that the median text size is 22 and that 30% of the test items have a font size of at most 16. Thus, these seem to be typical sizes for general annotations in a photobook. A text with a font size significantly exceeding these values may be a candidate to be a more important annotation such as an album title or the title for an event in an album. The distribution of font sizes over all text items in our test set is depicted in Figure 5.12a.

5.5.4 Vividness

So far we have analyzed the degree of socialness and expressiveness of photobooks. Another interesting characteristic is the vividness: Photobooks can be designed rather factual and *cold* or more lifely, e.g. by the use of many colors or strong variations in the visual layout. To some degree this is strongly related to the expressiveness of photobooks. However, unlike in the last section, we will focus more on the overall visual impression of photobooks rather than on the photobooks' textual contents.

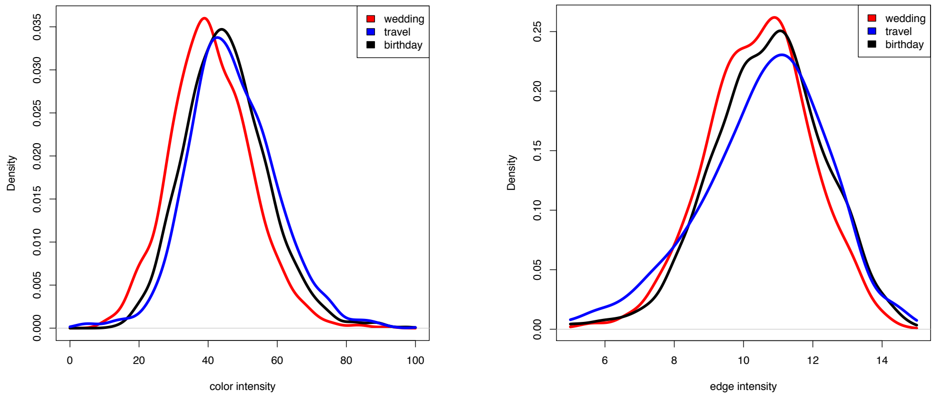
As indicators for the vividness we have chosen two features, the diversity of colors and the intensity of the photobook pages. We have derived these features for the same photobook classes as in the the last section.



(a) Distribution of font sizes

(b) Distribution the number of words per text item

Figure 5.12: Distribution of font sizes and text lengths over all photobooks



(a) Colorfulness

(b) Contrast

Figure 5.13: Degrees of colorfulness and number of edges for different classes of photobooks

Color Distribution

As way to rate the vividness of a photobook from a visual perspective is how diverse the colors of the photos throughout one page are. E.g. a photobook with sepia or gray photos or photos which often show the same scene would be perceived as much less vivid than photos with very diverse colors. We have analyzed all pages of our photobooks regarding this diversity in colors and have compared this to different classes of photobooks with the help of our photobook classifier. For the comparison of photo we have employed the Color Distribution Descriptor defined in the MPEG-7 standard. For every album we have determined the average pairwise similarity of photos on a page.

The result is depicted in Figure 5.13. Compared to travel and birthday photobooks, wedding photobooks show a much lower variation in their color layout throughout a page. This seems reasonable and backs up the impression from the manual, visual analysis of a number of wedding photobooks. These photobooks usually mainly incorporate very light and non-intensive colors and are limited to only a couple of colors. Thus, we can conclude that wedding albums are visually less vivid than other types of photobooks and create a rather calm, romantic and perhaps more intimate impression.

Intensity

As a second indicator for the vividness of photobooks we have identified the degree of smoothness or the intensity of the individual photobook pages. Thus we determined how many strong edges are present in the photos: A layout aiming at a more romantic style will often contain quite uniform areas with a strong use of depth-of-field smoothness. This effect is e.g. often used in professional portrait shots. Thus, we similarly analyzed our test set regarding the sharpness or contrast of contained photos and compared different classes of photobooks. For this we employed the MPEG-7 Edge-Histogram Descriptor and determined the percentage of non-edged pixels for every photobook page for different classes of photobooks.

The result is depicted in Figure 5.13b. Interestingly the degrees of edge intensity have a stronger variety in the travel photobooks compared to wedding or birthday photobooks. This can probably be explained with the variety of different events for travels. E.g. one could be on a winter skiing trip with a lot of smooth white areas in the photos or on a city trip where there are a lot of strong edges in the photos due to a lot of buildings. Also, again, the wedding photobooks shows significantly less strong edges than the other photobook types, which also backs up up our visual impression of wedding photobooks, which usually create a more intimate, less vivd and romantic impression is in line with the results of the color distribution analysis.

5.6 Summary

In this chapter we have presented the results of a large-scale analysis of real-world photobooks regarding the presence and degree of socialness, expressiveness and vividness.

For this we have distinguished between different types of photobooks and have developed a classifier to automatically decide about the type of a photobook. In conclusion we can say that photobooks are showing very strong social aspects and persons are very prominent in most photobooks. However, we have also observed significant differences for different types of photobooks. One example is that wedding photobooks aim to create a much more intimate and calm atmosphere than other types. Although we limited the categorization to only three important photobook types, our analyses have shown that photobooks a very interesting means to reveal more about peoples' incentives to design photobooks and how they express semantics and emotions for different types of events documented in these photobooks. We see the results in this chapter only as a first step. In the future we aim to identify additional photobook types and their special characteristics.

The main incentive for our analysis, besides to better understand peoples' behavior in photobook design, was to gain knowledge for being able to develop methods for automatic photobook design which take characteristics and implicit rules of real world photobooks into account. The main conclusion of our analysis is, that a meaningful layout system has to take into consideration the type of photobook which has to be created. E.g. different kinds of photos should be selected and a different layout should be chosen when designing a wedding photobook compared to a book documenting a journey.

6 Multimodal Semantic Photo Analysis

Prerequisite to be able to automatically select the right photos from a collection of photos and to be able to meaningfully place these photos in a photobook is a good semantic understanding of the photos. Looking at the related works in the field (See Section 3.2) we can conclude than neither pure content nor pure context analysis has been proven to be a success model. The approach we are following in this thesis is therefore a multimodal one where we combine context and content analysis.

6.1 Objectives and Approach

The methods for photo analysis developed in this thesis are aiming at solving different problems in the context of automatic photobook generation. In the following these different objectives are discussed:

Photo Retrieval

To be able to determine a reasonable selection of photos from one or more photo collections for a photo photobook, a good semantic knowledge about the photos is needed. This selection is influenced by various criteria, e.g. the quality or aesthetic of single photos but also the relation of photos and their semantics to each other. A more detailed discussion about these different factors is done in Chapter 7. The photo analysis methods developed in this thesis are partly targeting at providing measures or features which allow for deciding about the quality, relevance and importance of photo.

Content Enrichment

As elaborated in Chapter 2 a photo usually consists of content besides the user's photos. To be able to automatically decide if and what additional content could be added to a photobook, also a good understanding of the the relevant photos is needed, e.g. where a photo was shot, who is is shown on the photo or which kind of content could visually match specific photos.

Photobook Layout

To be able to automatically decide which photos should be placed on one page to best support the story represented in the photobook one has to be able to decide which photos do semantically belong together. Additionally, the visual character of photos, e.g. which colors are dominant in the image and are persons shown in the photo or significantly influences the visual character of a photobook. Also the decision of which parts of a photo can be trimmed or can be overlaid by other elements should be based on a good semantic understanding of the photos.

The goal of the system described in this chapter is to support these different requirements.

6.2 Photo Analysis Framework MetaXa

The goal of MetaXa[1] (Content-based and Context-driven metadata enhancement architecture) is to provide an architecture and infrastructure to automatically extract and enhance the metadata of digital photos and the initial metadata that comes with these photos. We aim at exploiting and enhancing the media content by using media analysis, multimodal retrieval techniques, and the integration of content-based and context-based metadata enhancement to create semantically rich personal media collections. Existing technology from content-based media analysis is advanced towards multimodal and context-driven enhancement of meta data extraction and deduction of higher-level semantic descriptions of the content. For this the architecture defines a plug-in infrastructure, an interface, and component definition as well as workflow specification for the flow of the extraction and enhancement steps.

6.2.1 Architecture

In this section, we briefly introduce the general architecture with its central elements and features which are illustrated in Figure 6.1.

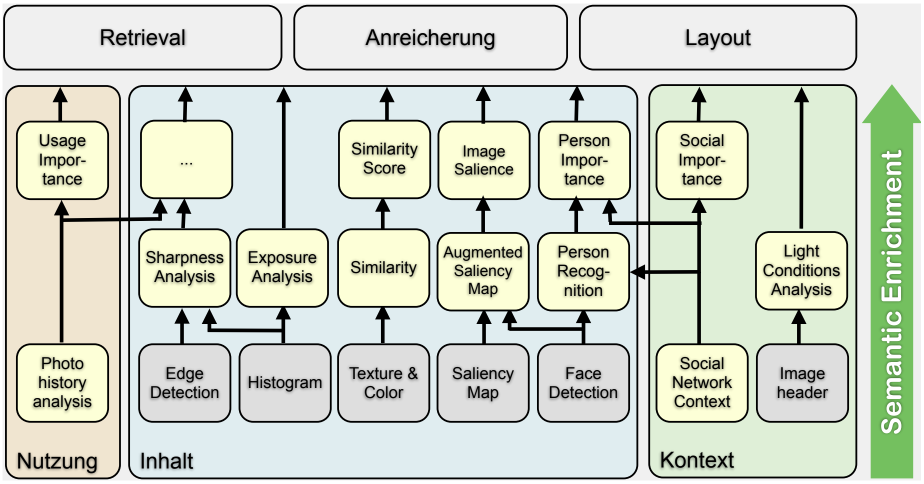


Figure 6.1: General metadata extraction and enhancement architecture

Input for the architecture are the images taken by a digital camera. These images, together with their contextual metadata that are available, e.g., within a so called EXIF header, enter the central *MetaXa Manager* for extraction and further enhancement of

[1] Metaxa is a Greek liqueur, a blend of brandy and wine, invented by Spyros Metaxa. In our system, Metaxa figuratively means an architecture that is a blend of content-based and context-based metadata enhancement.

the available metadata for the given image content. The image undergoes a sequence of extraction and enhancement steps in which iteratively the existing or derived metadata is input to the next enhancement step. This allows to modularize the metadata creation process into different steps. Increasing the amount of available metadata and also increasing the quality and confidence in the metadata each step contributes to a better semantic description of the personal photos. As illustrated in Figure 6.1, the raw image from a trip to Athens in Greece comes with its metadata to be analyzed both with regard to content-based but also context-driven features. Different extraction components can be employed to extract low level features such a color histogram or use edge detection to determine the degree of sharpness of a photo. Combination of features can be used to derive higher level metadata. For example, the analysis of pictures taken can be compared to an existing set of reference photos to get an understanding of the persons on the picture and to derive a semantic annotation for the photo. As we specifically aim at a combination of content-based and context-based metadata enhancement, there are also components that combine the signal analysis with existing context information to derive higher level metadata. For example, the analysis of an image and information about the exposure can be combined with information from the context such as the flash to understand if an image was taken indoors or outdoors. Also, components can be purely context-based: Time, date, location or a calendar can be used to derive the name of the place, the season and other relevant metadata from this.

The architecture provides an interface to plug-in new enhancement components. Each *extraction and enhancement component* which is plugged into the architecture reveals the input parameter it needs for the metadata enhancement and also which metadata is created or enhanced by the component. For example, the color histogram just needs the image and the data format and delivers the color histogram. An EXIF header extractor delivers all metadata that the EXIF header actually provides, given the image comes with such a header information. A component can also just increase or decrease a confidence value of metadata.

The different enhancement steps are driven by a workflow. This workflow configures the sequence of enhancement components that is carried out for each of the photos. This allows to configure the metadata extraction and enhancement for the actual application needs without having to change the system. The declarative workflow description identifies the different components and the metadata enhancement manager uses the workflow to drive the enhancement of the content.

6.2.2 Components in MetaXa

The different extraction and enhancement tasks are realized by software components. Software components encapsulate their implementation and interact with the environment by means of well-defined interfaces [SGM02]. Thus, a software component comes with a clear specification of what it requires and provides. As Java is used to realize the MetaXa architecture, which does not natively allow for such a detailed specification, the interface specification is implemented with Java Annotations. Java Annotations are a method to add metadata to Java source code which can be evaluated at run time. In MetaXa components two kinds of annotations are being used to specify the what kind of metadata a specific component expects (Pre-Condition) in order to function properly and which kind of metadata it provides (Post-Condition). The following listing shows an excerpt from such a file for the in-/outdoor enhancement component. This enhancement component requires the light status, time, brightness, and flash usage to determine whether the photo has been captured in-/outdoor.

```
@Precondition({
    @Datum(name = IMetadataEnrichment.LIGHTSTATUS),
    @Datum(name = IMetadataEnrichment.EXPOSURE_RATING),
    @Datum(name = IMetadataEnrichment.EXIF_HEADER)
})
@Postcondition(
  @Datum(name = IMetadataEnrichment.INDOOROUTDOOR)
)
public class InOutDoorEnrichment
  implements IMetadataEnrichment {...}
```

Listing 6.1: Interface description for a MetaXa component

With the help of these annotations it is possible to determine for a set of components, in which order photos are processed by these components. Furthermore it is possible to automatically parallelize certain components. Another benefit of this system is, that the same type metadata can potentially be provided by different methods. THese methods can be implemented in different components and easily be substituted.

6.3 Content-based Analysis

In the following the components of MetaXa are described which deal with the content of the photos.

Luminance Histogram

Histograms are useful as input for several analysis techniques. A histogram is the static distribution of color or brightness values in a picture and designates how often a color or brightness value occurs in the picture. In the histogram generation component four types of histograms are generated for each photo. First, separate histograms are built for each color channel in the RGB color space. Based on these histograms, we generate a brightness value histogram. According to [Ham92] the brightness of a pixel can be derived from its RGB values as follows:

$$Y = 0.299R + 0.587G + 0.114B \tag{6.1}$$

As seen, the colors are not treated equally. This is due to the fact that the human eye is for example more sensible for changes in the green channel than in the blue channel. Merging the three color histograms accordingly, we get a brightness histogram.

6.3.1 Edge histogram

Edge detection is used in computer vision to find regions of high spatial frequencies in an image. These regions visually correspond to edges. We use the sobel operator to detect this regions. This operator relies on the the mathematical operation convolution which is fundamental to many common image processing operators [Jäh06]. Let our input gray level image be represented by a matrix I_{mn} which consists of M rows and N columns. Then we convolute this image with *convolution kernel* K_{kl}. This kernel is typically much smaller than the input image. In the case of the sobel operator it has a size of 3×3. The actual convolution resulting in an output image O_{mn} is defined as follows:

$$O_{ij} = \sum_{k=1}^{m} \sum_{l=1}^{n} I_{i+k-1,j+l-1} K_{k,l} \tag{6.2}$$

The sobel operator is defined by to of these convolution kernels K_{kl}, namely:

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \tag{6.3}$$

By the use of these kernels two partial derivations of the input image are calculated, one for the horizontal and one for the vertical direction. These two pictures are used to calculate an overall edge direction histogram

6.3.2 Similarity analysis

Often a series of photos is taken from a motive to make sure that at least one photo is sharp or correctly exposed. We use a simple technique to detect similarities between certain photos and so to be able to detect such series. Therefore, we segment each photo into a 8×8-matrix and calculate the average RGB values for each of these fields. Assuming the RGB values each can consist of 256 different values similarity between two photos can be calculated with the according 8×8-matrizes p and q:

$$sim(p,q) = \frac{1}{8^2}\Bigg(0.299\sum_{i,j}\frac{|R(p_{ij}) - R(q_{ij})|}{256} \tag{6.4}$$

$$+0.587\sum_{i,j}\frac{|G(p_{ij}) - G(q_{ij})|}{256} \tag{6.5}$$

$$+0.114\sum_{i,j}\frac{|B(p_{ij}) - B(q_{ij})|}{256}\Bigg) \tag{6.6}$$

We weight the separate RGB channels with different shares, because the different colors are not sensed equally by the human eye.This results in a similarity value between 0 (very similar) and 1 (not similar at all). The size of the matrix has to be carefully chosen. If the segmentation is too coarse the differences between two pictures can not be reliably detected because too many details get lost by the calculation of averages. This can result in a high similarity value even if the two pictures are not considered as similar. However, if the segmentation is too fine grained, two pictures that e.g. only differ in a small horizontal or vertical shift, would possibly get a relatively small similarity value. From our experience we conclude that the best trade-off is a size of 8×8.

6.3.3 Saliency Map

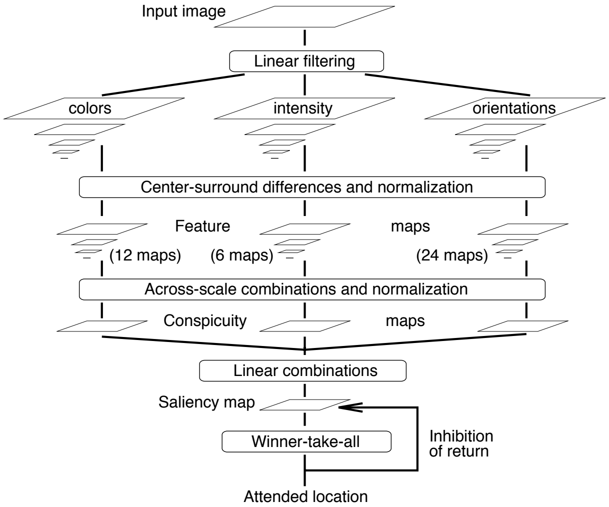


Figure 6.2: Original Itti & Koch Model for visual attention

The human visual system has a remarkable ability to interpret complex scenes in real time, despite the the humans brains inability to interpret all sensory information. There is a strong evidence that the human brain is able to distinguish between important and

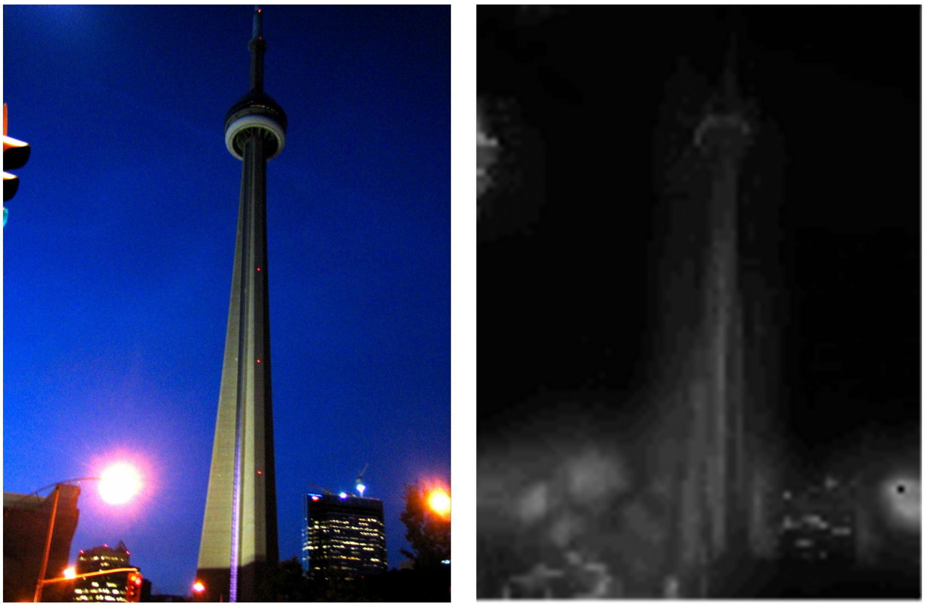


Figure 6.3: Original photograph and extracted saliency map

unimportant parts of a visual scene and tries to reduce the analysis complexity by focussing on the important parts [IKN98]. Studies have shown, that the human eye is only able to focus its attention to a limited area in the visual field and rapidly shifts this focus over the visual field. A couple of psychological works have tried to model this phenomena. Most models usually distinguish between a rapid, or bottom-up and a slower, task-driven top-down manner. Most systems for visual attention modeling follow are using a so-called saliency map for a visual scene where areas with high values designate areas of high visual attention in the corresponding location. Such saliency maps can thus also be a means to decide about interesting areas in a photograph.

In the following a method is described to extract such a saliency map from a photograph which closely follows the model developed by Itto & Koch [IKN98]. Unlike this model, in thesis the Lab color space is used in favor to the RGB color space. The Lab color space is known to better model the human visual cortex and thus is likely to lead to more realistic results than the RBG color space. In the following the model used in this thesis is briefly described.

A saliency map is built according to the method depicted in Figure 6.2. The basic idea behind this model is, that the human visual system is sensitive to various visual features which compete for attention. In [IKN98] three types of feature have been used, which are intensity, color, and orientation. These features are represented by different feature maps which are combined in a master saliency map. All features are computed by a set of linear *center-surround* operations: Visual neurons are most sensitive in a small region of the visual space, while stimuli in regions concentric to this center region (the surround) inhibit the neuronal response. Thus, objects which locally stand out from their surround are potentially areas of high saliency. Itti[IKN98] has modeled this effect by building the difference between coarse and fine scales of feature maps: For every input map a pyramid of 8 maps is constructed. Hereby every map has half the scale as its predecessor. The center-surround of a pixel is then defined as the difference between fine and coarse scales. Hereby the center is a pixel at scale $c \in \{2,3,4\}$ and the corresponding surround is the same pixel (relative to the scale) at scale $s = c + \delta$ whith $\delta \in 3,4$. The across-scale difference between such two maps is done by interpolating the coarser (c) to the finer scale (s) and building the point-by-point subtraction (denoted by $\ominus$): $A(c,s) = |A(c) \ominus A(s)|$.

The feature maps are built as follows:

Intensity: The intensity of a pixel basically represents, how bright it is perceived by the human eye. A set of 6 intensity maps and an across-scale combination is built as follows:

$$\mathcal{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(I(c,s)) \tag{6.7}$$

Color: The retina of the human eye consists of three types of color receptors (cones) which are bound to different bands of wavelenghts of light. This bands broadly correspond to red, green and blue. The human brain does not primarily interpret responses on these receptors separately but either detects differences in responses of these different types of receptors at the same place to perceive colors. These differences are reflected in the Lab color space by the a and b components. Thus, the Lab color space is much better complies to the human visual system than the RGB color space which was originally used by Itti & Koch. The color feature is thus reflected by 6 feature maps for the a color channel and 6 feature maps for the b color channel and an across-scale combination as follows:

$$\mathcal{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} (\mathcal{N}(a(c,s) + b(c,s)), \mathcal{C}_b = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} (\mathcal{N}(b(c,s)) \tag{6.8}$$

Orientation: Another cue for objects to stand out from their surrounding is a difference in their direction. To detect the orientation of of an image a series of gabor filters is applied to the image. Is of these filters broadly responses to one specific direction. The corresponding features maps and an across-scale combination is defined as follows:

$$\mathcal{O} = \sum_{\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}} \mathcal{N}(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} (\mathcal{N}(O(c,s,\theta))) \tag{6.9}$$

$\mathcal{N}(...)$ in the above formulas refers to a normalization operation which is done at each step. These three different feature maps are combined and again normalized to a final saliency map. An example of such a map and the corresponding image is depicted in Figure 6.3.

6.3.4 Face detection

For this component, we rely on a face detection method implemented in the OpenCV imaging library [Int, VJ01]. This method uses previously trained classifiers to rapidly detect object, in our case faces, in a picture. The classifiers were trained with several hundreds positive samples, namely faces, of the same size and also a few hundred negative examples, that means arbitrary images of the same size. The result was a *a cascade of boosted classifiers*. Cascade means, the resultant classifier consists of several simpler classifiers (stages) that are applied subsequently to a region of interest until at some stage the candidate is rejected or all the stages are passed. At each stage more classifiers are applied to the region of interest. This results in a very fast algorithm as regions that do not have the shape of a face are considered as not being a face at a very early stage, where only a small number of simple and therefore fast classifiers have been used.

6.4 Context-based Analysis

6.4.1 Photo header Metadata Extraction

Most consumer camera write an EXIF header to the photos they store. It is estimated that 60 percent of the photos of digital cameras do have an EXIF header, and this percentage increases constantly. The contents of this header vary from camera brand to camera brand but most cameras at least store information like a timestamp of the date and when the photo was taken, information about used ISO, aperture and exposure time settings, and if a flash was fired or not. In addition to that some cameras provide also information about the used focal length, the orientation of the camera or even location information if the camera is equipped with a GPS receiver. All of this information is extracted by an EXIF reader component.

6.4.2 Social Network Derivation

One outcome from the analysis of photobooks in Chapter 5 was, that persons are one of the main motifs shown. Thus, it is reasonable to also let the mere presence of a person, but also the relationships between different persons shown in a photo set influence their presence in an automatically designed photobook.

In the following a method is described to derive the social network of several persons from a set of photos where these persons are shown. The assumptions we stem from are the following:

Appearance in same photo

If two persons have a strong relationship, they **appear often together** in photos. Let n_i be the number of photos which show person i, n_j the number of photos showing person n_j and n_{ij} the number of photos showing both persons together. A function modeling the relationship between persons i and j based on this assumption can then be modeled as follows: $f_1(i,j) = \frac{n_{ij}}{n_i+n_j}$. This leads to a value between 0 and 1, 0 meaning the persons do no appear together in one single image and 1 meaning they only appear together in images. The function is only defined, if at least one person is present in at least photo.

Spatial distance

If two persons have a strong relationship, they often stand together on photos. The shorter this distance, the closer the relationship of these two persons is. This assumption is modeled as follows:

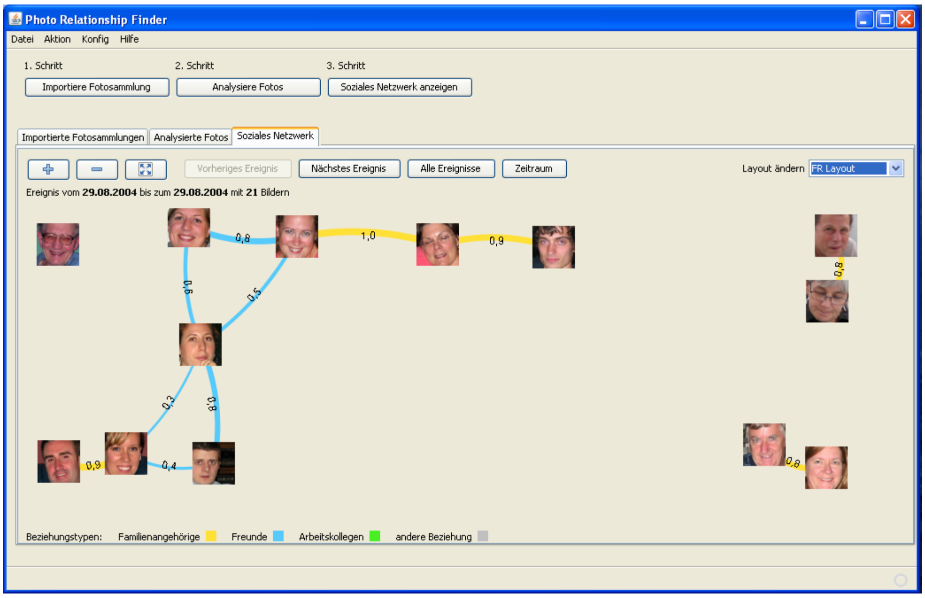
$$f_2(i,j) = \frac{1}{\frac{1}{n_{ij}}\sum_{k=0}^{N}(dist_{i,j}(k)+1} \quad ,$$

Again, n_{ij} ist the number of photos where both person i and j appear in an N is the overall number of photos in the set. $dist_{i,j}(k)$ models the distance from person i to person j in photo k. It takes into account the size of the detected faces and their distance from each other and from this provides a distance rating which is a logarithmic function dependent on the determined spatial distance. A distance of 0 cm leads to a rating of 1 and a distance of 1 m to a rating of 0.5.

Overall number of people

As a third function we take into account the overall number of people in a photo. The central assumption behind this is, that the more are shown in a photo the less strong the relationship between each two people in the photo is. Imagine, e.g. a group of 20 people where people are shown very close together to able to into the scene and photo of a couple. According to function f_2 both photos would lead to the same value based their distance in the photo, but one would assume that the couple in the second photo has a much stronger relationship than two people standing together in a group shot, which could barely know each other. This third assumption can be modeled in the following way: $f_3(i,j) = \frac{1}{\frac{1}{n_{ij}}\sum_{k=0}^{N}(numPers(k))+1}$ $numPers(k)$ designated the number of persons shown in photo k. This function leads to high values if only few persons are shown in photos with both person i and j and to low values if many persons are shown.

To derive a social network graph with the help of these three functions is a matter of combining the functions in a meaningful way. The overall social distance of two persons i and j can be calculated by the function $SocDist(i,j) = w_1 f_1(i,j) + w_2 f_2(i,j) + w_3 f_3(i,j)$. The values for the weightings have been determined on the basis of photo set of a wedding party showing 9 Persons. Table 6.1 shows an overview of the number of appearances of each person in the set. The social distances for each two persons have been assigned by a person knowing all persons in the photos and having attended the wedding party. As the optimal weighting factors to best match these assignments, the following values have been determined: $(w_1, w_2, w_3) = (1.1, 0.9, 2.2)$.



Ereignis vom **29.08.2004** bis zum **29.08.2004** mit **21** Bildern

Figure 6.4: Visualization of a social network derived from a set of photos.

	Ha	EH	He	UH	JF	WF	SF	NF	HW	VW	IN
count	22	23	5	7	9	4	6	9	5	5	4

Table 6.1: Number of appearances of persons in the test data set for determining the parameters for the social distance function.

6.5 Higher Level Semantics Derivation

In the following higher level components are described which build upon the semantics of the before described components.

6.5.1 Exposure Analysis

One indicator to decide about the quality of a photo in most consumer photos is to look at the exposure. The central parameters to control the exposure in a photo camera are the lens aperture and shutter speed. These two parameters have to be balanced in order achieve, speaking in technical terms, a *correctly exposed* photo. If a photo os overexposed, important bright parts of the photo are *washed out* or completely white, an an effect which is also referred to *clipped whites*. Underexposure on the other side is visible in a loss of detail in the dark parts of an image (*clipped blacks*). Sometimes over- or underexposure is intentionally chosen by the (advanced) photographer to achieve a specific visual impression. In most cases, especially for conventional consumer shots, it is a matter of operating a photo camera in a wrong way or the inability of a camera to automatically choose the right parameters. A common is, e.g. that the built-in flash of a camera is not able to light to sufficiently light the scene.

The exposure of a photo can be derived from its luminance histogram. An overexposed photo leads to high values in the higher bins of the histogram whereas an underexposed photo exhibits high values in the lower bins. A correctly exposed photo is usually characterized by an even or gaussian-like distribution in the luminance histogram. To be able to decide about the exposure of a photo we use the following method:

- We transform the luminance histogram of the photo into one with 3 and one with 8 bins (lum_3 and lum_{32})
- Initialize a rating variable with $R_{exp} = 0$.
- If $\sum_{i=0}^{3} lum_{32}[i] < 0.01$ then $R_{exp}+ = 0.22$
- If $\sum_{i=28}^{31} lum_{32}[i] > 0.25$ then $R_{exp}+ = 0.32$
- If $lum_3[0] < 0.125$ and $lum_3[2] > 0.60$ then $R_{exp}+ = 0.46$
- If $\sum_{i=28}^{31} lum_{32}[i] < 0.01$ then $R_{exp}- = 0.22$
- If $\sum_{i=0}^{3} lum_{32}[i] > 0.25$ then $R_{exp}- = 0.32$
- If $lum_3[0] > 0.6$ and $lum_3[2] < 0.125$ then $R_{exp}- = 0.46$

This simple analysis of a luminance histogram leads to rating $1- >= R_{exp} <= 1$ where -1 designates a strong underexposure and 1 a strong. overexposure.

6.5.2 In- and Outdoor Derivation

We have developed a simple yet powerful method to detect, if a photo was taken indoors or outdoor. Therefore, we rely on previously extracted metadata namely the light condition information, daytime, information, if a flash was fired or not, and the rating for the exposure. The first three metadata entries are generated from context information and the last from the image content.

daytime and dark picture	2
very bright picture	-2
flash fired and dark picture	-1
flash fired, dark picture and night time	1
flash fired and night time	1

Table 6.2: In-/Outdoor detection

For each photo each of the *rules* shown on the left in Table 6.2 is evaluated. If a photo fulfills a certain rule, the value shown on the right is added to an overall classification value. If the resulting value is negative, the photo is considered as being taken outdoors.

If the value is positive it is considered as being taken inside. If it is zero, no assumption is made whether it was taken inside or outside.

A similar method has been proposed in [BL05], where Support Vector Machines (SVMs) are used to classify a photo as taken indoors our outdoors. Unlike our approach only context information is used to classify the photos. In contrast, we aim at combining context and content information to achieve an even higher precision.

We have evaluated our approach with a typical photo set from a four week vacation trip consisting 437 photos from which 68 are indoor and 369 are outdoor shots. The test set also consists some ambiguous photos like photos made indoor out of a window. The component misclassified 0.5% of the outdoor shots as indoor and 10.1% of the indoor shots as outdoor. 10.1% of all pictures were classified as ambiguous. We are are aiming at improving the classification accuracy. One way would be to take object distance metadatum into account which is stored by most cameras in the EXIF header. An object that is very far away from the camera (e.g. over 30 m) can be considered as being outdoor.

6.5.3 Lightstatus Enrichment

Information about the light conditions when a photo was taken are useful both for search purposes and as information for further analysis of the photo. With light conditions we mean if the scene where and when the photo was taken was well illuminated or not. One possibility would be to measure the light explicitly with a dedicated sensor on the camera. Most cameras do this to be able to find the right values for setting aperture and exposure time but do not store this information in the EXIF header. This information can alternatively be derived from the photographic information aperture and exposure. In [ISSE02] a method is described to calculate an exposure value from the given values in the EXIF header for aperture ($FNumber$) and exposure time ($ETime$):

$$\begin{aligned} E_v &= A_v + T_v \\ &= 2 * \log_2(FNumber) + (-1 * \log_2(ETime)) \end{aligned}$$

The value E_v should be proportional to the brightness in the scene and therefore is a good indicator for the light conditions.

6.5.4 Importance Map

To be able to determine a meaningful layout of photos in a photobook it is important to, know parts of a photo are perceived as more and which are perceived as less important by humans. A good starting point is a saliency map as described before. However, these saliency maps only consider the general case of visual scenes for digital photos, and especially for consumer photos often shot by lay users, which are present in digital photobooks, we have made two important observations:

- The main object is most of the time placed in the middle of the photo making this part more important to the user than the rest of the photo.
- Person are the most important motif in photos. Thus, emphasis should be put on persons in a photo.

These two observations have lead us to extend the general saliency by Itti&Koch in the following way: To model the first observation we determine a map of the same size as the saliency map following a gaussian distribution with the origin placed in the center of the map. By this, the most important region is shown in the middle and the importance is decreasing towards the sides and corners of the input image. Additionally we determine a map where regions corresponding to a face detected via the face detection component, are showing the maxim value (256) of the map and regions without a face the minimum value (0). The result are three maps which are combined into a combined importance map. Mathmatically, the importance map can be described as follows:

$$Imp(x,y) = w_{face} M_{face}(x,y) + w_{sal} M_{sal}(x,y) + w_{gauss} 255 \exp^{-(\frac{(x-\frac{X}{2})^2+(y-\frac{Y}{2})^2}{2s^2}}$$

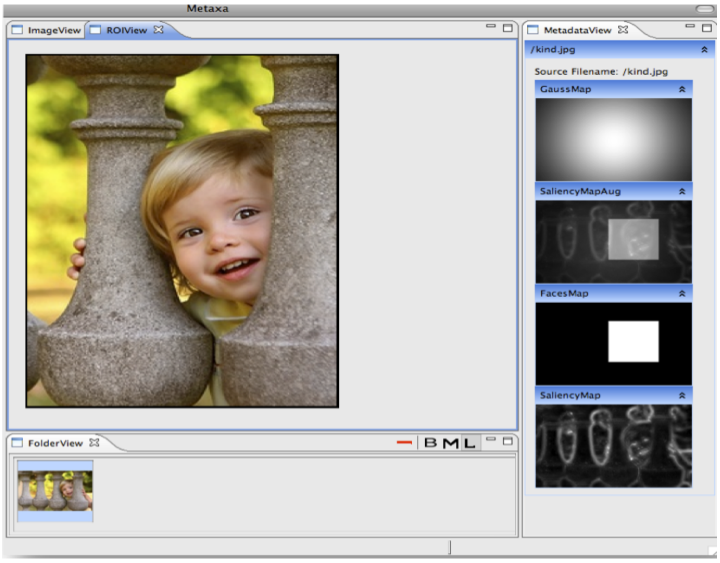
This function can be adjusted by the three weights for each each map with $1 = w_{face} + w_{sal} + w_{gauss}$ and the *steepness* of the gauss map adjusted by parameter s. A higher value for s gives more emphasis for the inner part of the photo. We have empirically determined that the values $w_{face} = 0.3, w_{sal} = 0.5, w_{gauss} = 0.2, s = 130$ are a good tradeoffs for most applications and photos.

An example of such an importance map in depicted in Figure 6.5. It is screenshot of the MeTaxa workbench visualizing the different maps. Based on the combined importance map the input is automatically cropped by selecting the part of the image which corresponds to maximum mean value of pixels in the corresponding importance map. On the right visualizations for the different maps are shown. From top to bottom these are the Gaussian map, the combined importance map, the face map and the saliency map.

6.6 Summary

In this chapter we have described a novel software architecture for the component-based analysis of digital photos. This architecture is in first place meant as one component in the photobook creation process and nevertheless easily be used as a general architecture for other photo analysis applications. The key idea behind this architecture is to intelligently divide the analysis process into smaller components which can easily be rearranged and substituted by other components. For this, every component provides which kind of metadata it expects and which kind of metadata it provides as output for other components or applications.

Additionally, a couple of novel methods have been described in this chapter which are important for the automatic generation of digital photobooks. AMong these are methods



Source Filename: /kind.jpg

Figure 6.5: Determining an importance by extending the Itti-Koch saliency model with a gaussian map and a face map. The image is automatically cropped based on the combined importance map.

for In- and Outdoor detection based on multimodal metadata, Importance Map generation as an extension of the Itti-Koch saliency model, and social network derivation.

The MetaXa photo analysis architecture is the basis for the automatic retrieval and also the layout system which are described in the following chapters.

7 Retrieval

Selecting a meaningful subset of photos from a large photo collection is an important step when designing a digital photobook. A typical scenario is e.g. the selection of the *best* photos from a holiday trip, where several thousands of photos have been shot. Usually, the user does not want to use all of these photos for a photobook, but rather a selected and meaningful subset, which best represents the underlying event and is visually pleasing. A large amount of photobooks (See Chapter 5) covers photos from a period of about 1 year. Typically, all photos shot by a person in this year are way too much for a typical photobook and thus a selection of the *best* photos of the is needed. From the analysis of real world photobooks we can see that users tend to use mechanisms for the automatic selection of photos. This both means that photos *are* selected and that there is a need for automating this process.

But, what is a good selection of photos and more specifically what is it for a photobook? Several works have been published for determining the aesthetic value of photographs (see Section 3.2.3). However, only few works actually approached the problem of sub-set selection in general and the selection for photobooks in particular. Recently Yeh et. al. [YHBO10] have proposed a system for personalized photo ranking and selection. However, as with most approaches, this work also only considers the content of photos to determine their importance for the user. As it very early became clear that “One way to resolve the semantic gap comes from sources outside the image ...” [SWS$^+$00] surprisingly no work so far has considered to also take these sources outside the image into account for photo selection. One result of the analysis of photo usage in Chapter 4 is that for a meaningful selection of photos one has to not only consider the single photo, but set this photo in context to other photos in the set. Also this aspect has so far seldom been considered.

Following this considerations in this chapter we propose a clustering and selection framework for digital photos. In contrast to other works in the field, it follows a hybrid approach to consider both the content and the context of photos and follows a two-step selection process by both considering the single photos, but also the importance of the event this photo belongs to.

The approach is driven by the following assumptions for a good selection of photos for a photobook:

1. All photos should be aesthetically pleasing and show a good quality.
2. The underlying event or events should be covered as good as possible. This means that every sub-event should at least be represented by one photo, even if this photo or other photos of the same event are of comparably low quality.
3. Near-Duplicate photos should be avoided. Often, an important scene is photographed more often than once to be sure that at least one good picture is taken. Usually, not all pictures should then be taken for a good selection, but rather the *best* out of this series.

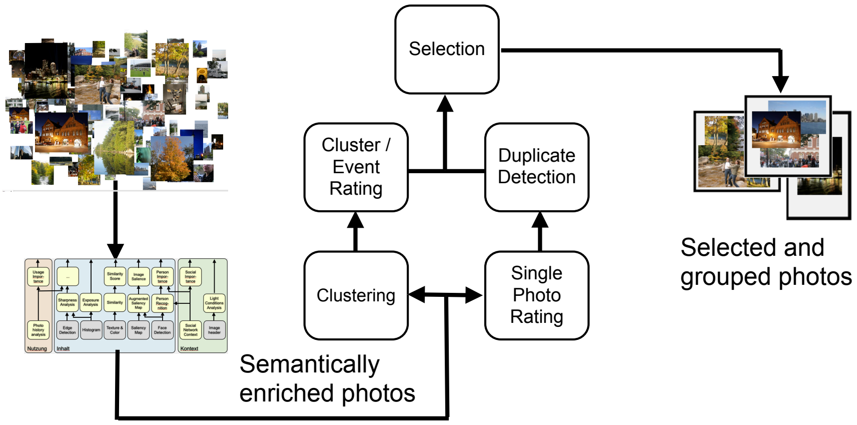


Figure 7.1: Architecture for photo selection

4. From the analysis in Section 5 we can conclude, that the presence of people is very important for a good selection for a photobook. Thus, photos with people should be favored over photos not showing people.

These assumptions were compiled by interviewing several people on their photo selection behavior as well reviewing the results of the study on human image appeal by Savakis et al. [SEL00]. The authors have performed a ground truth experiment on 11 persons to rank photos according to their relative appeal. One of the main outcomes was, that image appeal does not directly coincide with image quality, but is rather influenced by the presence of people in the photos and the relation between photos regarding support to represent the underlying event.

An overview of the proposed system is shown in Figure 7.1.The underlying principle is the distinction between the rating of events or groups of photos and the rating of single photos. This meets both assumptions (1) and (2). The ratings for the single photos and the corresponding cluster are then combined to derive an overall score for every photo. In an intermediate step near-duplicate photos are detected and rated accordingly. The result of a selection process is a subset of of the input photo set grouped into separate events. In the following the details of the proposed selection system are described.

7.1 Event Rating

One outcome of [SEL00] was, that image appeal is not only based on a single photo but also on the context it is bound to. Thus, the photo itself is considered for a selection, but rather the underlying event. This hypothesis is also backed up by the results presented in Section 4.3.1 where a correlation between the rating of photos and corresponding view durations in a slideshow of photos of the same event could be found. Thus, as one part of the overall photo rating the proposed system considers the importance of the underlying event.

To be able to perform such an event rating on a photo set, two task have to be done: (1) The identification of events and (2) the rating of events. For the identification of events in a photo set, several algorithms have been proposed (see Section 3.4.1 for an overview). The main means to distinguish between different events in a collection are time and place [NSPGM04, COT06]. However, only a very small portion of photos today is equipped with a location tag. Thus, for the determination of events in a photos the proposed system relies on a combination of time information and content-based similarity, developed in the PhotoTOC system [PCF02]. For this, the photos are first ordered according to their tim stamps present in the Exif header (it is assumed that every photo contains a valid Exif header). The problem of finding clusters is broken down to the problem of detecting noticeable gaps in the photo creation times. These gaps are assumed to correspond to a change of an event. The proposed system follows the same adaptive approach of Platt where a gap is considered as an event break when its length is much longer than the local gap average. Because gaps follow a very large dynamic range, e.g. can range from seconds to months, the detection algorithm works on logarithmically transformed gap times. It is assumed, that for every photo p from a temporally sorted set $P = \{p_1, \dots, p_n\}$ of photos a time stamp $t(p)$ can be be extracted. The gap between two photos is defined as $g_i = t(p_{i+1} - t(p_i))$. A gap defines a cluster break if the following condition is met:

$$log g_i \geq K + \frac{1}{2d+1} \sum_{j=-d}^{d} log(g_{i+j}) \tag{7.1}$$

K is a suitable threshold (empirically chosen to $log(17)$) and d defines the window size (chosen to be $d = 10$).

Having determined a set photo clusters from an input photo set, the clusters are individually rated. For this, again, several assumptions are made which influence the importance of a specific event. In the following it is assumed, that a given set of photos P is clustered into a set of photo clusters $\mathcal{C} = \{c_1, \dots, c_n\}$ where $P = c_1 \cup \ldots \cup c_n$. For every photo $p_i \in P$ a unique cluster $C(p_i) \in \mathcal{C}$ exists.

Nr. of photos

If one event comprises of comparably many photos it is assumed that this event is more important than other events with fewer photos. As a measure for this the we define

$$R_{c1}(p) = \frac{\#C(p)}{max(\#c_1, \dots, \#c_n)} \tag{7.2}$$

Cluster Distance

If an event is temporally isolated compared to other events, it is assumed that this specific event is more important than events which are close together. In our system this is modeled as follows: For every cluster $c_i \in \mathcal{C}$ a gap g_i is defined which is the temporal distance of its first photo to the last photo of the preceding cluster. Then a rating for a photo $p \in c_i$ in this cluster is defined by the following formula:

$$R_{c2}(p) = \frac{g_i + g_{i+1}}{2 * max(g_1, \dots, g_n)} \tag{7.3}$$

Intra-cluster similarity

If photos are very similar throughout an event, we assume that one specific motive was shot more than once and thus was especially important to the photographer. As a rating for the similarity inside a cluster the average, pair-wise similarity of all photos in a cluster is taken. As a similarity measure the MPEG-7 Color Descriptor is chosen. The scores (similarity measures) are normalized to $0 \dots 1$ where 0 designates the cluster with the lowest similarity and 1 with the highest. The rating is defined as R_{c3}.

7.2 Single Photo Rating

The perceived quality of a photo depends on several factors and the scientific community has developed several methods for deciding about this (see Section 3.2.3). All approaches have in common, that they usually rely on a combination of different ratings for different aspects to compile an overall score. The approach presented in the following follows this same basic approach, but differs in the choice of different aspects of features. The driving paradigm is the same as presented in some of the newer research outcomes [DLW07, YHBO10, CI10] which rate a photo according to the conformance to common photographic rules and aesthetic principles. However, these works solely concentrate analyzing the content of photos. Additionally, the proposed system takes also the context of photos and even the combination of content and context into account. In the following, the different aspects are described in more detail.

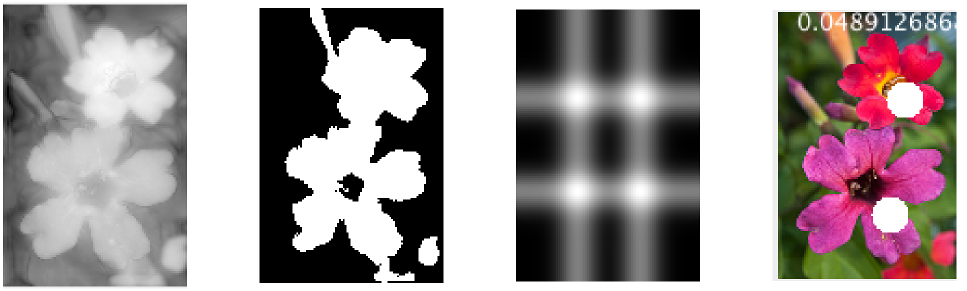
Rule of Thirds

The rule of thirds is one the most well-known photographic composition principles [Kra05, GS90]. The main idea is, that the main object(s) should not be placed in the middle of a photograph, but roughly at one of the intersections of imaginary lines dividing the photograph equally into 3 horizontal and 3 vertical parts.

In the selection system, two methods are used to rate the degree of conformance to this rule. The first method employs a importance map (See Section 6) to detect the most important points of the image (Figure 7.2a). This map is then binarized with an adaptive threshold filter to eliminate noise, and small and less important parts of the image. This binarized image is then rated according to the following formula:

$$R_{s1}(p) = \sum_{s \in bin(sal(p))} s wmap(s_x, s_y) \tag{7.4}$$

s designates the value of a point in the binary map which is either 1 or 0 and s_x, s_y the coordinates of this point. $wmap$ is a weight map which basically rates a pixel higher



(a) Importance Map

(b) Binary Map

(c) Weight Map

(d) Detected Attention Points

Figure 7.2: Rule of thirds Rating

the closer it lies to one of the key points conforming to the rule of thirds. The map is depicted in Figure 7.2c. It is built as follows:

$$wmap(x,y) = e^{-\frac{(y-\frac{Y}{3})^2}{x^2+y^2}} + e^{-\frac{(x-\frac{X}{3})^2}{x^2+y^2}} + e^{-\frac{(y-\frac{2Y}{3})^2}{x^2+y^2}} + e^{-\frac{(x-\frac{2X}{3})^2}{x^2+y^2}} \tag{7.5}$$

X and Y designate the height and width of the input photo.

Color Harmony

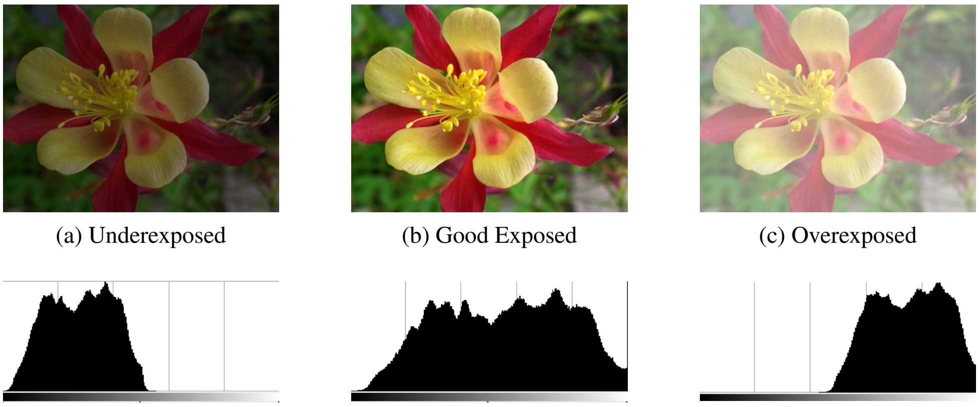
Besides the overall composition of photos regrading the position of dominant objects, another important factor to rate a photo is the color composition of a photo. Certain color combinations are perceived by the human eye as more pleasing than others and thus humans generally prefer harmonic color combinations. The harmony of colors is not decided by specific colors but rather by their relative position in color space [COSG$^+$06]. Professionals generally intuitively chose harmonic combinations of colors, but looking closely at pleasant combinations of colors, several rules can be derived. Itten [Itt97] has proposed several rules or color schemas which are considered as harmonic to the human eye (See Figure 9.2). These color schemas can also be used to decide about the color harmony of an image.

Given a harmonic scheme (m, α) a function $F(p,(m,\alpha))$ is defined which measures the color harmony of a picture p with respect to this scheme [COSG$^+$06]:

$$F(p,(m,\alpha)) = \sum_{x \in p} ||H(x) - E_{T_m(\alpha)}(x)||\dot{S}(x)\dot{S}al(x), \tag{7.6}$$

where H and S denote the hue and saturation. The original formula from Cohen is extend with an additional factor $Sal(x)$ which corresponds to the saliency of the respective pixel (See Section 6.5.4) The distance $||\ldots||$ corresponds to arc-length distance in radians on the hue color wheel to the closest border of a sector in the corresponding template.

Intensity



(a) Underexposed (b) Good Exposed (c) Overexposed

Figure 7.3: Different levels of exposure and corresponding intensity histograms

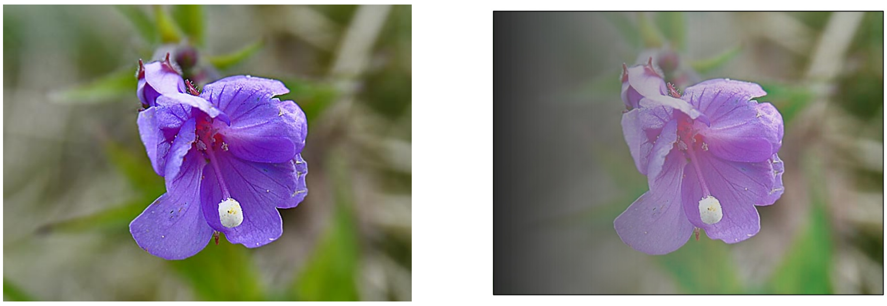
A common problem in typical photos shot by amateurs is over- and underexposement, thus photos are often much too light or much too dark. The intensity of a pixel is, e.g. encoded by the Hue Channel in the HSV color space. Sometimes it is wanted that certain parts of a photo are very dark or very light, e.g. a wedding dress of photo showing a wedding couple. Such photos are not perceived as under- or overexposed. However, calculating the average pixel intensity of such a photo would result in a relatively high or low value respectively. A good way to decide about a pleasant exposure is to look at the intensity histogram which should ideally show gaussian like distribution over the whole spectrum. Examples and corresponding intensity histograms are shown in Figure 7.3. Thus to rate to the intensity of an image, the distance of its intensity histogram to an ideal Gaussian distribution can be taken:

$$R_{s3}(p) = \frac{1}{\sqrt{\sum_{i=1}^{k}(hist_i(p) - \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{i}{k})^2})^2}} \tag{7.7}$$

Intensity Balance

Besides an adequate overall exposure of an image, a good photo creates an impression of balance, which is also a fundamental principle in photographic composition. Thus, ideally all objects in a photo should be equally distributed over the photo and the photo should evenly be exposed throughout the photo. Figure 7.4 shows examples of a balanced and an unbalanced image. This can be modeled by comparing different areas of an image regarding their intensity histograms. To simplify this task, we only compare the right and the left part of an image:

$$R_{s4}(p) = \frac{1}{\sqrt{\sum_{i=1}^{k}(hist_i(p_{left}) - hist_i(p_{right}))^2}} \tag{7.8}$$



(a) Balanced

(b) Unbalanced

Figure 7.4: Examples for balanced and unbalanced intensity within an image

Sharpness

Unsharp pictures are a common problem in consumer photo shots, e.g. caused by shaking the camera having chosen a long exposure time or due to a malfunction in the autofocus of the camera. In most cases, a photo is perceived as sharp when a large amount of sharp edges is present in the photo. Thus, a rating for the sharpness of a photo can be done by counting the number of detected edges from its sharpness histogram (See Section 6.3.1):

$$R_{s5}(p) = 1 - sharphist_{nonedge}(p) \tag{7.9}$$

Presence of people

According the analysis of existing photobooks in Chapter 5, people are present on more than 70% of all photos on average in a photobook. Thus, the presence of people is an important way to select photos. Rather than solely taking the number of people as an indicator, the face-covered area is taken as a feature. This avoids that photos are favored, where many people are shown, but rather small or in the background and thus are not the main object in an image:

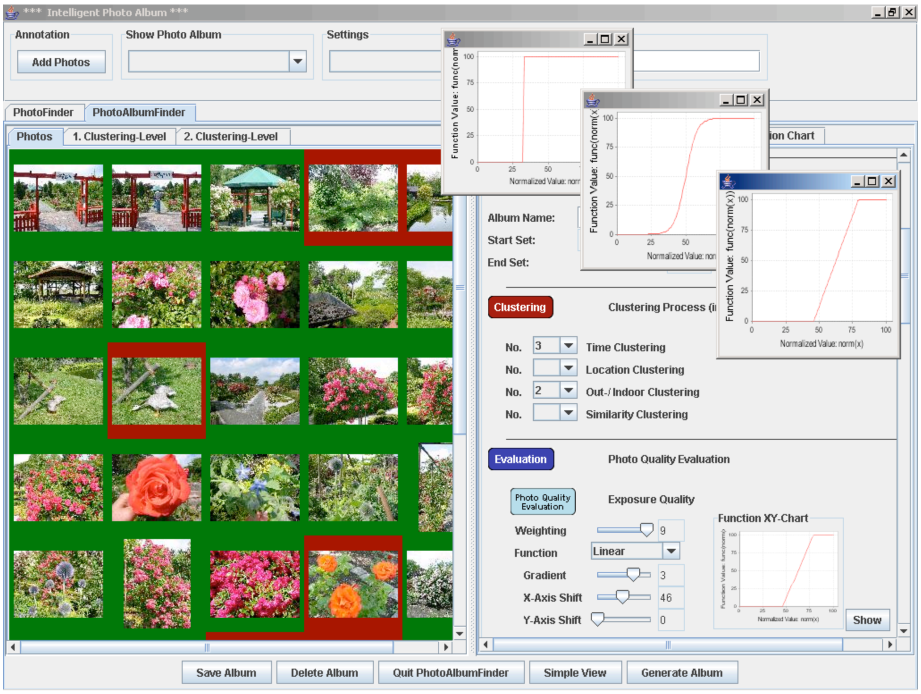
$$R_{s6}(p) = \frac{FaceCoveredArea(p)}{Area(p)} \tag{7.10}$$

7.3 Combination and Selection

As depicted in Figure 7.5 the actual photo selection is done in several steps. First an input photo set is clustered according to the clustering algorithm described above. The individual clusters are then rated according to the different cluster features and the additionally single photos are rated separately according to the different single rating features. An overall rating for a photo is then determined as follows:

$$R(p) = \sum_{i=1}^{3} wc_i R_{ci}(p) + \sum_{i=1}^{6} ws_i R_{si}(p) \tag{7.11}$$

This overall rating function is parametrized by the 9 weighting factors $wc_1, \ldots, ws_3$ and $ws_1, \ldots, ws_5$. As an additional step, the photo clusters are scanned for near-duplicate photos. These are clusters where the intra-cluster similarity exceeds a certain degree. From these clusters, the weighting of the best photo, according to the single photo rating, is increased, while the rating of other photos in the cluster is decreased.



No. 3 Time Clustering
No. Location Clustering
No. 2 Out-/Indoor Clustering
No. Similarity Clustering

Weighting 9
Function Linear
Gradient 3
X-Axis Shift 46
Y-Axis Shift 0

Figure 7.5: Screenshot of retrieval workbench to manually tune selection parameters

In Figure 7.5 a screenshot of a workbench is shown to manually adjust the parameters for the retrieval system. In our experience, the choice of suitable parameters is highly dependent on the kind of photos and the application of the retrieval system. E.g. for the application of automatic photobook creation one has to take into account the number of desired pages, the size of the individual pages and the maximum number of desired photos per page. Thus, the we did not aim to find an *optimal* set of parameters for the selection process but rather to provide a flexible toolset which takes into account the most important aspects which drive the human selection process.

7.4 Summary

In this Chapter we have provided a novel system for the automatic retrieval of photos from a larger set of photos which is flexible enough to be used in various retrieval tasks. The system both provides means to select the most important photos from an input set as well clustering these photos in a meaningful way. Its primary application is the use in the automatic photobook creation process.

The novelty of the proposed retrieval lies both in the explicit consideration of multimodal metadata driving the selection and introducing and explicit cluster rating step into the selection process which is motivated by modeling the human photo taking behavior which is usually driven by events that lead to photos taken in bursts.

8 Augmentation

With the increasing availability of online communities and the trend to collaboratively tag and organize shared multimedia content, a great potential lies in utilizing this ever growing pool of knowledge and media to augment personal photobooks and add collaborative content and knowledge to the personal media collection. Imagine, e.g., having spent a nice holiday in Paris but having missed the chance to take a decent picture of the Eiffel tower. This missing photo might be searched on a photo community site like Flickr and added to the personal media set. For photobooks, a day in Rome could be augmented by a description of the Colosseum from a Web 2.0 travel blog. The Internet is an ideal pool to search and find collaboratively created content to augment the personal media collection.

The sources for the photobook contents are not only solely the users' own hard disks any more. The results of the analysis of real-world photobooks in Chapter 5 shows, that most photobooks already contain photos from more than one camera. With the growing success of photo sharing platforms and photo sharing in social networks, photobooks are more and more getting a representation of the users' social interactions: Important photos not only sit in the users' personal accounts, but are are spread widely over their social network [RSB10]. This social aspect is also present in the users' photobooks: Typical events seem to incorporate people to a quite large degree and photobooks are often a reflection of the authors' social life.

One of the key features of the Web 2.0 is the fact, that it contains a vast amount of user generated content and often interfaces on standardized web technologies (REST, JSON, RSS, ...) to easily access this content from applications. Users can also shape the content in the Web 2.0 world themselves: They can add, delete, comment, organize and alter content. A good example for this are photo communities: One can add or delete a photo, tag it in various ways add it to groups or comment it. Several of the provided Web 2.0 services are of interest for the creation of personal photobooks.

One way to employ the Web 2.0 is to directly add content to a personal photobook. Semantics, like descriptions or locations, extracted from the photobook can be used to formulate adequate search requests to, e. g., photo community sites like Flickr or online encyclopedias like Wikipedia.

Besides photos, other visual media can be an attractive means to enrich personal photobooks. With the availability of location information from the photobook, for example, geographical maps from services like Google Maps or Yahoo Maps can be retrieved. A map showing the route taken in a 4-week trip to Mexico can, e. g., be added to the beginning of the photobook.

Also textual content can augment one's personal photobook. For example, web services like *geonames.org* provide a geographical name, like a city or country, for a given GPS-Position. This can be used to, e. g., be added as a label to a specific photo in the album, which is equipped with a GPS datum. But it can also act as a search request to other services. The result can be, e. g., a Wikipedia article describing the place the photos were shot. An excerpt, perhaps the first paragraph, can be used as a description, which can be added to the photobook. Also photo communities often provide rich textual descriptions of photos.

In this Chapter a generic system is described which leverages contextual information which is available during a photobook design process to retrieve content from the internet as a sensible extension to the individual photobook. For this generic system a couple of modules are provided to support geographical maps, photos and textual content. The system is meant to be integrated into the generic photobook design process.

8.1 Rule-based Content Augmentation

The core idea of the photobook augmentation system is to automatically decide which kinds of contents are sensible for a specific part of a photobook (page, double page, chapter, ...) and then to try to retrieve this content from the web. In a nutshell, for an input photobook, the system formulates a number of search requests which results are then integrated into the photobook. To be able to perform adequate search requests on web services a good semantic understanding of the photobook and its content is needed. One way to achieve this is to extract metadata from the photos itself. For this the photo analysis system presented in Section 6 is used. Additionally, semantics are derived from the structure of photobook, e.g. the number and size of photos on a page. The more information is available the easier it is to formulate adequate search requests. To not blindly flood the photobook with content when certain information are available, it is necessary to limit the amount of added content by putting restrictions on the augmentation process. For this we follow a rule-based approach. Each rule stands for the definition of one specific kind of content which is subject to augment the photobook. A rule consists of a set of preconditions which have to be met in order to have the rule evaluated, a description of what content should be queried with what parameters and how this content should be integrated into the photobook. An example for such a precondition can be, e. g., that the amount of photos on a page should not exceed a certain number to prevent the photobook page from appearing too packed. For actually performing a request on a web service, definitions for such services have to be provided which have to be registered to the system.

Figure 8.1 shows an overview of how rules are used to integrate Web 2.0 content. An existing set of photos or a photo album together with relevant semantics is used to determine which rules are applicable for the photobook, that means for which rules the preconditions are met. These rules are evaluated, that means it is determined if a service is registered that provides the requested content. It is possible that more than one service has to be called for this. A good example for this is a query to Flickr to retrieve photos which are tagged with the name city other photos on a page are tagged with. When these tags are not available but the photos contain GPS data, the corresponding city name can be retrieved from a web service such as *geonames.org*.

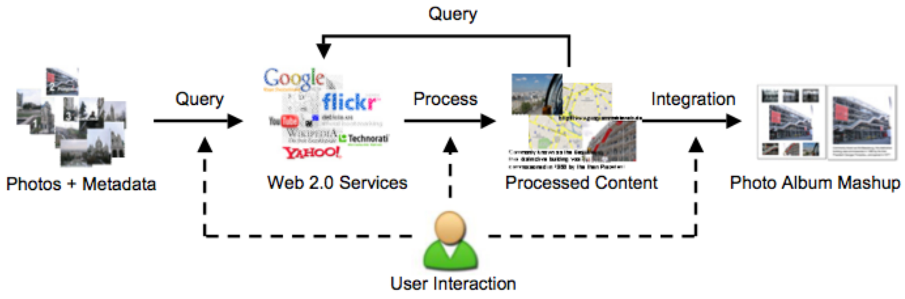


Figure 8.1: Steps of integrating Web 2.0 content into a photobook

The resulting content is integrated into the photobook. How this is done is defined by a set of rules. The photobook user can, but does not have to be integrated into this process. He can, e. g., control which parameters are used for querying external sources, perhaps by manually adding annotations to a specific photobook page. It is also possible that rules do not define how content should be added to the photobook but that the users can choose from a list of automatically retrieved content.

8.2 Augmentation Rules

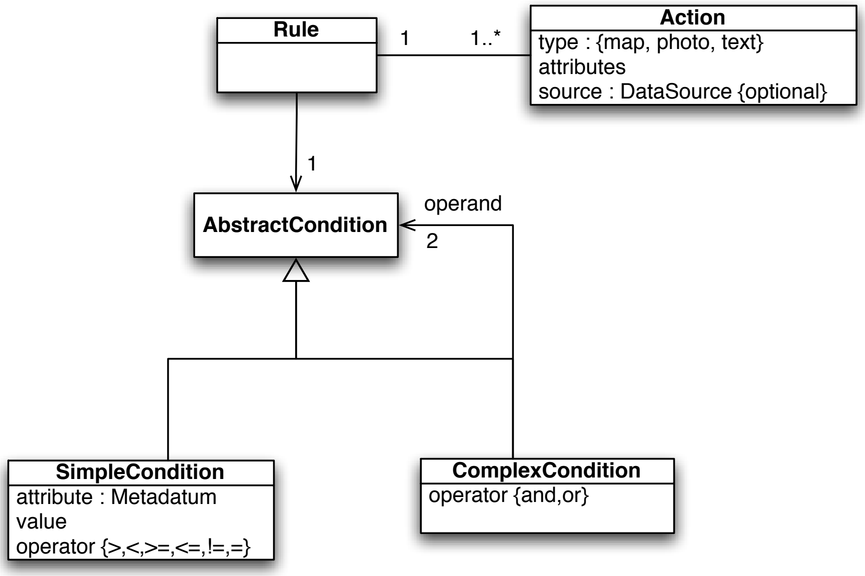


Figure 8.2: UML Class Diagram for he representation of rules for querying external data sources

For the definition of rules we distinguish between four granularity levels: A rule applies either to a photo, a photobook page, a photobook section or a whole photobook. A rule consists of a condition and an action. The condition itself can be a basic condition or a complex condition which combines two rules by a boolean operator. A basic condition defines a simple restriction, e. g. the number of pictures on a page is three or more. The action part of a rules defines what should be retrieved if the condition are fulfilled. This is the type of media (a map, a text, image, ...) and a set of parameters for this media type. These are static parameters like a text length restriction. The data source from which the content should be retrieved is an optional attribute. If not present, the pool of defined data sources is searched for a suitable web service. An overview over this model is depicted in Figure 8.2.

We briefly describe each rule and present a semi-formal definition based on OCL[1].

Evaluate Location

A rule does not necessarily have to lead to instant visible changes in the photobook. The following rule enriches a photo which consists of a GPS information with the according location names.

```
context Photo::evaluateLocationEnrichmentRule()
pre: captureLocation -> notEmpty()
pre: countryOfCapture() -> empty()
pre: cityOfCapture() -> empty()
pre: streetOfCapture() -> empty()

body: let
        parameters = captureLocation()
      in
        countryOfCapture =
          ServiceConnector::
            getCountry(parameters).content)
        cityOfCapture =
          ServiceConnector::
            getCity(parameters).content)
        streetOfCapture =
          ServiceConnector::
            getStreet(parameters).content)
```

Evaluate Map

This rule applies to a photobook page. It says that if the number of photos on a page is not larger than two and if at least one photo consists of a location information, a map should be retrieved with these locations as parameters and the map should be added to the page.

```
context Page::evaluateMapRule()
```

[1] Object Constrained Language

```
pre: photos -> size() < 3
pre: photos -> size() > 0
pre: photos -> exists(photo |
       photo.captureLocation -> notEmpty())
body: let
        parameters = photo.captureLocation()
      in
        photos = photos@pre.union(
          ServiceConnector::
           getMap(parameters).content)
```

Evaluate Location Label

This rule employs information from the photo level rule defined above can use this information to add a descriptive label to the beginning of the album episode, if all pictures have been taken in the same city.

```
context AlbumEpisode::evaluateLocationLabelRule()
pre: pages.first().pageLabel -> empty()
pre: pages.photos ->
       forAll(p1, p2 |
         p1.cityOfCapture = p2.cityOfCapture)
body: let
         label = pages.photos.first().cityOfCapture
      in
        pages.first().pageLabel =
          'Visiting '.concat(label)
```

Evaluate Title Picture

The following rule looks for an additional photo for the title page when a title is present in a photo album.

```
context Album::evaluateTitlePictureRule()
pre: episodes.first().
       pages.first().photos -> empty()
pre: title -> notEmpty()
body: episodes.first().
        pages.first().photos@pre.union(
          ServiceConnector::
            getPhoto(title))
```

Connectors to Web 2.0 sources

The rules in the previous paragraph do not define, how content is actually retrieved. Instead a single service connector is called for this. The idea behind this is to keep the rules independent from specific services. Concrete definitions for services can be registered at this service connector. An example for such a service is:

```
context GoogleMapsService::getContent(parameter)
pre: parameter.captureLocation -> notEmpty()
post: result.type('Map')
```

This definition specifies that the provided parameters should consist of at least one location metadatum and that content of the type *Map* is provided. Based on these preconditions and the provided output the right service for a rule can be chosen.

8.3 Integration into photobook creation process

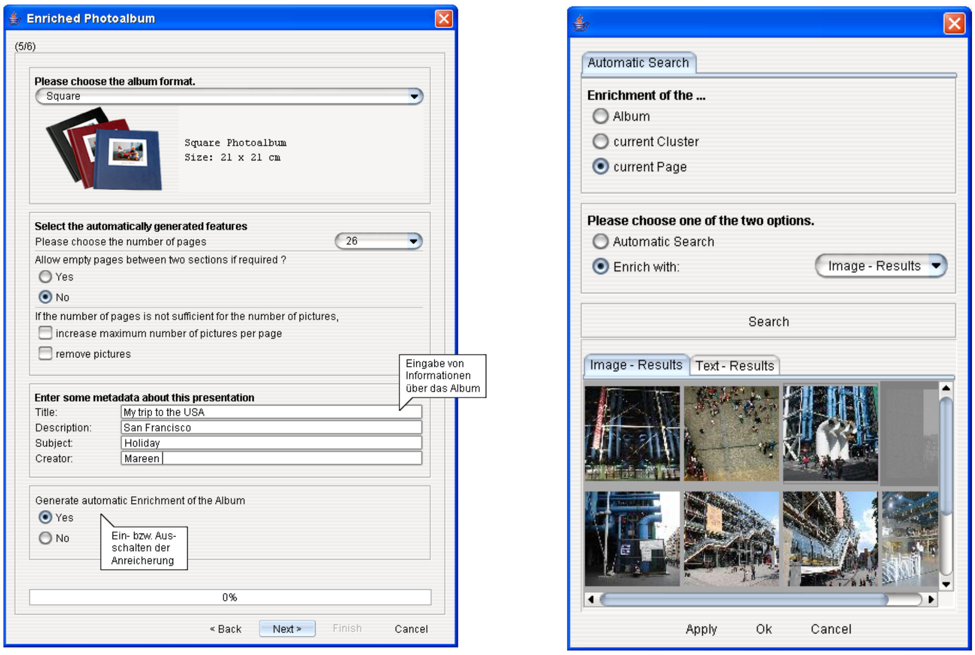
The described system and exemplary rule sets are meant to be integrated in into the automatic photobook creation process. This can either be done in a fully automatic or an interactive way:

In a fully automatic approach the system is integrated as shown in Figure 2.1. Here the augmentation is done after selecting and clustering, but before actually placing the photos on the individual photobook pages. Based on the defined rule and available information the individual photo clusters are augmented with additional content. A screenshot as part of a wizard for the automatic photobook generation is shown in Figure 8.3a. Here the user can choose, if the photobook should be augmented with additional content or not.

In addition the augmentation is automated in the manual photobook design process when the user is adjusting an automatically designed photobook or has started a photobook from scratch. In this approach the user can choose which part of the photobook (current page, episode or album) should be augmented and with what kind of content. An example of such an interaction with the system is shown in Figure 8.3b. In this example photos are shown which are relevant for the current page. For this the augmentation rules are chosen which correspond the page level and result in the media type image. In this concrete example some photos on the corresponding page were equipped with a location stamp in the Exif header and a rule to search for nearby photos at Flickr has been chosen.

8.3.1 Examples

Figures 8.4 - 8.7 show examples of the application of the augmentation system. Figures 8.4 and 8.5 are the results of a manual application of augmentation rules according to



Enriched Photoalbum

(5/6)

Please choose the album format.

Square

Square Photoalbum
Size: 21 x 21 cm

Select the automatically generated features

Please choose the number of pages: 26

Allow empty pages between two sections if required ?

○ Yes
● No

If the number of pages is not sufficient for the number of pictures,

□ increase maximum number of pictures per page
□ remove pictures

Enter some metadata about this presentation

Title:	My trip to the USA
Description:	San Francisco
Subject:	Holiday
Creator:	Mareen

Generate automatic Enrichment of the Album

● Yes
○ No

0%

Eingabe von Informationen über das Album
Ein- bzw. Aus- schalten der Anreicherung
< Back
Next >
Finish
Cancel

(a) Automatic

Automatic Search

Enrichment of the ...

○ Album
○ current Cluster
● current Page

Please choose one of the two options.

○ Automatic Search
● Enrich with: Image - Results

(b) Semi-automatic

Figure 8.3: Dialogues as part of a wizard for the automatic generation of a photobook. On the automatic variant the user can only decide, of the photobook should be augmented with additional content or not. In the semi-automatic variant he can augment parts of the photobook within manual authoring process.

the dialogue shown in Figure 8.3b. The user has chosen to augment the the current photobook part with additional content. As some of the photos are augmented with location information the rule *Evaluate Map* is applied to the page. In the system a map service connector is defined which retrieves a map from Google Maps where the locations of all photos on the page are shown. This map is then added to the current page and the design of the page is then rearranged (See Chapter 9 for details on this). The second example in Figure 8.5 shows the application of a rule to retrieve additional photos from the same location from FLickr when the there is only one photo on the current page which consists of location information. The retrieved photos are added to the page and the layout is again rearranged.

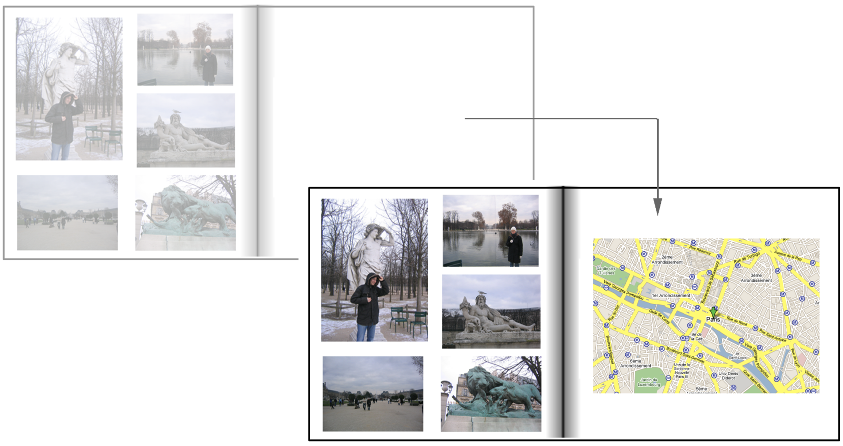


Figure 8.4: Application for a rule to augment a double page with a suitable geographical map if location data is available.

The example in Figure 8.6 shows the integration of the augmentation system in a web-based photobook design tool based on Microsoft Silverlight. The photos on the shown page were all taken near the Louvre and are equipped with location information. The has been augmented with a map retrieved from OpenStreetmap containing thumbnails of the photos at the respective places they were taken. Additionally an excerpt from a Wikipedia article was added with the help of the web service DBPedia. DBPedia is a *semantic layer* on top of Wikipedia which enables developers to access the metadata of articles through a webservice. In this case the location information attached to the Wikipedia article for *Louvre* was used to retrieve the article describing a place which is the closest to the pictures shown on the page.

8.4 Summary

We have described a method for the automatic and interactive augmentation of digital photobooks with the web. This approach is based on the definition of rules which intelligently combine semantic information present in the photos and the photobook and

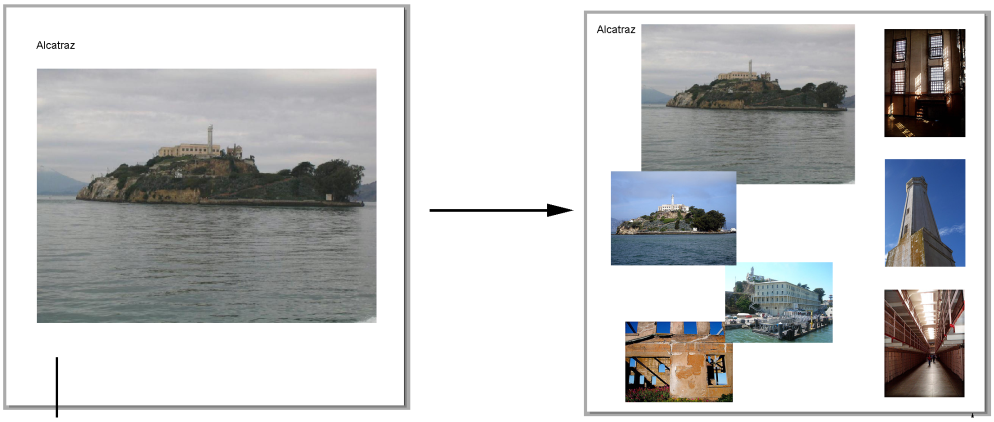
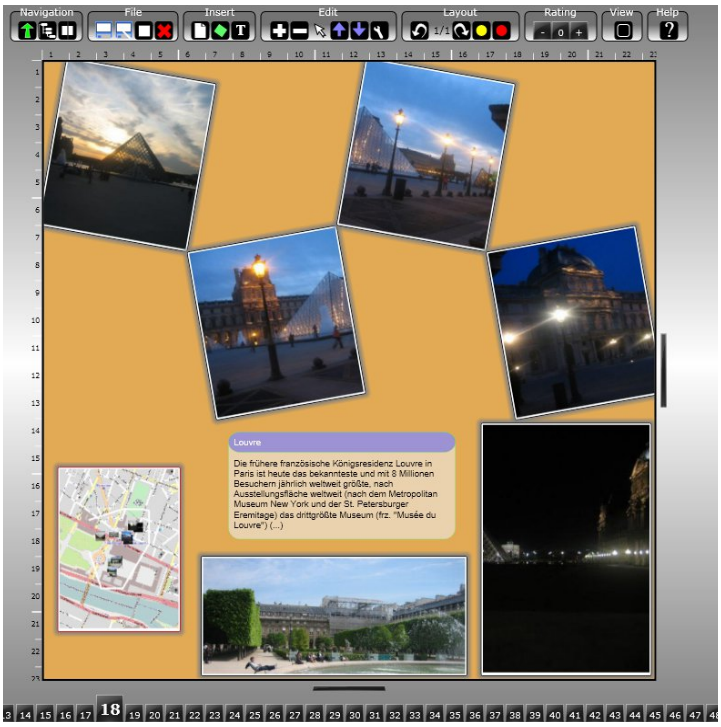


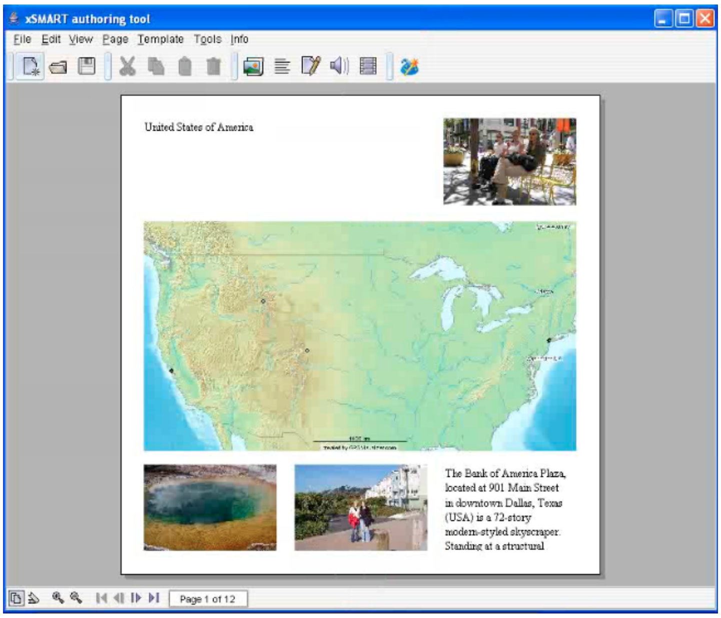
Figure 8.5: Example of augmenting a photobook page with Flickr photos based of text on the photobook page.



Louvre

Die frühere französische Königsresidenz Louvre in Paris ist heute das bekannteste und mit 8 Millionen Besuchern jährlich weltweit größte, nach Ausstellungsfläche weltweit (nach dem Metropolitan Museum New York und der St. Petersburger Eremitage) das drittgrößte Museum (frz. "Musée du Louvre") (...)

Figure 8.6: Integration of augmentation system in a prototype based on Microsoft Silverlight. Automatically added contents are a generated map showing where the photos on the page were shot and an excerpt from a Wikipedia article describing the visited place (Louvre in Paris).



United States of America

The Bank of America Plaza, located at 901 Main Street in downtown Dallas, Texas (USA) is a 72-story modern-styled skyscraper. Standing at a structural

Figure 8.7: Integration of augmentation system in the xSmart authoring system. The title page of an album is augmented with a geographical map showing places visited in a USA trip and a descriptive text retrieved from Wikipedia.

webservices suitable for the augmentation. In this chapter we have proposed connecters to exemplary webservices providing image, geographical maps and text content, which are common additions for photobooks. However, our system is flexible enough to be extended by additional connectors.

9 Design and Layout

Crucial for visually appealing compositions of photos are the layout and the design of the photos in the 2D space. By the layout and style of the composition on the users aim to create visually appealing presentations which should reflect his or her experience of the event captured in the photos. The *layout* of these composition typically defines the size, rotation and placement of content elements. The *design* includes aspects such as color schemes, additional design elements in the presentation, or decorations of the content.

People appreciate if their composition of photos looks nice and makes a nice present or souvenir. The creation of an appealing photo composition, however, can become a very tedious task. Many users do not have the compositional skills and the technical skills needed in a creative authoring process with today's commercial editing programs. Research and industry started addressing the demand of photo compositions by automating certain aspects of the design and layout processes. Applications such as the SmileBook photobook creator[1] or Apple's photo management system iPhoto[2] provide rich sets of professionally designed templates defining the layout and design of photobooks, calendars, greeting cards and so on. The underlying templates are prepared by skilled designers with a sufficient knowledge about principles of layout and design. With the respective authoring tool the end user can then add photos and select the desired template. A few clicks and these tools fills placeholders with the user's personal photos and create the photo composition. Presentations created this way usually lead to a pleasing result but rather factual and uniform presentation, due to the limited amount of templates for a given number of photos. Skilled users might exploit all the different options of professional graphics authoring tools and typically spend much time to manually achieve an if at all satisfying result. Depending on the user's abilities the results might be great but given the skills and the time the author needs to create the composition this is addressing only a few expert users.

The dilemma is that the unexperienced and potentially technically uninterested user has no way of becoming a good designer by the tool. The latter, however, is a crucial factor when it comes to the commercialization of photo products such as photobooks, calendars, or posters. In an age in which the number of digital photos is exploding, companies in the field of photo finishing and photo services are searching for new products and business models to convert these photos into visually appealing digital presentations or physical products. We argue that the driving factor for success will be to *bring design to the people*, i.e., to transfer knowledge from visual arts and design into a wizard-like authoring environment. We can literally hear the outcry by professional designers saying that layout and design will never be automatable and that the end result will never be meeting the standards of a manually and professionally curated presentation. Well, let's give it a try and see how far we can get for the end consumer.

[1] `http://www.smilebooks.com`

[2] `http://www.apple.com/ilife/iphoto/`

In this chapter, we present an approach for the automatic generation of photo compositions based on fundamental design and layout principles. We follow a generative approach in which the photo composition is created from a set of underlying rules. Each composition is unique and arranged according on the individual photo set.

Existing approaches are either purely template-based or provide algorithms that producing layout that are usually not following any sophisticated, aesthetic design principles. We take a different approach and specifically take aesthetic rules as our starting point. Additionally, we explicitly consider not only photos but also media types such as texts into account for the resulting layout and accommodate for the specific characteristics of these. Finally, we take the content of photos into account to a degree not present in other approaches by considering features such as the color layout, the saliency, and the presence of faces in the photos for the design of backgrounds, the spatial composition and the placement of photos in the foreground and background.

Our technical solution heavily relies on the use of genetic algorithms as these very much comply to the nature of our problem. Therefore we want to quickly review the rationale behind genetic algorithms and how they apply to our approach. Genetic algorithms do not mimic or model a specific process, but create solutions more or less randomly and evaluate the results according to certain criteria. This very much fits our problem for album page layout as it is very hard to come up with an algorithmic solution to produce a pleasant layout. It is however possible to find measures for certain characteristics and provide ratings for the compliance to certain rules. The general idea behind genetic algorithms is to simulate evolution by generating several potential solutions (populations) and then to combine (crossover) the best solutions (survival of the fittest) resulting in a better population of better solutions which is fed back into the evolution cycle (illustrated in Figure 9.1). As in evolution, also inferior solutions can survive by chance or a solution can mutate randomly. When applying this general concept to a specific problem, three important tasks have to be performed:

- The solutions to a problem have to be *encoded* somehow in a specific data structure or genotype. A specific instance of such a genotype is called a chromosome which consists of several genes, possibly organized in a specific structure. Another important task is to provide methods to translate a chromosome into the actual solution of the problem (phenotype).
- Based on a defined genotype specific genetic operators have to be defined which combine two chromosomes (crossover) or mutate a given chromosome by altering the values of its genes or altering the structure of the chromosome.
- The most important task is define a *fitness function*. This fitness function rates a specific chromosome and thus decides if it is kept for the next evolution or not and determines the best resulting solution.

The rest of this chapter is organized as follows. We identify relevant design and layout rules from the literature and examples of professionally designed photo albums which

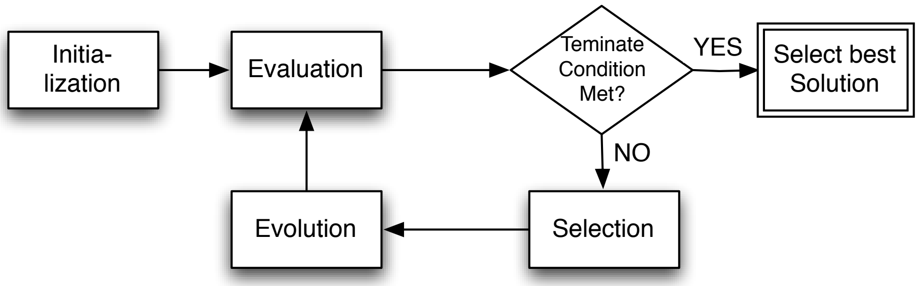


Figure 9.1: General layout of genetic algorithms

are the basis for the development of our automatic layout system described in Section 9.1. These rules are the basis for our system for automatic photobook layout presented in Section 9.2. This system is implemented in two applications described in Section 9.3 before closing with a conclusion.

9.1 Aesthetic Principles for photobooks

One does not need to be a skilled designer to distinguish between an appealing photobook presentation and an unaesthetic photo presentation. Some photobooks instantly catch the viewer's eye and others are not really a pleasure to look at. Looking more closely at such photobooks reveals certain patterns and applied rules regarding layout and design in the more appealing books. Many of these rules have been known for a very long time and are employed by skilled artists and designers. In this section, we review selected rules for visual layout from the literature and discuss how they can be mapped to the automatic generation of photobooks. We also analyze samples of photo compositions that have been built by skilled designers to verify the usage and applicability of these rules in the context of our domain.

9.1.1 Principles of layout and design

The rules and principles we consider for visual layout are mainly adopted from the the works presented in [Itt97, LHB03]. Lidwell [LHB03] gives a good overview over universal design principles which are approved over many years by many designers. Additionally Itten [Itt97] provides a good overview over different models for color combinations. From [LHB03] we identified those principles which are in our opinion applicable for photobooks. We divide these principles into the categories spatial layout, color layout, and highlighting. We show how these principles are eligible and applicable for photobook designs.

9.1.1.1 Spatial layout

Two of the central aspects of a layout are the size and position of the contained items. A well-known underlying principle creating the impression of balance and stability is the *golden ratio* (divine proportion) which describes a rule for the relation between two lengths: If a line ac is divided by a point b into the segments ab and bc the resulting sections follow the golden ratio if $bc/ab = ab/ac = (\sqrt{5} - 1)/2$. A similar proportion can be achieved by employing the *Fibonacci sequence*, which is defined recursively:

$$Fib(i) = \begin{cases} 1 & i = 0, i = 1 \\ Fib(i-2) + Fib(i-1) & i > 1 \end{cases}$$

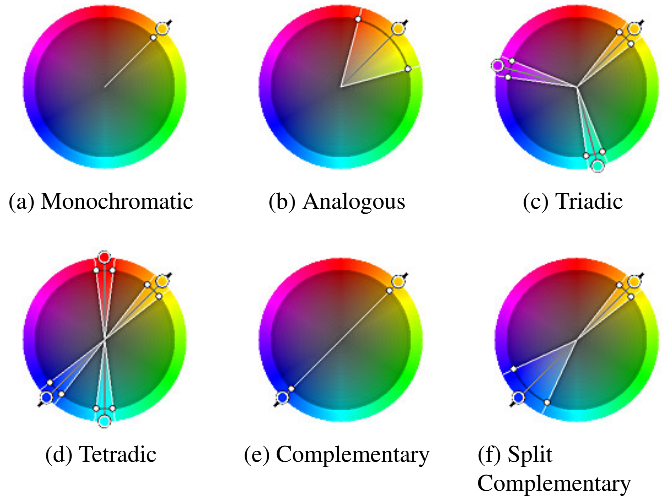
The proportion defined by two consecutive Fibonacci numbers is similar to the golden ratio. By taking more elements of a Fibonacci sequence into account, additional visually appealing proportions can be determined. These principles can be applied to many aspects in the photo composition such as the proportions of the size of different sub-areas or the position and relation of width and height of items on a page. Layouts following these principles are usually perceived as more balanced and appealing than layouts that do not. Golden ratio and Fibonacci sequence are well-known for the design of still images, e.g., by placing the horizon or the main object in a picture not in the middle of the image but at a point according to the golden ratio or the Fibonacci sequence. When thinking of photo collages these principles can be employed when placing photos on a page, e.g., at or along sections of virtual horizontal and vertical lines following the golden ratio. Similarly, the aspect ratios of photos and more general logical subareas in a collage should follow these rules to create a balanced and visually pleasing layout.

Another principle for distributing objects and partitioning areas are different forms of *symmetry*. According to [LHB03], symmetry creates an impression of health and balance. Lidwell distinguishes between different kinds of symmetry: *reflection* symmetry, *rotation* symmetry, and translation symmetry. Reflection symmetry refers to the mirroring of an equivalent element around a central axis or mirror line. Rotation symmetry is referred to the rotation of equivalent elements around a common center. Translation symmetry refers to the location of equivalent elements with the same orientation and size in different areas of space. Translation symmetry is often combined with additional restrictions such as the placement of several elements along a line. These symmetry principles can, e.g., be seen in nature: A butterfly or a human's face exhibit reflection symmetry, a sunflower exhibits rotation symmetry in its petals around the center of the blossom. These symmetries can be used to distribute photos over a page and also to partition the page in subareas that are filled with different photos.

9.1.1.2 Color layout

When different colors are applied to the same design they usually lead to different perceived emotions [LHB03]. An important guideline is to limit the number of colors to the amount which can be processed at one glance. [LHB03] suggests about 5 colors for this. In the following, we present the principles for combinations of colors and the choice of a color in different situations.

Color combinations: We can observe that some color layouts are perceived as more appealing than others. Looking more closely at such appealing color combinations one can observe a number of patterns. [Itt97] has structured such combinations of colors in different color schemes. One of the main attributes of a decent color layout is the limitation of colors. [Itt97] suggests not to consider more than three and has structured such color combinations into six schemes which are depicted in Figure 9.2. The color wheels are a flat representation of the HSV color scheme where the angle designates the hue value and the distance to the center the saturation of the color. One can see that different saturations of the same hue can always be combined which is shown in the monochromatic scheme (a)) which only allows the same hue value for all colors. Other combinations incorporate adjacent colors (analogous, b)), colors at the corners of a symmetrical polygon circumscribed in the color wheel (triadic,c) and tetradic,d)) or opposing colors (complementary,e)). The split-complementary scheme (f)) is a variation of the complementary scheme where additional, adjacent colors of the two opposing colors are allowed. These color combinations can not only be found in artificial design but are also present in nature, e.g., in color combinations in the petals of flowers.


(a) Monochromatic (b) Analogous (c) Triadic

(d) Tetradic (e) Complementary (f) Split Complementary

Figure 9.2: Different color schemes [Itt97]

We employ these color schemes to rate the overall color layout of a photobook page

Color choice: Color schemes provide an easy way to algorithmically determine combinations of matching colors. To employ these rules for photo collages we have to decide for which colors we want to determine matching colors. We opted to determine the colors based on those colors appearing in the pictures contained in the presentation for two reasons: Colors that appear in nature, such as in landscape images, are perceived as harmonic and natural [Itt97], which is a good starting point for finding matching colors. Second, colors in the images are part of the overall color layout of a photo composition anyway. Thus, it is reasonable to consider the dominant colors in the photos of a photo composition as a starting point for the determination additional, matching colors for the color layout of a composition.

9.1.1.3 Highlighting

Not all elements in a presentation have the same importance. One means to reflect this is the distinction between background and foreground. [LHB03] suggests that the foreground should be clearly distinct from the background to create a more appealing impression. They present various means to highlight elements on the one side and on the other side making other objects less salient. Regarding color, objects are generally perceived as more salient if they incorporate a couple of vivid and rich colors. The reverse is that lowering these factors and reducing the number of colors results in a less salient impression. Another way of highlighting objects is to create the visual effect of physically separating objects from the their surroundings. This can, e.g., be achieved by employing a drop-shadow effect which creates a three-dimensional effect and visually separates an element such as a photo from its background.

9.1.2 Professionally authored photo albums

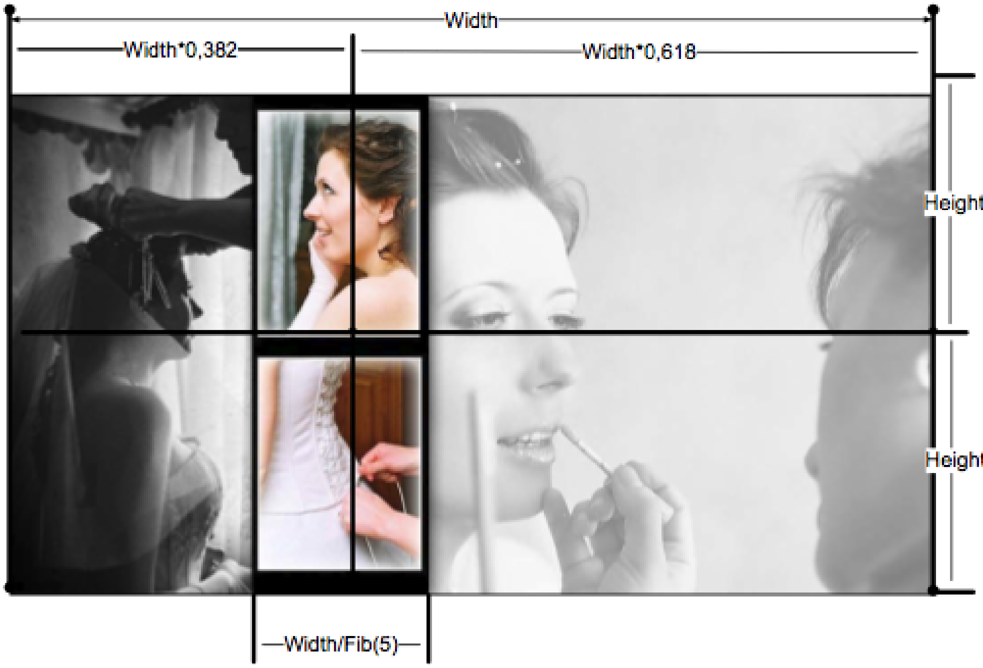


Figure 9.3: Example of professionally designed wedding photo album, the preparation of the bride following principles of color choice and layout according to golden ratio and the Fibonacci series, and symmetry

In addition to the review of relevant design principles applicable to photo compositions, we analyzed several professionally presentations regarding the application of these rules. We carried out a qualitative analysis of selected photo compositions from public album sites on the Web[3][4][5]. Our findings show that the principles we identified for design and layout hold true for most of these examples. An exemplary page of a wedding album is shown in Figure 9.3. The page shows the scene of the preparation of the bride. The photos show the finishing of the make up, the putting on of the earrings, the dressing of the corsage and bridal veil. From a design and layout perspective we see that the focus is put on the images in the foreground by maintaining the photos in color while the background photos are reduced to black and white. One can also see that the layout is symmetric in many ways. The page is reflection symmetric along a horizontal line the middle. The two foreground photos are placed in translation symmetry and reflection symmetry. The ratio of the horizontal aspects of the foreground photos and the whole canvas follows the rule of the Fibonacci sequence. The two foreground photos are vertically aligned along an axis according to the golden ratio. An interesting characteristic of this and most other photo albums we analyzed is that often one or two photos are used for the background and a comparably small percentage of the area is highlighted with single photos which coincides with the highlighting principle presented by [LHB03].

9.2 Automatic Photo Album Layout

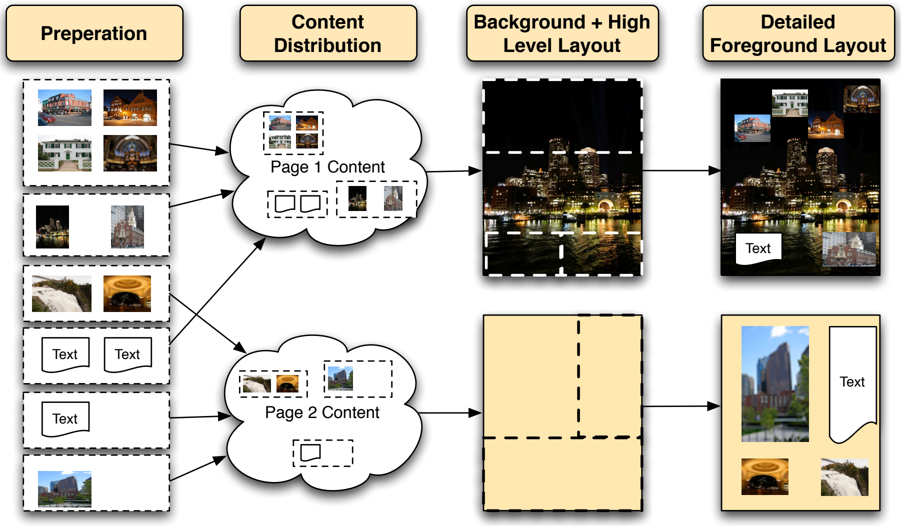


Figure 9.4: System overview for our automatic photo album layout

The aesthetic principles for photo album layout discussed in the previous section form the basis for our proposed system for automatic layout of digital photobooks. As input

[3] http://customalbumdesign.com
[4] http://www.embassyprobooks.com
[5] http://www.rhemastudio.com

to our system we assume an ordered set of photos and text elements. Usually the order of the photos is derived from their time stamps but can also depend on the application. The text elements are assumed to be assigned to one photo or a series of photos. An origin of these kind of media could be a travel blog with text and images from which one would like to create a photobook. But also other type of sources like photo management tools or social media would come with photos and additional text in some kind of order.

Based on the input media the system creates an automatic layout in five consecutive steps which are illustrated in Figure 9.4. The *Preprocessing* analyzes and structures the input elements into different groups that form the basis for the different pages. These groups are then assigned to the different pages in the *Content Distribution*. The creation of the layout begins with the determination of the *Background Layout* of the individual pages. The creation of the layout of the foreground is divided into two steps: For a High-Level Foreground Layout the pages are divided into separate areas according to certain layout principles, one for each element group. These areas are then laid out detailed in the *Detailed Foreground Layout* step. The different steps of the automatic layout creation are presented in the following.

9.2.1 Preprocessing

To be able to assign the different image and text items to the single pages, the content distribution step expects the elements to be structured in a certain way. This is accomplished in the preprocessing step. Basically two operations are performed on the input photos and text elements:

- In case no pre-defined order and structure is present in the photo set, they are ordered and clustered according to their time stamps. For this we employ the algorithm presented in [PCF02]. By this we keep semantically close photos also together in the final photobook and thereby support the story-telling aspect of a photo album. The parameters of the clustering algorithm are chosen in a way that at 5 photos form a group and the majority of groups only consists of one photo.
- Large text blocks are divided into several text blocks. The threshold number of characters used for the split depends on the resulting size of the pages. The reason behind this is to also allow longer text passages in the photo while avoiding cluttering the pages with text and maintaining a minimum threshold for the font size. Longer text passages can thereby be split over different parts of a page or even different pages in the resulting album. Additionally, if not already present, a distinction between headings and descriptive text is made. This categorization can usually be derived from the application. If no distinction is present, we simply categorize text elements with 5 or fewer words as headings.

The intermediate result is an ordered set of sets containing either one ore more photos, a long text, or a heading text. These sets are assigned to the single pages in the following step.

9.2.2 Content Distribution

The purpose of the content distribution step is to distribute the preprocessed elements over the photobook pages. For this we assume that the number of pages is predefined, which is reasonable in practical applications: For commercial photobook printing services, due to production limitations, usually only a fixed number of page numbers (e.g. 8, 16, 32, ...) and page sizes are offered. An optimal distribution of photos and text elements is defined by best meeting the following restrictions in our system:

R1 The predefined order of items should also be kept in the layout of the photobook.

R2 The pages should be visually balanced, this means that roughly all pages should contain the same amount of items.

R3 The text to photo ratio should be roughly the same on all pages. Therefore, the system should avoid to have pages only containing text or only photo items.

R4 The content distribution should not extend the defined maximum of pages but at the same time avoid empty pages. This restriction stems from practical facts in some of our applications. Usually, when designing a printed photobook, one can only choose to increase the number of pages in multitudes of pages of, e.g. eight. A person who wants to order such an album usually wants to avoid having empty pages in the photobook but also wants to limit the number of pages as the price of the printed photobook increases with the number of pages.

R5 Ensure that all page elements do really fit the page. Images should not be scaled down too much to ensure their visibility on the page and the font size of a the element should not be too tiny to be easily readable. This leads to minimum sizes both text and images.

R6 The color layout of the page should be balanced, this means combinations of photos corresponding to one of the color schemes presented in Section 9.1 should be preferred.

The problem of content distribution can be seen as an optimization problem of which an optimal solution is that solution which best meets these partially competing conditions. We opted to solve this problem with the help of a genetic algorithm. For our application of the genetic algorithm, we defined the chromosome as a simple list of genes (`nextPage`), one for every for every element group. These genes have boolean types and decide if the corresponding element group is placed on the current or the next page. The phenotype or interpretation is then done simply defined by going to the ordered list of genes, placing the corresponding element groups on the current page (starting with the first page) and switching to the next page, if a corresponding gene has a positive value. Hereby we implicitly meet the restriction R1. The genes are simply initialized by evenly distributing all elements over the pages so that every page contains the same amount of groups. The interesting part of the genetic algorithm is the definition of the fitness function. This function rates a specific distribution of groups over the pages of a photobook and consists of the sum of the following ratings. We also indicate in brackets [] which rating addresses which restriction from the list above.

- $r_1 = 1/(1 - (\#ElementsOnPage - avgNrElements)^2)$, a rating how close the number of elements meets the average number of elements per page of the book [R2]
- $r_2 = \#PagesWithTextAndPhotos/\#Pages$ [R3]
- $r_3 = 1 - (\#EmptyPages/\#Pages)$ [R4]
- $r_4 = \#PagesExceedingPageArea/\#Pages$, for this all pages are determined where the containing elements exceed the overall page area by more then 80 $\%$. The preferred spatial extension is defined for every photo by a minimum size of $4cm^2$ and for each text element derived from the number of characters and the font size. [R5]
- A rating r_5 is determined by checking, if the dominant colors of all photos on the page adhere to one of the presented color schemes. If such a color schema is met, a rating of 1 is given, otherwise 0. The dominant colors of a photo are determined by building their color histograms in the HSV color space and selecting the bins exceeding a threshold value. [R6]

The overall rating function is defined as $r_dist = \sum_{i=1}^{5} r_i$. The maximum number of evolutions (cycles in the graph, see Fig. 9.1) was heuristically set to 400. As genetic operator only crossover was considered. By this, two chromosomes of a population are combined.

A good tradeoff between the computation time and the quality of the solution could be found at 200 evolutions in experiments. The result of this step is a set of pages with assigned groups of elements.

9.2.3 Background Layout

After distributing all elements over the pages, the single pages are laid out individually. First for every page a background is determined as follows: An important principle regarding background vs. foreground of presentations is that the background should not distract from the foreground. We also have found out that the majority of photobooks is designed by placing a suitable photo in the background. We therefore wanted to prefer layouts where a non-distracting photo is used as the background for a page and if no suitable photo can be determined, a color complementing the colors of the photos on the page should be taken. An overview the algorithm is depicted in Figure 9.5.

In our system, distraction is modeled by two components: *Color* and *Saliency*. We assume that photo with a lot of different colors distract the viewers eye more than photos only consisting of a few colors. We also assume that photos having a lot of salient regions like e.g. a photo showing a rough sea scenery do distract the viewer more than

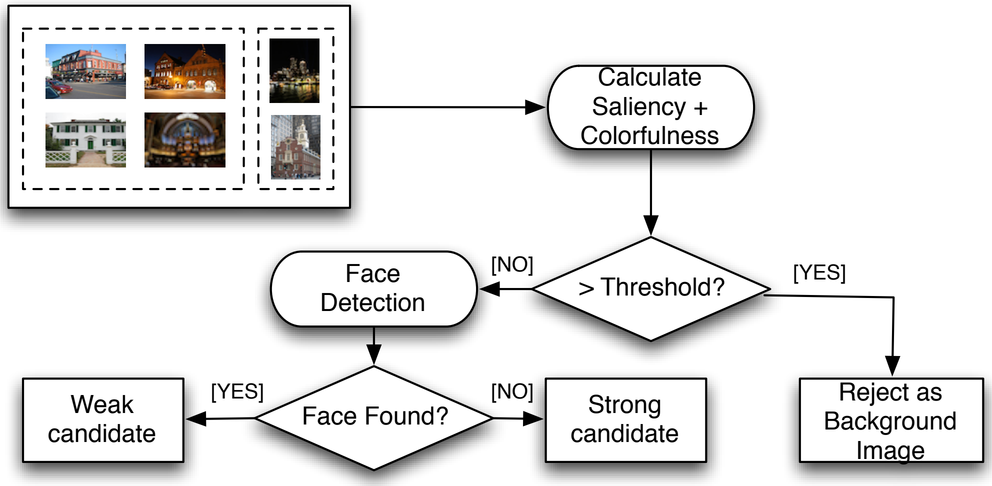


Figure 9.5: Algorithm for the background determination

e.g. a photo with a very calm sea scenery. Thus, we have modeled these two aspects as follows:

- *Color:* To determine the degree of colorfulness of a photo we analyze the histogram of the photo in the HSV Color space. We only consider the histogram of the H component and count the bins which exceed a certain threshold value. The degree of colorfulness is then determined by the number of bins exceeding this value.
- *Saliency:* For every image a saliency map is determined, an example is shown in Figure 9.8. This map indicates regions in a photo which potentially catch the attention of the viewer. For this map we employ an extended version of the algorithm presented in [IKN98]. The saliency map from the original algorithm is overlaid with a gauss map, which amplifies regions near the middle of the photo. This stems from our observations, that the important parts of an image are usually placed in the middle of the photo. Additionally, we employ the face detection algorithm of [VJ01] to amplify regions containing faces. A score for the overall saliency of the image is determined by simply calculating the average saliency value of the resulting saliency map.

For every photo on the page we determine a distraction value according to these scores. Photos exceeding a certain threshold value are automatically rejected as a background photo candidate. For the remaining candidates additionally the number of faces in each photo is determined. Then the photo with the least number of faces is chosen as the background photo. If there are more than one photo with the least number of faces the one is chosen with the least distraction rating. If there are no remaining photos a uniformly colored background is generated.

When a background image is determined, usually some parts are nevertheless interesting to the viewer's eye and should not be overlaid with other elements. To ensure this a additional, empty dummy group is added to the page, when a background photo is determined. Thus the result of this step is a layout for the background for each page together with a set of groups of elements assigned to each page determined from the content distribution step, possibly augmented with a single empty group.

9.2.4 High-Level Foreground Layout

The design of the page's foreground is divided into two phases, a rough and a detailed layout process. In this first step the page area is divided into several rectangular areas, one for each group. The spatial layout of these areas follows several, partly competing restrictions:

RL1 All elements of one group have to fit in the assigned area.

RL2 The pre-defined order of connected groups (text groups) should be reflected in the layout.

RL3 Keep text items within their preferred spatial extension.

RL4 Prefer overall layouts following principles of golden section and symmetry.

RL5 Prefer aspect ratios following the golden ratio for text items.

RL6 Avoid changing the aspect ratio of groups with one photo.

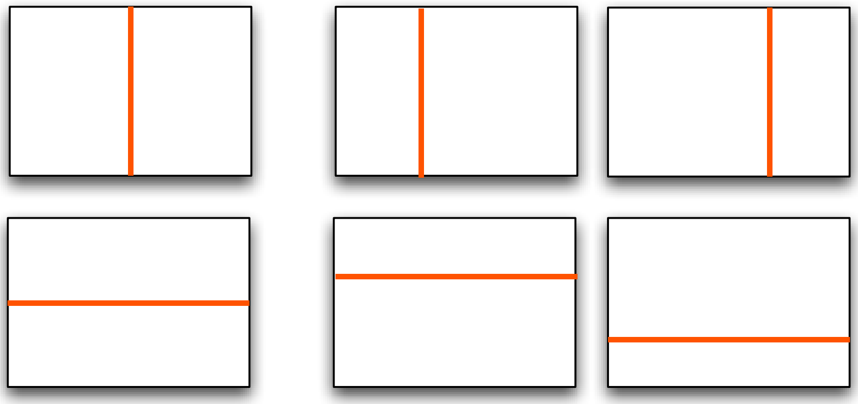
RL7 Prefer keeping headings in the upper part of the page.

RL8 Do not cover salient areas of the background.

RL9 Items should be evenly distributed over the page and thus the visual weight should be placed in the middle of the page.

To meet these rules we again formulated the automatic design employing a genetic algorithm with a fitness function rating designs according to these rules. The generation of rough layouts is accomplished by recursively dividing the page into rectangular areas following slice principles depicted in Figure 9.6. This is again accomplished by employing genetic algorithms. For this, a page is encoded by a chromosome consisting of a series of compound genes, each these compound genes representing an element group. A compound gene is built up by three genes:

- **hOrV:** This boolean gene decides, of the current available space is split horizontally or vertically.
- **whichSide:** This boolean gene decides, in which half of the split area the corresponding element group is placed.
- **splitAmount:** This gene decides, at which percentage the available area is split. The possible values correspond to the slice primitives depicted in Figure 9.6.



(a) Reflection Symmetry

(b) Golden Ratio

Figure 9.6: Slice primitives for rough foreground layout employing the principles of (a) reflection symmetry and (b) golden ratio

The corresponding to a specific chromosome is simply computed by building a graph by interpreting the series of compound genes. Two samples of such graphs and their visual representation in a page layout are depicted in Figure 9.7.

Heuristically, a population size of 50 and a fixed number of 100 evolutions was chosen. As genetic operators only a simple mutation operator was. This mutation operator randomly alters one to three genes in one chromosome. The fitness function evaluating the resulting layout consists of two parts. The first part consists of functions testing for mandatory conditions to the layout. If these conditions are not met, the corresponding chromosome is deleted from the population. These conditions consist of the above mentioned restrictions RL1- RL3. If these restrictions are met, the chromosome is rated according to the following definition (we also indicate in brackets [] which restriction is addressed):

- $r_1 = \frac{\#bord_{rat}}{\#bord}$: This determines the percentage of borders in the layout, which follow the golden section and the Fibonacci Series. By following we mean we determine x position for vertical and the y position of horizontal and set this in relation to overall height or width of the page. Considering the nested nature of constructed layouts, it can happen, that borders do not follow this rule despite they are split according to the pre-defined slice primitives. This can also happen when considering aspect ratios of pages, which do not follow the golden section. [RL4]

- $r_2 = \frac{\#text_{gs}}{\#text}$: This determines the percentage of text items on the page, which aspect ratios follow the golden section. [RL5]

- $r_3 = \frac{\sum_{i \in images} coveredArea_i}{\#images}$: This determines, to which percentage area can be covered by the image for single image groups. This favors layouts following the aspect

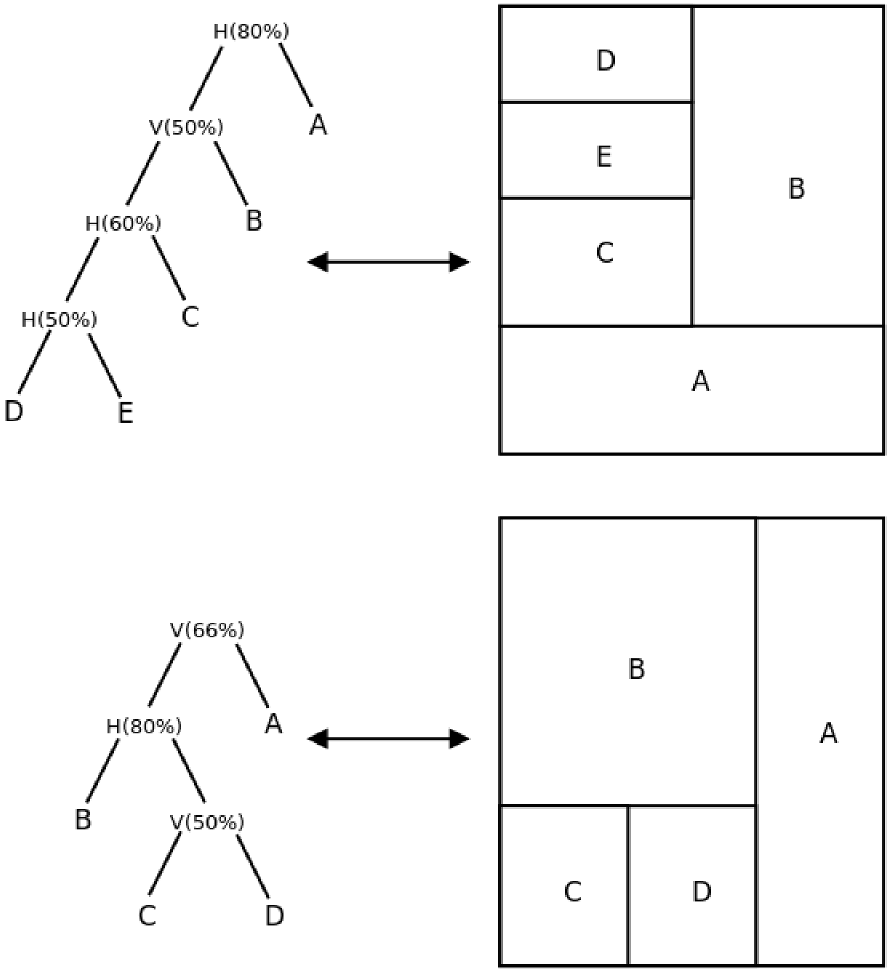


Figure 9.7: Examples of resulting layouts according to recursive algorithm

ratio of the image and downgrades areas being degraded to long stripes or not following the image's aspect ratio. [RL6]

- $r_4 = \frac{\sum_{h \in headings} distanceToBottom_h}{\#headings}$: For Every heading in the page it's distance to bottom of the page is determined favoring headings being in the upper part of the page. [RL7]
- $r_5 = salienceSum(freeArea)$: As mentioned before, if a page consists of a background photo, one empty group is added to the page to allow keeping parts of the page free showing a very salient background. This rating determines the average salience value of the part of the background being covered with this empty group. If the page does not contain a background image, this part of the rating is discarded. [RL8]
- $r_6 = 1 - \frac{\sum_{e \in elements_{page}} \frac{e_x - width_{page}/2}{width_{page}} + \frac{e_y - height_{page}/2}{height_{page}}}{\#elements_{page}}$: This determines a rating for the visual balance of the page. By e_x and e_y the center of an element group e is defined. The closer the average of all of these is placed to the center of the page, the higher the corresponding rating for the visual balance is. [RL9]

The overall rating function is defined as $r_{layout} = \sum_{i=1}^{6} r_i$. This genetic algorithm is performed for every page. The result is a high level layout for each page with a group of elements assigned to each of the resulting sub-areas of the page. These sub-areas are further laid out in the next step.

9.2.5 Detailed Foreground Layout

Giving a high level layout for a page, the different sub-areas are further laid out depending on the kind of their content.

Areas containing of a **single photo** a filled entirely with this photo while keeping a small border. If the photo does not have the same aspect ratio as the assigned sub-area, it is cropped accordingly. For this the saliency map of the photo is determined and the photo is cropped in a way that the resulting photo corresponds to the aspect ratio of the sub-area while keeping the most salient regions of the photo. An example of cropping a photo according to its saliency map is depicted in Figure 9.8.

Areas consisting of **text items** filled entirely with this text. Additionally we ensure, that the text is easily readable on the background. For this, as described in Section 9.2.3, we analyze the color histogram of the covered area of the background. If the background consists of only a few dominant colors (we chose 2), a color for the text is chosen which best complements these colors. If the background consists of too many dominant colors, the part of the background is overlaid with a translucent white area and the text color is set to black ensuring easy readability (see examples in Figure 9.9).

Areas consisting of groups of photos are laid out according to a set of pre-defined layout primitives. These are not templates but rather a set of visual rules to define the

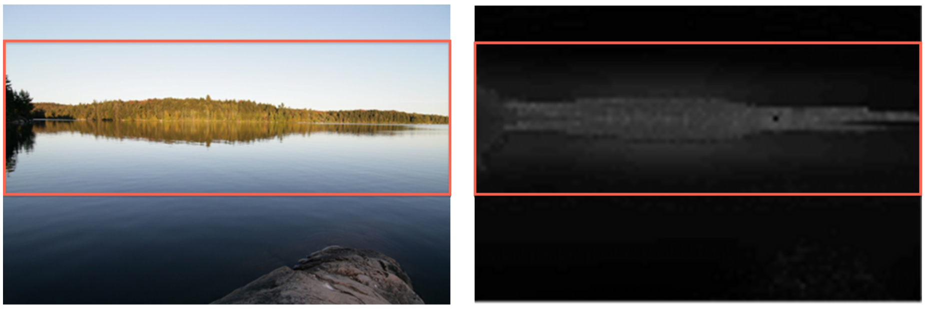
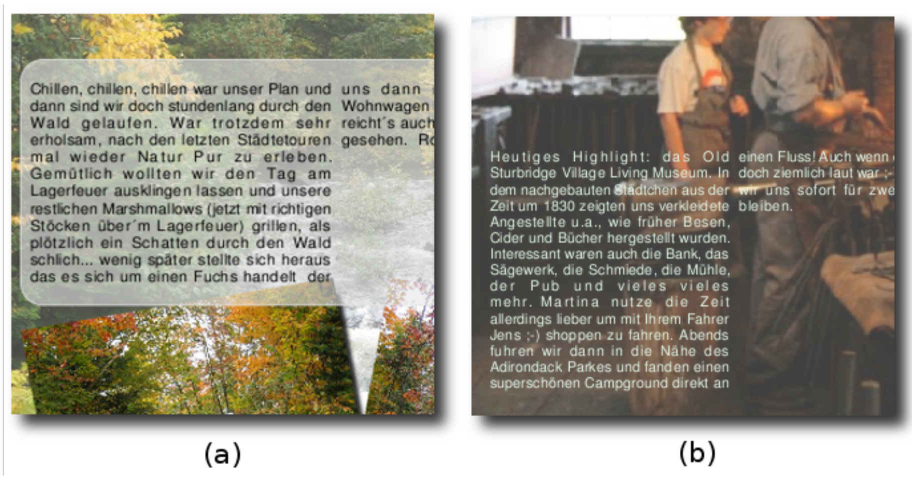


Figure 9.8: Example of cropping an image according to its saliency map



Chillen, chillen, chillen war unser Plan und dann sind wir doch stundenlang durch den Wald gelaufen. War trotzdem sehr erholsam, nach den letzten Städtetouren mal wieder Natur Pur zu erleben. Gemütlich wollten wir den Tag am Lagerfeuer ausklingen lassen und unsere restlichen Marshmallows (jetzt mit richtigen Stöcken über´m Lagerfeuer) grillen, als plötzlich ein Schatten durch den Wald schlich... wenig später stellte sich heraus das es sich um einen Fuchs handelt der uns dann Wohnwagen reicht´s auch gesehen. Ro

(a)

Heutiges Highlight: das Old Sturbridge Village Living Museum. In dem nachgebauten Städtchen aus der Zeit um 1830 zeigten uns verkleidete Angestellte u.a., wie früher Besen, Cider und Bücher hergestellt wurden. Interessant waren auch die Bank, das Sägewerk, die Schmiede, die Mühle, der Pub und vieles vieles mehr. Martina nutze die Zeit allerdings lieber um mit Ihrem Fahrer Jens ;-) shoppen zu fahren. Abends fuhren wir dann in die Nähe des Adirondack Parkes und fanden einen superschönen Campground direkt an einen Fluss! Auch wenn doch ziemlich laut war ;- wir uns sofort für zwe bleiben.

(b)

Figure 9.9: Examples of text areas where the text color was automatically set to the best complementing color black (left) and where it was chosen to add a white shaded area to ensure easy readability

spatial relationships of a small set of photos. Figure 9.10 illustrates some of the layout primitives for our application which follow the principles of symmetry and the golden ratio. The figure shows only examples for a specific aspect ratio, the primitives are adjusted accordingly for different aspect ratios. The primitives are either examples of primitives found in professionally designed photobooks or developed by us following the rules of golden ratio and symmetry. Row (a) depicts four examples where photos are laid out on an invisible line incorporating reflection symmetry along two axes. In the layout primitives shown in row (b) photos are placed along an invisible diagonal line and therefore following the principle of translation symmetry. Row (c) depicts various forms of rotation symmetry and in row (d) the concepts of translation symmetry are present. We limited the the number of photo per primitive to at most four as we experienced from early prototypes that more photos lead to photos which were perceived as too small for the overall layout.

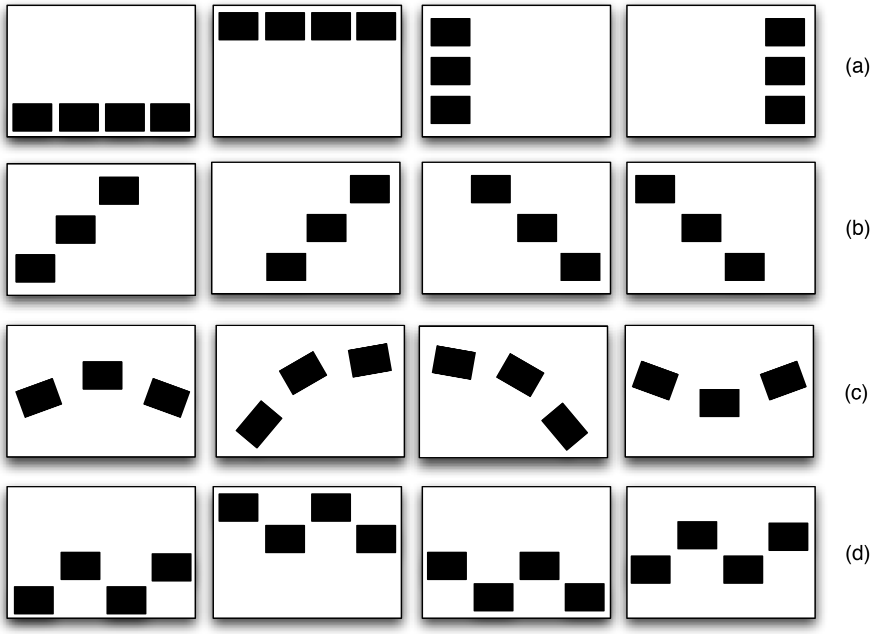


Figure 9.10: Examples of layout primitives used for positioning photos inside a sub-area of a page

Examples of resulting layouts according to our algorithm are depicted in Figure 9.11. The figure shows pages that have applied different of the layout primitives to the different page sub-regions.

9.3 Application

One characteristic of our system is its ability to consider images along with photos as input. Additionally, an already existing structure on these image and text items is also

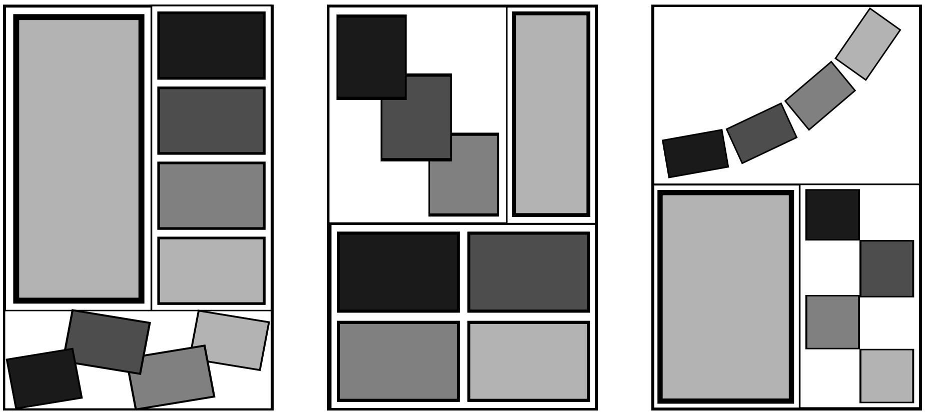


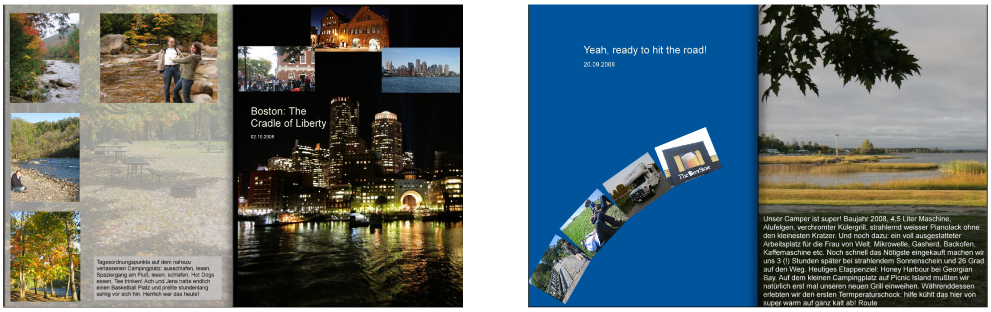
Figure 9.11: Abstract examples of overall page layouts after having completed the detailed foreground layout step.

reflected in the resulting layout. This enables the system to be used in applications where these structures are present and textual content is associated together with the photos. Examples for this are e.g. travel blogs or photos hosted on social network platforms. User who want to convert for example a photo into a photobooks would appreciate if their valuable editing effort be maintained and reflected in the automatic layout. On the other side, our system is flexible enough to also provide sufficient results when this information is not available and can e.g. also work on raw photo sets. In the following, we are presenting two implemented applications of our systems. One application is the automatic creation of photobooks from a travel blog. The one one is the integration of the layout mechanism into our online photo album authoring client.

9.3.1 Photo Blogs

Personal blogs are a popular mean to document a person's life and for some have replaced the old paper-based diary. Especially interesting when considering digital photobooks, are blogs containing of many photos, such as travel blogs. The beauty of such blogs, from a technical perspective, is their inherent structured presence of text and images: Blogs consist of several entries, each consisting of a heading, one or more text blocks and images, possibly annotated with descriptions. Thus, one of the applications of our system is the automated transformation of these structured media sets into a digital photobook, e.g. allowing for turning the digitally documented memory to a nice holiday into a physical counterpart in the form of a digital photobook.

Figure 9.12 shows the result of such an automatic transformation of a travel blog documenting a trip through Canada and the USA into a digital photobook. Shown are four pages. On three pages a photo was chosen as the background and on the fourth page a matching, uniformly colored background was added. Headings from the original blog have resulted in headings and the entries' text and image contents have resulted in text and images on the pages laid out by our system.



Tagesordnungspunkte auf dem nahezu verlassenen Campingplatz: ausschlafen, lesen, Spaziergang am Fluß, lesen, schlafen, Hot Dogs essen, Tee trinken! Ach und Jens hatte endlich einen Basketball Platz und prellte stundenlang sehlig vor sich hin. Herrlich war das heute!

Boston: The Cradle of Liberty

02.10.2008

Yeah, ready to hit the road!

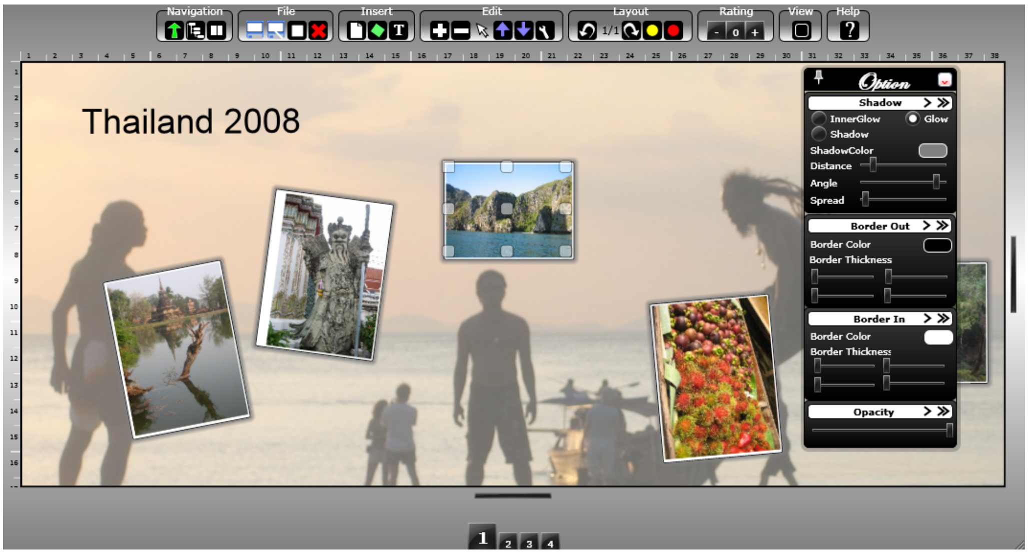
20.09.2008

Unser Camper ist super! Baujahr 2008, 4,5 Liter Maschine, Alufelgen, verchromter Külergrill, strahlernd weisser Pianolack ohne den kleinesten Kratzer. Und noch dazu: ein voll ausgestatteter Arbeitsplatz für die Frau von Welt: Mikrowelle, Gasherd, Backofen, Kaffemaschine etc. Noch schnell das Nötigste eingekauft machen wir uns 3 (!) Stunden später bei strahlendem Sonnenschein und 26 Grad auf den Weg. Heutiges Etappenziel: Honey Harbour bei Georgian Bay. Auf dem kleinen Campingplatz auf Picnic Island mußten wir natürlich erst mal unseren neuen Grill einweihen. Währenddessen erlebten wir den ersten Termperaturschock: hilfe kühlt das hier von super warm auf ganz kalt ab! Route

Figure 9.12: Example of two double-pages generated from a blog documenting a journey through Canada and the USA

9.3.2 Web-based Authoring Tool

We implemented our system for automatic album page layout in a web-based authoring application on the basis of Microsoft's Silverlight Technology. With this application, users are able to upload their photos to a server and to design a multi-page photo album from their photos. The photos can be freely distributed over the pages, rotated, resized, clipped and so on. As a means for additional decoration, rectangles and text boxes can be added to the pages and freely rotated and resized. All objects can be put into background or foreground and their opacity can be adjusted. Additionally, the borders of all objects can be adjusted regarding color, thickness and the amount of shadows. A screenshot of the application is depicted in Figure 9.13. Hence, the system provides a manual authoring environment for photo albums.



Thailand 2008

Figure 9.13: Web-based photo album authoring tool

We integrated our automatic layout approach into the system in two ways: By means of a wizard the user can chose a set of images from his photo library. He or she can specify the preferred number of pages and the page size. From this, a suggestion for a complete photobook layout is automatically generated. The user can though still alter the pages add text descriptions or add or delete photos. The second integration is much more embedded in the actual album design process. For this we assume, that the photos might have been manually distributed and placed on the pages and the background of the pages has been chosen. The user can now employ our system to receive different layout suggestions for the elements on the page (images and texts), thus the steps of content distribution and background layout are skipped. This is done by pressing a respective button in the application's user interface. Due to the nature of the genetic algorithm and some random decisions in the details of the layout step of the algorithm every time the user hits the button a slightly different layout is produced and offered. The user still has the choice and go back and forth trough the different layouts and any time.

9.4 Summary

In this chapter we presented an approach for automatic layout of photo compositions which incorporates the knowledge about aesthetic design principles. We identified and analyzed rules and principles of design in literature and from existing usage in professional photo album pages for their applicability in the domain of photo compositions. The discussed principles of layout and design were systematically integrated the design of our automatic layout system. We implemented our approach both in a web-based rich media application for the manual and automatic creation of photo compositions and a system for the automatic transformation of blogs into photo books.

With our automatic design aid end users are now able to create layouts that are appealing with very low effort and limited design skills. The users can quickly create one design after the other until the result meets the personal expectations of a nice design. The effectiveness of our system could be validated by an informal user study with 10 subjects which showed that the advantages of an automatic creation of appealing compositions from photos were appreciated by the users. The study supports the proposed kind of design support and shows that the resulting presentation are to the users' satisfaction.

Although primarily integrated in the automatic photobook design process our approach can be applied in the many cases in which a set of photos needs to be nicely arranged (and printed) such as for collage postcards, for posters, calendar pages and so on. In times of billions of photos captured every year which reside on the users' hard disks, our approach provides an important incentive to create, view, share, and give photos as appealing compositions in many forms and will push commercialization of digital photo services.

10 Conclusions

In this concluding chapter we summarize the research which has been presented in this thesis. Section 10.1 provides a summary and Section 10.2 lists the concrete results and scientific contribution of our work. The closes with an outlook to potential extensions to our system and potential future research topics which have not been addressed in this thesis in Section 10.3.

10.1 Summary

In this thesis we have described a method and system for the the automatic generation of digital photobooks. We have partly based this system on the analysis of thousands of real world photobooks and by analyzing the usage of photos is the context of digital photobooks. The method for photobook creation is split into the parts analysis, retrieval, augmentation, and layout. Though meant to be integrated into the overall photobook creation process they are also designed to be used in other applications in the context of personal photos.

In Chapter 3 a thorough analysis of existing work in the field is described which identifies works which relate to the research topics addressed in this thesis. Several works exist which deal with the topics of photo analysis, selection and photo collage layout in the context of personal photos. However, none of the works so far specifically deals with problems and questions which arise from creation of digital photobooks.

Our approach bases on the analysis of a large set of real-world photobooks which is described in Chapter 5 and the analysis of photo usage in the context of digital photobooks described in Chapter 4. The analysis of photobooks is done on thousands of real-world photobooks which have been ordered over the period of about one year at Europe's largest photo finishing company CEWE Color. We have analyzed both the structure of these photobooks as well as the semantic layer of these books. The usage analysis of photos in Chapter 4 starts by structuring and a general model for different forms of photo usage which leads to an abstract model for the capturing of usage information for digital photos. One form of this photo usage, watching a photo slideshow is further examined by performing two experiments on two distinct groups of people attending a slide show. In this experiment we analyzed the relation between the viewing time of photos in such a slide show and the later selection of photos from the same set for a photobook.

Chapter 6 introduces our framework MetaXa for the multimodal analysis of digital photos. Based on this framework a couple of components are developed and described to analyze and semantically enrich digital photos. The results of these analyses in form of semantically rich metadata acts as input and basis for the retrieval of photos and the augmentation and layout of digital photobooks described in the next chapters.

In Chapter 7 we describe our method for the multi-stage selection and grouping of photos from a potentially large set of photos. This approach combines the rating of events and single photos. The retrieval system uses and rates metadata attached to the photos by our photo analysis system and provides a set selected and grouped photos which acts as input to the photobook layout system.

Our photobook augmentation system is introduced in Chapter 8. Based on a set of rules and definitions for relevant sources the system interprets the metadata attached to the photos and to the photobook to gather additional content from web which smoothly fits into the resulting photobook. This system is motivated by the insight from the photobook analysis in Chapter 5 that people tend to augment their photobooks with photos from other people, text and other types of content, like geographical maps. The augmentation system is meant to be integrated either in a fully automatic process or in a semi-automatic way during the manual photobook authoring process.

Chapter 9 describes the rule-based photobook layout system which is implemented with the help of genetic algorithms. We identify and derive a couple of layout rules both from the literature and from the analysis of photobooks and implement these rules in our system. The system uses the grouping information from the photo analysis system to keep semantically related photos together in the final layout. It also uses the content analysis information of photos to decide, which photos can act as backgrounds, which parts can be covered and which photos should be placed more prominent than others. The layout system is designed to be dynamic and thus the final photobook page layout is not based on a pre-defined layout.

The different parts of our system have been implemented in couple of prototypes which can be seen in figures throughout this thesis.

10.2 Results and Contribution

Photobooks are one of the oldest and popular means to preserve ones personal history captured in photos. Interestingly their digital counterparts are still the most popular methods to preserve such memories for the future (the CEWE Photobook is the companies best-selling selling product [COL12]). This thesis is giving insight into the human photobook authoring behavior and provides means to assist the user in designing their photobooks by automating the relevant steps as far as possible. The central results can be summarized as follows:

Photobook authoring behavior

Several works exist on the analysis of photo related activities or photowork[KSRW06]. This thesis however specifically focusses on activities in the context of photobooks. To our knowledge we are the first to provide insights into photobook authoring behavior by having structurally and semantically analyzed thousands of real-wold photobooks. Furthermore we have analyzed the connection between photo usages such as slide watching and photo selection for photobooks and have shown that such usage information gives valuable hints for the importance of photos for the user and thus their importance for a photobook. To our knowledge we are the first to consider usage information as a source for the semantic analysis of photos and for this have provided a general photo usage metadata model which can act as the basis for a formal metadata format. [SB11b, SB11a, SBMB10, SB09, STB08]

Photo Analysis

Our multimodal photo analysis framework MetaXa intelligently combines information from the content and context of photos to derive information which is more semantically rich then it would be with content- or context-analysis alone. From a scientific perspective we have shown, that such a multimodal really works and have provided a couple of novel methods based on the MetaXa framework that follow this idea. From a technical perspective we have provided a software framework along with a rich set of photo analysis components which is easily extensible for, but not limited to the applications of digital photobook creation. [SB11a, SB09, BSST07, BSSW07]

Photo Retrieval

We have provided a method group and select photos, based on a rich set of semantic metadata. This method is specifically designed to be used for the creation of digital photobooks. In this method we combine information about the single photo as well the event this photo belongs to. From the usage analysis we found out that metadata attached to a single photo not necessarily provides semantic hints for this single photo but rather the event it belongs to. This result has driven the design of our novel photo selection and grouping system. [SB11a, SBF08]

Dynamic photobook layout

This thesis introduces a novel layout method for digital photobooks which easily can also be applied to other multimedia presentation applications. In contrast to other approaches it specifically takes into account the content of the photos, their importance and their relation to each other, which is reflected in the resulting layout. In contrast to other approaches it does not solely rely on pre-defined layouts but takes a simple set of general layout primitives together with carefully chosen set of layout rules to dynamically layout the individual pages. As outlined in Chapter 3 other approaches either follow a fully template-based approach or generate individual layouts which are not based on agreed-upon aesthetic principles for. In addition to this our approach specifically includes content besides photos, like texts or maps, into the layout process. [SREB11, SRB10]

10.3 Directions for future work

Throughout the thesis a couple of open research questions and challenges arose. In the following we will highlight the most promising ones and will give hints for potential approaches to solve them.

Social Network Integration

Multimedia in social networks is one of the current hot topics in research. With the rise of social networks also the human photo capturing and sharing behavior has significantly changed, also due to the increasing quality of cameras in mobile devices. Nowadays people are able to always capture important moments of their lifes and instantly share them with others. This also affects the kind of photos taken and the quality and amount of metadata which is attached to the single photo. In contrast to only basic photographic metadata automatically captured by the digital cameras, with photos captured and shared by smartphones it is comments, annotations, tags, and the sharing history of each single photo. This rich pool of semantic information has a high potential to be exploited for the selection and layout of digital photobooks. First research in this direction has already been started, e.g. in [RSB10] and [RSB11].

Usage Data

In Chapter 4 we have proposed a model for the simple integration of usage data into existing metadata formats like XMP. This model is general enough to also capture complex usages like e.g. sharing activities in social networks or the use in photobooks. However, more research could be done on investing which kinds of information are actually most important to contribute to a better semantic understanding of the photos and thus are worth to be captured. Also, there is the open question where this information should be stored. Should it be stored together with the photo in its header? Should it be stored at a central metadata storage? How can we deal with copies of photos? How can be ensured that usage metadata from these copies are preserved and synchronized for every other copy? Who owns the single photo and who is allowed to contribute to the usage history and also to view this history? These are only some open questions which arise when thinking of capturing and using the history of a photo and leads to open and challenging research topics as well and technical challenges.

Distributed photo collections

In this thesis we have proposed a set of systems which are meant for the application of photobook creation. For this, we have focussed mostly on the technical aspect and only barely touched the question to answer what kind of content from a semantic perspective summarizes an event. Several works exist on summarization of personal photo collections (See Section 3.3.3 for details), but very few consider the diversity and distribution of photos: Today personal photographs not only reside on the peoples hard drives anymore but are rather distributed over various devices, web platforms and people with different types and quality of metadata. This provides many new challenges and opportunities for the retrieval and layout of personal photos for photobooks.

Automatic Learning of retrieval parameters

In this thesis we have provided both a system for the retrieval of photos for photobooks. The result of this retrieval system can be tuned by various retrieval parameters which adjust the weights for the separate rating components. For our application we have developed a workbench to manually tune these parameters which can also be used to adjust the system to different applications. What however is missing is a way to reliably distinct between different use cases which require different selection parameter sets. On way to achieve this is to try to learn from existing manual photo selections and to learn such parameters depending on the photo sets these selections where retrieved from, e.g. with state-of-the-art machine learning techniques. Similar to the analysis of large sets of photobooks in Chapter 5, existing real-world photobooks could be used as training sets for this.

Figures

Bibliography

[ACM] ACM: *Multimedia Grand Challenge*. http://comminfo.rutgers.edu/conferences/mmchallenge/

[Ado05] ADOBE: XMP Specification / Adobe Systems Inc. San Jose, CA, September 2005 (1). – Technical Report

[AEN$^+$09] AMES, Morgan ; ECKLES, Dean ; NAAMAN, Mor ; SPASOJEVIC, Mirjana ; VAN HOUSE, Nancy: Requirements for mobile photoware. In: *Personal and Ubiquitous Computing* (2009)

[al.05] AL., Chabane D.: 3rd Special Workshop on Multimedia Semantics. Pisa, Italy, June 2005

[ALCV08] ARDIZZONE, E. ; LA CASCIA, M. ; VELLA, F.: Mean shift clustering for personal photo album organization. In: *Proc. of ICIP 2008. 15th IEEE International Conference on Image Processing*, 2008. – ISSN 1522–4880, P. 85 –88

[App06] APPLE INC.: *iPhoto*. 2006. – http://www.apple.com/de/ilife/iphoto/

[ASB11] *Chapter* Geospatial Web Image Mining. In: AHLERS, Dirk ; SANDHAUS, Philipp ; BOLL, Susanne: *Internet Multimedia Search and Mining*. Bentham Science, 2011

[ASS02] AVCIBAŞ, İ. ; SANKUR, B. ; SAYOOD, K.: Statistical evaluation of image quality measures. In: *Journal of Electronic Imaging* 11 (2002), P. 206

[Atk04] ATKINS, B.C.: Adaptive photo collection page layout. In: *Image Processing, 2004. ICIP '04. 2004 International Conference on* 5 (2004), Oct., P. 2897–2900 Vol. 5. – DOI 10.1109/ICIP.2004.1421718. – ISSN 1522–4880

[Atk08] ATKINS, C. B.: Blocked recursive image composition. In: *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–303–7, P. 821–824

[ATSC08] ARNI, T. ; TANG, J. ; SANDERSON, M. ; CLOUGH, P.: Creating a test collection to evaluate diversity in image retrieval. In: *Beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments* (2008)

[AXO08] ANGUERA, Xavier ; XU, JieJun ; OLIVER, Nuria: Multimodal photo annotation and retrieval on a mobile phone. In: *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–312–9, P. 188–194

[AY00] ASLANDOGAN, Y.A. ; YU, C.T.: Diogenes: A Web search agent for content based indexing of personal images. In: *Proceedings of ACM SIGIR*, 2000, P. 481–482

[AZP96] AIGRAIN, P. ; ZHANG, H.J. ; PETKOVIC, D.: Content-based representation and retrieval of visual media: A state-of-the-art review. In: *Representation and Retrieval of Visual Media in Multimedia Systems* (1996), P. 3–26

[BBL06] BOUTELL, M. ; BROWN, C.M. ; LUO, J.: Exploiting context for semantic scene classification. In: *Univ. Rochester, Rochester, NY, Tech. Rep* 894 (2006)

[BE04] BANERJEE, S. ; EVANS, B.L.: Unsupervised automation of photographic composition rules in digital still cameras. In: *Proceedings of SPIE* Bd. 5301, 2004, P. 364–373

[BE07] BOYD, Danah ; ELLISON, Nicole B.: Social network sites: Definition, history, and scholarship. In: *Journal of Computer-Mediated Communication* 13 (2007), Nr. 1

[BHW09] BALINSKY, Helen Y. ; HOWES, Jonathan R. ; WILEY, Anthony J.: Aesthetically-driven layout engine. In: *Proceedings of the DocEng '09*, ACM Press, 2009. – ISBN 978–1–60558–575–8, P. 119–122

[BL04a] BOUTELL, M. ; LUO, J.: Bayesian Fusion of Camera Metadata Cues in Semantic Scene Classification. In: *Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 2, 2004, P. 623–630

[BL04b] BOUTELL, Matthew ; LUO, Jiebo: Photo Classification by Integrating Image Content and Camera Metadata. In: *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*. Washington, DC, USA : IEEE Computer Society, 2004. – ISBN 0–7695–2128–2, P. 901–904

[BL05] BOUTELL, Matthew R. ; LUO, Jiebo: Beyond pixels: Exploiting camera metadata for photo classification. In: *Pattern Recognition* 38 (2005), Nr. 6, P. 935–946

[BLHL$^+$01] BERNERS-LEE, T. ; HENDLER, J. ; LASSILA, O. u. a.: The Semantic Web. In: *Scientific american* 284 (2001), Nr. 5, P. 28–37

[BMH06] BENTLEY, Frank ; METCALF, Crysta ; HARBOE, Gunnar: Personal vs. commercial content: the similarities between consumer use of photos and music. In: *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. New York, NY, USA : ACM, 2006. – ISBN 1–59593–372–7, P. 667–676

[BNG10] BECKER, Hila ; NAAMAN, Mor ; GRAVANO, Luis: Learning similarity metrics for event identification in social media. In: *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*. New York, NY, USA : ACM, 2010, P. 291–300

[BPS+05] BLOEHDORN, S. ; PETRIDIS, K. ; SAATHOFF, C. ; SIMOU, N. ; TZOUVARAS, V. ; AVRITHIS, Y. ; HANDSCHUH, S. ; KOMPATSIARIS, Y. ; STAAB, S. ; STRINTZIS, M.G.: Semantic annotation of images and videos for multimedia analysis. In: *The Semantic Web: Research and Applications* (2005), P. 592–607

[BPWK06] BURNETT, Ian S. ; PEREIRA, Fernando ; WALLE, Rik Van d. ; KOENEN, Rob: *The MPEG-21 Book*. John Wiley & Sons, 2006. – ISBN 0470010118

[Bry07] BRYANT, R.E.: Data-intensive supercomputing: The case for DISC. In: *School of Computer Science, Carnegie Mellon University, Tech. Rep. Technical Report CMU-CS-07-128* (2007)

[BSAT06] BOLL, Susanne ; SANDHAUS, Philipp ; ANSGAR, Scherp. ; THIEME, Sabine: Multimedia Information Retrieval aus der Perspektive eines Fotoalbums. In: *Datenbank Spektrum* 18 (2006), August, P. 33–40

[BSST07] BOLL, Susanne ; SANDHAUS, Philipp ; SCHERP, Ansgar ; THIEME, Sabine: MetaXa—Context- and Content-Driven Metadata Enhancement for Personal Photo Books. In: CHAM, Tat-Jen (Hrsg.) ; CAI, Jianfei (Hrsg.) ; DORAI, Chitra (Hrsg.) ; RAJAN, Deepu (Hrsg.) ; CHUA, Tat-Seng (Hrsg.) ; CHIA, Liang-Tien (Hrsg.): *Advances in Multimedia Modeling* Bd. 4351, Springer, Jan 2007 (LNCS). – ISBN 3–540–69421–8, P. 332–343

[BSSW07] BOLL, Susanne ; SANDHAUS, Philipp ; SCHERP, Ansgar ; WESTERMANN, Utz: Semantics, Content, and Structure of Many for the Creation of Personal Photo Albums. In: *ACM Multimedia*. Augsburg, Bavaria, Germany : ACM Press, September 2007. – ISBN 978–1–59593–702–5, P. 641

[CC10] CEWE COLOR, GfK-Panel S.: *Nutzungsverhalten Digitalfotografie 2010. Eine repräsentative Befragung von Digitalkamerabesitzern in deutschen Haushalten durch die GFK*. 2010

[CCK+06] CHEN, Jun-Cheng ; CHU, Wei-Ta ; KUO, Jin-Hau ; WENG, Chung-Yi ; WU, Ja-Ling: Tiling slideshow. In: *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*. New York, NY, USA : ACM Press, 2006. – ISBN 1–59593–447–2, P. 25–34

[CCS03] CUSANO, C. ; CIOCCA, G. ; SCHETTINI, R.: Image annotation using SVM. In: S. SANTINI & R. SCHETTINI (Hrsg.): *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2003, P. 330–338

[CF80] CHANG, Ning-San ; FU, King-Sun: Query-by-Pictorial-Example. In: *IEEE Trans. Softw. Eng.* 6 (1980), Nr. 6, P. 519–524. – DOI 10.1109/TSE.1980.230801. – ISSN 0098–5589

[CFGW05] COOPER, Matthew ; FOOTE, Jonathan ; GIRGENSOHN, Andreas ; WILCOX, Lynn: Temporal event clustering for digital photo collections. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 1 (2005), Nr. 3, P. 269–288. – DOI 10.1145/1083314.1083317. – ISSN 1551–6857

[Cha87] CHALFEN, Richard: *Snapshot Versions of Life*. Bowling Green University Popular Press, 1987

[CI10] CHANDRAMOULI, Krishna ; IZQUIERDO, Ebroul: Semantic structuring and retrieval of event chapters in social photo collections. In: *MIR '10: Proceedings of the international conference on Multimedia information retrieval*. New York, NY, USA : ACM, 2010. – ISBN 978–1–60558–815–5, P. 507–516

[CKL+07] CHU, C.T. ; KIM, S.K. ; LIN, Y.A. ; YU, Y.Y. ; BRADSKI, G. ; NG, A.Y. ; OLUKOTUN, K.: Map-reduce for machine learning on multicore. In: *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference* The MIT Press, 2007, P. 281

[CLH08] CAO, Liangliang ; LUO, Jiebo ; HUANG, Thomas S.: Annotating photo collections by label propagation according to multiple similarity cues. In: *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–303–7, P. 121–130

[COL12] COLOR, CEWE: *Fact Book*. `http://www.cewecolor.de`. Version: April 2012, Last checked: 16.04.2012

[Com99] COMITÉ INTERNATIONAL DES TÉLÉCOMMUNICATIONS DE PRESSE: IPTC - NAA Information Interchange Model Version 4. Version: July 1999. `http://www.iptc.org/IIM/`. 1999 (1). – Technical Report

[COSG+06] COHEN-OR, Daniel ; SORKINE, Olga ; GAL, Ran ; LEYVAND, Tommer ; XU, Ying-Qing: Color harmonization. In: *ACM SIGGRAPH 2006 Papers*. New York, NY, USA : ACM, 2006 (SIGGRAPH '06). – ISBN 1–59593–364–6, P. 624–630

[COT06] CHEN, Chufeng ; OAKES, Michael ; TAIT, John: Browsing Personal Images Using Episodic Memory (Time + Location). In: *Advances in Information Retrieval* 3936 (2006), P. 362–372

[CPSS04] CHIANESE, A. ; PICARIELLO, A. ; SANSONE, L. ; SAPINO, M. L.: Managing Uncertainties in Image Databases: A Fuzzy Approach. In: *Multimedia Tools Appl.* 23 (2004), Nr. 3, P. 237–252. – DOI 10.1023/B:MTAP.0000031759.22145.5d. – ISSN 1380–7501

[CRM04] CRABTREE, Andy ; RODDEN, Tom ; MARIANI, John: Collaborating around collections: informing the continued development of photoware. In: *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*. New York, NY, USA : ACM, 2004. – ISBN 1–58113–810–5, P. 396–405

[DBDFF06] DUYGULU, P. ; BARNARD, K. ; DE FREITAS, J. ; FORSYTH, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: *Computer Vision—ECCV 2002* (2006), P. 349–354

[DE05] DIAKOPOULOS, Nicholas ; ESSA, Irfan: Mediating photo collage authoring. In: *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*. New York, NY, USA : ACM, 2005. – ISBN 1–59593–271–2, P. 183–186

[DG08] DEAN, Jeffrey ; GHEMAWAT, Sanjay: MapReduce: simplified data processing on large clusters. In: *Commun. ACM* 51 (2008), Nr. 1, P. 107–113. – DOI 10.1145/1327452.1327492. – ISSN 0001–0782

[dig01] DIGITAL IMAGING GROUP: DIG35 Specification - Metadata for Digital Images – Version 1.1. 2001. – Technical Report

[DJLW06] DATTA, Ritendra ; JOSHI, Dhiraj ; LI, Jia ; WANG, James Z.: Studying Aesthetics in Photographic Images Using a Computational Approach. In: LEONARDIS, Ales (Hrsg.) ; BISCHOF, Horst (Hrsg.) ; PINZ, Axel (Hrsg.): *ECCV (3)* Bd. 3953, Springer, 2006 (Lecture Notes in Computer Science). – ISBN 3–540–33836–5, P. 288–301

[DLW07] DATTA, Rittendra ; LI, Jia ; WANG, James Z.: Learning the Consensus on Visual Quality for Next-Generation Image Management. In: *Proceedings of the ACM Multimedia Conference*, 2007

[DLW08] DATTA, R. ; LI, J. ; WANG, JZ: Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition. In: *15th IEEE International Conference on Image Processing*, 2008, P. 105–108

[DVKG$^+$04] DAMERA-VENKATA, Niranjan ; KITE, Thomas D. ; GEISLER, Wilson S. ; EVANS, Brian L. ; BOVIK, Alan C.: Image Quality Assessment Based on a Degradation Model. In: *IEEE Transactions on Image Processing* 9 (2004), apr, Nr. 4, P. 636–650. – DOI 10.1109/83.841940

[DWGW07] DATTA, R. ; W. GE, J. L. ; WANG, J. Z.: Image Retrieval: Ideas, Influences, and Trends of the new age. In: *ACM Computing Surveys, 2007* (2007)

[EDR06] EISENTHAL, Yael ; DROR, Gideon ; RUPPIN, Eytan: Facial Attractiveness: Beauty and the Machine. In: *Neural Comput.* 18 (2006), Nr. 1, P. 119–142. – DOI 10.1162/089976606774841602. – ISSN 0899–7667

[EF95] ESKICIOGLU, A.M. ; FISHER, P.S.: Image quality measures and their performance. In: *Communications, IEEE Transactions on* 43 (1995), Dec, Nr. 12, P. 2959–2965. – DOI 10.1109/26.477498. – ISSN 0090–6778

[ES03] ENSER, P. ; SANDOM, C.: Towards a comprehensive survey of the semantic gap in visual image retrieval. In: *Int. Conf. Image and Video Retrieval* 2728 (2003), P. 291–299

[Esk00] ESKICIOGLU, A.M.: Quality measurement for monochrome compressed images in the past 25 years. In: *IEEE International Conference on Acoustics Speech and Signal Processing* Bd. 4, 2000

[FKP+02] FROHLICH, David ; KUCHINSKY, Allan ; PERING, Celine ; DON, Abbe ; ARISS, Steven: Requirements for photoware. In: *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work.* New York, NY, USA : ACM Press, 2002. – ISBN 1581135602, P. 166–175

[FS09] FAGETH, Reiner ; SANDHAUS, Philipp: The Picture to Print Value Chain. In: *Proceedings of the International Symposium on Technologies for Digital Photo Fulfillment 2009 (TDPF '09).* Las Vegas, Nevada, USA, Feb 2009

[FSA96] FRANKEL, Charles ; SWAIN, Michael J. ; ATHITSOS, Vassilis: WebSeer: An Image Search Engine for the World Wide Web / University of Chicago. Chicago, IL, USA, 1996. – Technical Report

[FSN+95] FLICKNER, Myron ; SAWHNEY, Harpreet ; NIBLACK, Wayne ; ASHLEY, Jonathan ; HUANG, Qian ; DOM, Byron ; GORKANI, Monika ; HAFNER, Jim ; LEE, Denis ; PETKOVIC, Dragutin ; STEELE, David ; YANKER, Peter: Query by Image and Video Content: The QBIC System. In: *Computer* 28 (1995), Nr. 9, P. 23–32. – DOI 10.1109/2.410146. – ISSN 0018–9162

[FWK08] FROHLICH, D.M. ; WALL, S.A. ; KIDDLE, G.: Collaborative Photowork: Challenging the Boundaries Between Photowork and Phototalk. In: *Proc. of CHI Workshop on Collocated Social Practices Surrounding Photos*, 2008

[FXL+05] FAN, Xin ; XIE, Xing ; LI, Zhiwei ; LI, Mingjing ; MA, Wei-Ying: Photo-to-search: using multimodal queries to search the web from mobile devices. In: *MIR '05: Proceedings of the 7th ACM SIGMM inter-*

national workshop on Multimedia information retrieval. New York, NY, USA : ACM, 2005. – ISBN 1–59593–244–5, P. 143–150

[GAC+03] GIRGENSOHN, Andreas ; ADCOCK, John ; COOPER, Matthew D. ; FOOTE, Jonathan ; WILCOX, Lynn: Simplifying the Management of Large Photo Collections. In: *INTERACT*, 2003

[Gao09] GAO, Yuli et a.: MagicPhotobook: Designer Inspired, User Perfected Photo Albums. In: *Proc.of ACM MM*. Beijing, China : ACM Press, 2009. – ISBN 978–1–60558–608–3, P. 979–980

[GAW04] GIRGENSOHN, Andreas ; ADCOCK, John ; WILCOX, Lynn: Leveraging face recognition technology to find and organize photos. In: *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*. New York, NY, USA : ACM Press, 2004. – ISBN 1–58113–940–3, P. 99–106

[GC04] GIRGENSOHN, A. ; CHIU, P.: Stained glass photo collages. In: *Proc. ACM Symposium on User Interface Software and Technology* (2004), P. 13–14

[GDT02] GARGI, Ullas ; DENG, Yining ; TRETTER, Daniel R.: Managing and Searching Personal Photo Collections / HP Laboratories, Palo Alto. 2002. – Technical Report

[GGMPW02] GRAHAM, Adrian ; GARCIA-MOLINA, Hector ; PAEPCKE, Andreas ; WINOGRAD, Terry: Time as essence for photo browsing through personal digital libraries. In: *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA : ACM Press, 2002. – ISBN 1–58113–513–0, P. 326–335

[GJL+05] GURRIN, Cathal ; JONES, Gareth J. F. ; LEE, Hyowon ; O'HARE, Neil ; SMEATON, Alan F. ; MURPHY, Noel: Mobile access to personal digital photograph archives. In: *MobileHCI '05: Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. New York, NY, USA : ACM, 2005. – ISBN 1–59593–089–2, P. 311–314

[GL03] GEIGEL, Joe ; LOUI, Alexander: Using Genetic Algorithms for Album Page Layouts. In: *IEEE MultiMedia* 10 (2003), Nr. 4, P. 16–27. – DOI 10.1109/MMUL.2003.1237547. – ISSN 1070–986X

[Goo06] GOOGLE, INC.: *Picasa*. 2006. – http://picasa.google.com/

[GS90] GRILL, Tom ; SCANLON, Mark: *Photographic Composition*. Amphoto Books, 1990. – ISBN 9780817454272

[GUC05] GONG, Zhiguo ; U, Leong H. ; CHEANG, Chan W.: Web Image Semantic Clustering. In: *Proceedings of the 4th International Conference on Ontologies, Database and Applications of Semantics (ODBASE 2005).* Agia, Napa, Cyprus : Springer, 2005 (Lecture Notes in Computer Science)

[Ham92] HAMILTON, Eric: JPEG File Interchange Format Version 1.02 / C-Cube Microsystems. 1992. – Technical Report

[HCW$^+$07] HE, Xiaofei ; CAIA, Deng ; WEN, Ji-Rong ; MA, Wei-Ying ; ZHANG, Hong-Jiang: Clustering and Searching WWW Images Using Link and Page Layout Analysis. In: *ACM Trans. on Multimedia Comput. Commun. Appl. (TOMCAP)* 3 (2007), May, Nr. 10

[HLES07] HARE, Jonathon S. ; LEWIS, Paul H. ; ENSER, Peter G. B. ; SANDOM, Christine J.: Semantic Facets: An in-depth Analysis of a Semantic Image Retrieval System. In: *ACM CIVR 2007: The 6th International Conference on Image and Video Retrieval*, ACM, 2007

[HNO$^+$05] HARDMAN, Lynda ; NACK, Frank ; OBRENOVIC, Zeljko ; KERHERVE, Brigitte ; PIERSOL, Kurt: Canonical Processes of Media Production. In: *ACM Workshop on Multimedia for Human Communication - From Capture to Convey*, 2005

[HSL$^+$06] HARE, JS ; SINCLAIR, PAS ; LEWIS, PH ; MARTINEZ, K. ; ENSER, PGB ; SANDOM, CJ: Bridging the semantic gap in multimedia information retrieval: top-down and bottom-up approaches. In: *3rd European Semantic Web Conference*, 2006

[HSLN08] HARE, Jonathon S. ; SAMANGOOEI, Sina ; LEWIS, Paul H. ; NIXON, Mark S.: Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In: *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval.* New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–070–8, P. 359–368

[HSWW03] HOLLINK, L. ; SCHREIBER, A.T. ; WIELEMAKER, J. ; WIELINGA, B.: Semantic annotation of image collections. In: *Knowledge Capture*, 2003, P. 41–48

[Hun01] HUNTER, J.: Adding Multimedia to the Semantic Web building an Mpeg-7 Ontology. In: *International Semantic Web Working Symposium (SWWS)*, 2001

[HwSG$^+$05] HALASCHEK-WIENER, C. ; SCHAIN, A. ; GOLBECK, J. ; PARSIA, B. ; HENDLER, J.: A flexible approach for managing digital images on the semantic web. In: *th International Workshop on Knowledge Markup and Semantic Annotation*, 2005

[IKN98] ITTI, L. ; KOCH, C. ; NIEBUR, E.: A model of saliency-based visual attention for rapid scene analysis. In: *IEEE transactions on pattern analysis and machine intelligence* 20 (1998), Nr. 11, P. 1254–1259

[Ino09] INOUE, M.: Image retrieval: Research and use in the information explosion. In: *Progress in Informatics* 6 (2009), P. 3–14

[Int] INTEL CORP.: *Open Source Computer Vision Library (OpenCV)*. http://www.intel.com/technology/ computing/opencv/index.htm,

[ISSE02] IT STORAGE SYSTEMS, Technical Standardization C. a. ; EQUIPMENT: Exchangeable Image File Format for Digital Still Cameras: EXIF Version 2.2 / Japan Electronics and Information Technology Industries Association. 2002. – Technical Report

[Itt97] ITTEN, Johannes: *The Art of Color: The Subjective Experience and Objective Rationale of Color*. Wiley & Sons, 1997

[Jäh06] JÄHNE, Bernd: *Digital Image Processing*. 6. Springer, 2006

[JLM03] JEON, J. ; LAVRENKO, V. ; MANMATHA, R.: Automatic image annotation and retrieval using cross-media relevance models. In: *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA : ACM, 2003. – ISBN 1–58113–646–3, P. 119–126

[JNTD06] JAFFE, Alexander ; NAAMAN, Mor ; TASSA, Tamir ; DAVIS, Marc: Generating summaries for large collections of geo-referenced photographs. (2006), P. 853–854. – DOI 10.1145/1135777.1135911. ISBN 1–59593–323–9

[JYH08] JIA, Jimin ; YU, Nenghai ; HUA, Xian-Sheng: Annotating personal albums via web mining. In: *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–303–7, P. 459–468

[KKL07] KU, W. ; KANKANHALLI, M.S. ; LIM, J.H.: Using Camera Settings Templates to Classify Photos. In: *International Workshop on Advanced Image Technology*, 2007

[KN08] KENNEDY, Lyndon S. ; NAAMAN, Mor: Generating diverse and representative image search results for landmarks. In: *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–085–2, P. 297–306

[KNA$^+$07] KENNEDY, Lyndon ; NAAMAN, Mor ; AHERN, Shane ; NAIR, Rahul ; RATTENBURY, Tye: How flickr helps us make sense of the world: context and content in community-contributed media collections. In:

MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia. New York, NY, USA : ACM, 2007. – ISBN 978–1–59593–702–5, P. 631–640

[Kra05] KRAGES, Bert: *Photography: The Art of Composition*. Allworth Press, 2005. – ISBN 9781581154092

[KS05] KUSTANOWITZ, Jack ; SHNEIDERMAN, Ben: Meaningful presentations of photo libraries: rationale and applications of bi-level radial quantum layouts. In: *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA : ACM Press, 2005. – ISBN 1–58113–876–8, P. 188–196

[KSRW06] KIRK, David ; SELLEN, Abigail ; ROTHER, Carsten ; WOOD, Ken: Understanding photowork. In: *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. New York, NY, USA : ACM Press, 2006. – ISBN 1–59593–372–7, P. 761–770

[KTJ06] KE, Yan ; TANG, Xiaoou ; JING, Feng: The Design of High-Level Features for Photo Quality Assessment. In: *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA : IEEE Computer Society, 2006. – ISBN 0–7695–2597–0, P. 419–426

[KZB04] KHERFI, M. L. ; ZIOU, D. ; BERNARDI, A.: Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. In: *ACM Comput. Surv.* 36 (2004), Nr. 1, P. 35–67. – DOI 10.1145/1013208.1013210. – ISSN 0360–0300

[L$^+$02] LI, X. u. a.: Blind image quality assessment. In: *Proc. IEEE Int. Conf. Image Proc* Bd. 1, 2002, P. 449–452

[LB02] LAFON, Yves ; BOS, Bert: *Describing and retrieving photos using RDF and HTTP*. Version: 2002. http://www.w3.org/TR/photo-rdf/

[LBB06] LUO, Jiebo ; BOUTELL, M. ; BROWN, C.: Pictures are not taken in a vacuum. In: *IEEE Signal Processing Magazine* 23 (2006), Nr. 2, P. 101–114

[LC08] LUX, Mathias ; CHATZICHRISTOFIS, Savvas A.: Lire: lucene image retrieval: an extensible java CBIR library. In: *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–303–7, P. 1085–1088

[LDKT08] LINDLEY, Siân E. ; DURRANT, Abigail C. ; KIRK, David S. ; TAYLOR, Alex S.: Collocated social practices surrounding photos. In: *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*.

New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–012–X, P. 3921–3924

[Lew00] LEW, Michael S.: Next-Generation Web Searches for Visual Content. In: *IEEE Computer* 33 (2000), Nr. 11, P. 46–53. – DOI 10.1109/2.881694. – ISSN 0018–9162

[LFN04] LOK, Simon ; FEINER, Steven ; NGAI, Gary: Evaluation of visual balance for automated layout. In: *IUI '04: Proceedings of the 9th international conference on Intelligent user interface*. New York, NY, USA : ACM Press, 2004. – ISBN 1–58113–815–6, P. 101–108

[LFSBS08] LACERDA, Y.A. ; FIGUEIREDO, H.F. de ; SOUZA BAPTISTA, C. de ; SAMPAIO, M.C.: PhotoGeo: A Self-Organizing System for Personal Photo Collections, 2008, P. 258 –265

[LHB03] LIDWELL, William ; HOLDEN, Kritina ; BUTLER, Jill: *Universal Principles of Design*. Rockport Publishers, 2003. – ISBN 1592530079

[LMJ04] LAVRENKO, V. ; MANMATHA, R. ; JEON, J.: A model for learning the semantics of pictures. In: *Advances in Neural Information Processing Systems* 16 (2004)

[Lou00] LOUI, Alexander C.: Automatic Image Event Segmentation and Quality Screening for Albuming Applications. In: *ICME*, 2000, P. 1125–1128

[LS03] LOUI, Alexander C. ; SAVAKIS, Andreas E.: Automated Event Clustering and Quality Screening of Consumer Pictures for Digital Albuming. In: *IEEE Transactions on Multimedia* 5 (2003), September, Nr. 3, P. 390–402

[LSDJ06] LEW, Michael S. ; SEBE, Nicu ; DJERABA, Chabane ; JAIN, Ramesh: Content-based multimedia information retrieval: State of the art and challenges. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 2 (2006), February, Nr. 1, P. 1–19. – DOI 10.1145/1126004.1126005. – ISSN 1551–6857

[Lux09] LUX, Mathias: Caliph & Emir: MPEG-7 photo annotation and retrieval. In: *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*. New York, NY, USA : ACM, 2009. – ISBN 978–1–60558–608–3, P. 925–926

[LW99] LU, G. ; WILLIAMS, B: An Integrated WWW image retrieval system. In: *Proceedings of the AusWeb99*. Lismore, Australia, 1999

[LW08] LI, Jia ; WANG, James Z.: Real-time computerized annotation of pictures. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (2008), Nr. 6, P. 985–1002

[LZLM07] LIU, Ying ; ZHANG, Dengsheng ; LU, Guojun ; MA, Wei-Ying: A survey of content-based image retrieval with high-level semantics. In: *Pattern Recogn.* 40 (2007), Nr. 1, P. 262–282. – DOI 10.1016/j.patcog.2006.04.045. – ISSN 0031–3203

[ME07] MILLER, Andrew D. ; EDWARDS, W. K.: Give and take: a study of consumer photo-sharing culture and practice. In: *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA : ACM, 2007. – ISBN 978–1–59593–593–9, P. 347–356

[Met09] METADATA WORKING GROUP: Guidelines for Handling Image Metadata – Version 1.0.1 / Metadata Working Group. 2009. – Technical Report

[MHMK09] MANSOOR, Atif B. ; HAIDER, Maaz ; MIAN, Ajmal S. ; KHAN, Shoab A.: A Hybrid Image Quality Measure for Automatic Image Quality Assessment. In: *SCIA '09: Proceedings of the 16th Scandinavian Conference on Image Analysis*. Berlin, Heidelberg : Springer-Verlag, 2009. – ISBN 978–3–642–02229–6, P. 91–98

[MHN06] MAEKAWA, Takuya ; HARA, Takahiro ; NISHIO, Shojiro: Image classification for mobile web browsing. In: *WWW '06: Proc.of the 15th Intl. Conf. on World Wide Web*. New York, NY, USA : ACM Press, 2006. – ISBN 1–59593–323–9, P. 43–52

[Mil95] MILLER, G.A.: WordNet: a lexical database for English. In: *Communications of the ACM* 38 (1995), Nr. 11, P. 41

[ML03] MULHEM, Philippe ; LIM, Joo-Hwee: Home Photo Retrieval: Time Matters. In: *Proc. of the Second International Conference on Image and Video Retrieval (CIVR)*. Urbana-Champaign, IL, USA : Springer, LNCS, July 24-25, 2003, P. 321–330

[MMB05] MARCHAND-MAILLET, S. ; BERETTA, G.: *The Benchathlon Network*. `http://www.benchathlon.net`. Version: 2005

[MMMM$^+$03] MÜLLER, H. ; MÜLLER, W. ; MARCHAND-MAILLET, S. ; PUN, T. ; SQUIRE, D.M.G.: A framework for benchmarking in CBIR. In: *Multimedia Tools and Applications* 21 (2003), Nr. 1, P. 55–73

[MMMP02] MÜLLER, Henning ; MARCHAND-MAILLET, Stéphane ; PUN, Thierry: The Truth about Corel - Evaluation in Image Retrieval. In: *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*. London, UK : Springer-Verlag, 2002. – ISBN 3–540–43899–8, P. 38–49

[MMS+01] Müller, Henning ; Müller, Wolfgang ; Squire, David M. ; Marchand-Maillet, Stéphane ; Pun, Thierry: Performance evaluation in content-based image retrieval: overview and proposals. In: *Pattern Recogn. Lett.* 22 (2001), Nr. 5, P. 593–601. – DOI 10.1016/S0167–8655(00)00118–5. – ISSN 0167–8655

[MMW06] Marchand-Maillet, Stéphane ; Worring, Marcel: Benchmarking image and video retrieval: an overview. In: *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval.* New York, NY, USA : ACM, 2006. – ISBN 1–59593–495–2, P. 297–300

[MO06] Monaghan, Fergal ; O'Sullivan, David: Automating Photo Annotation using Services and Ontologies. In: *Mobile Data Management, IEEE International Conference on* (2006), P. 79–79. – DOI 10.1109/MDM.2006.39. – ISSN 1551–6245

[MO07] Monaghan, Fergal ; O'Sullivan, David: Leveraging Ontologies, Context and Social Networks to Automate Photo Annotation. In: *SAMT*, 2007, P. 252–255

[MSS02] Manjunath, B. S. ; Salembier, Philippe ; Sikora, Thomas: *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons, 2002. – ISBN 0471486787

[MT01] Munson, Ethan V. ; Tsymbalenko, Yelena: To Search for Images on the Web, Look at the Text, Then Look at the Images. In: *Proceedings of the First International Workshop on Web Document Analysis (WDA2001)*, 2001

[MTO99] Mori, Y. ; Takahashi, H. ; Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: *Proceedings of the First Interna- tional Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, 1999

[MWH+06] Mei, T. ; Wang, B. ; Hua, X.S. ; Zhou, H.Q. ; Li, S.: Probabilistic multimodality fusion for event based home photo clustering. In: *IEEE International Conference on Multimedia and Expo*, 2006, P. 1757–1760

[NGP08] Negoescu, Radu A. ; Gatica-Perez, Daniel: Analyzing Flickr groups. In: *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval.* New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–070–8, P. 417–426

[NH02] Nack, Frank ; Hardman, Lynda: Towards a Syntax for Multimedia Semantics / Centrum voor Wiskunde en Informatica (CWI). Amsterdam, 2002. – Technical Report

[NHWP04] NAAMAN, Mor ; HARADA, Susumu ; WANG, QianYing ; PAEPCKE, Andreas: Adventures in Space and Time: Browsing Personal Collections of Geo-Referenced Digital Photographs / Stanford InfoLab. Stanford, 2004 (2004-26). – Technical Report

[NMH03] NACK, Frank ; MANNIESING, Amit ; HARDMAN, Lynda: Colour picking: the pecking order of form and function. In: *ACM MM '03*, ACM Press, 2003, P. 279–282

[NNY09] NOV, O. ; NAAMAN, M. ; YE, C.: Motivational, Structural and Tenure Factors that Impact Online Community Photo Sharing. In: *Third International Conference on Weblogs and Social Media (ICWSM)*. San Jose, California, May 2009

[NRD05] NAIR, Rahul ; REID, Nick ; DAVIS, Marc: Photo LOI: browsing multiuser photo collections. In: *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*. New York, NY, USA : ACM, 2005. – ISBN 1–59593–044–2, P. 223–224

[NSPGM04] NAAMAN, Mor ; SONG, Yee J. ; PAEPCKE, Andreas ; GARCIA-MOLINA, Hector: Automatic organization for digital photographs with geographic coordinates. In: *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA : ACM Press, 2004. – ISBN 1–58113–832–6, P. 53–62

[OAOO09] OBRADOR, Pere ; ANGUERA, Xavier ; OLIVEIRA, Rodrigo de ; OLIVER, Nuria: The role of tags and image aesthetics in social image search. In: *WSM '09: Proceedings of the first SIGMM workshop on Social media*. New York, NY, USA : ACM, 2009. – ISBN 978–1–60558–759–2, P. 65–72

[OBMCP00] ORTEGA-BINDERBERGER, Michael ; MEHROTRA, Sharad ; CHAKRABARTI, Kaushik ; PORKAEW, Kriengkrai: *WebMARS: A Multimedia Search Engine*. 2000

[OGL$^+$05] O'HARE, Neil ; GURRIN, Cathal ; LEE, Hyowon ; MURPHY, Noel ; SMEATON, Alan F. ; JONES, Gareth J.: My digital photos: where and when? In: *Proceedings of the 13th annual ACM International Conference on Multimedia (MM)*. New York, NY, USA : ACM Press, 2005. – ISBN 1–59593–044–2, P. 261–262

[OT02] OLIVA, A. ; TORRALBA, A.: Scene-centered description from spatial envelope properties. In: *Biologically Motivated Computer Vision* 2525/2010 (2002), P. 263–272

[PCF02] PLATT, John C. ; CZERWINSKI, Mary ; FIELD, Brent A.: PhotoTOC: Automatic Clustering for Browsing Personal Photographs / Microsoft Research. 2002 (MSR-TR-2002-17). – Technical Report

[Pet94] PETERSEN, T.: *Art and architecture thesaurus*. New York: Oxford, 1994

[PG05] PIGEAU, A. ; GELGON, M.: Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices. In: *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*. New York, NY, USA : ACM, 2005. – ISBN 1–59593–044–2, P. 141–150

[Pig10] PIGEAU, Antoine: MyOwnLife: incremental and hierarchical classification of a personal image collection on mobile devices. In: *Multimedia Tools Appl.* 46 (2010), Nr. 2-3, P. 289–306. – DOI 10.1007/s11042–009–0373–x. – ISSN 1380–7501

[Pla00] PLATT, J.: *AutoAlbum: Clustering Digital Photographs Using Probabalistic Model Merging*. 2000

[PPT07] PENTA, Antonio ; PICARIELLO, Antonio ; TANCA, Letizia: Towards a definition of an Image Ontology. In: *International Workshop on Database and Expert Systems Applications* 0 (2007), P. 74–78. – DOI 10.1109/DEXA.2007.83. – ISSN 1529–4188

[PW09] PENG WU, Dan T.: Close & Closer: Social Cluster and Closeness from Photo Collections. In: *Proceedings of MM'09*, 2009, P. 709

[RBHB06] ROTHER, Carsten ; BORDEAUX, Lucas ; HAMADI, Youssef ; BLAKE, Andrew: AutoCollage. In: *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*. New York, NY, USA : ACM, 2006. – ISBN 1–59593–364–6, P. 847–852

[RC09] REN, Kan ; CALIC, Janko: FreeEye - Interactive Intuitive Interface for Large-scale Image Browsing. In: *ACM Multimedia*, 2009, P. 757

[RSB10] RABBATH, Mohammad ; SANDHAUS, Philipp ; BOLL, Susanne: Automatic Creation of Printable Photo Books out of your Stories in Social Networks. In: *2nd SIGMM Workshop on Social Media (WSM2010)*. Florenze, Italy, Oct 2010

[RSB11] RABBATH, Mohammad ; SANDHAUS, Philipp ; BOLL, Susanne: Semantic Photo Books: Leveraging Blogs and Social Media for Photo Book Creation. In: *Proc. of SPIE/ Electronic Imaging 2011-Imaging and Printing in a Web 2.0 World*, 2011

[RW03] RODDEN, Kerry ; WOOD, Kenneth R.: How do People manage their Digital Photographs? In: GILBERT COCKTON, Panu K. (Hrsg.): *Proceedings of the Conference on Human Factors and Computing Systems, ACM*, 2003, P. 409–416

[San05] SANDHAUS, Philipp: *Entwicklung eines komplementären, kameraunterstützten Sensorsystems für Lageregelungsaufgaben von Kleinstflugzeugen*. Oldenburg, Germany, Universität Oldenburg, Diplomarbeit, Oct. 2005

[SB05] SHIRAHATTI, Nikhil V. ; BARNARD, Kobus: Evaluating Image Retrieval. In: *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*. Washington, DC, USA : IEEE Computer Society, 2005. – ISBN 0–7695–2372–2, P. 955–961

[SB09] SANDHAUS, Philipp ; BOLL, Susanne: From usage to annotation: Analysis of personal photo albums for semantic photo understanding. In: *Proceedings of the First SIGMM Workshop on Social Media co-located with ACM Multimedia conference*. Bejing, China, October 2009

[SB11a] SANDHAUS, Philipp ; BOLL, Susanne: Semantic analysis and retrieval in personal and social photo collections. In: *Multimedia Tools and Applications* 51 (2011), P. 5–33. – ISSN 1380–7501. – 10.1007/s11042-010-0673-1

[SB11b] *Chapter* Social Aspects of Photobooks: Improving Photobook Authoring from Large-scale Multimedia Analysis. In: SANDHAUS, Philipp ; BOLL, Susanne: *Social Media Modeling and Computing*. Springer, 2011

[SBC05] SHEIKH, H.R. ; BOVIK, A.C. ; CORMACK, L.: No-reference quality assessment using natural scene statistics: JPEG2000. In: *Image Processing, IEEE Transactions on* 14 (2005), Nr. 11, P. 1918 –1927. – DOI 10.1109/TIP.2005.854492. – ISSN 1057–7149

[SBC06] SCHERP, Angar ; BOLL, Susanne ; CREMER, Holger: Emergent Semantics in Personalized Multimedia Content. In: *Fourth special workshop on Multimedia Semantics (WMS)*. Chania, Greece, 2006

[SBF08] SANDHAUS, Philipp ; BOLL, Susanne ; FAGETH, Reiner: Employing a photo's life cycle for multimedia retrieval. In: *MS '08: Proceeding of the 2nd ACM workshop on Multimedia semantics*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–316–7, P. 56–59

[SBMB10] SANDHAUS, Philipp ; BAUMGARTNER, Hannah ; MEYER, Jochen ; BOLL, Susanne: That was my life - Creating Personal Chronicles at the End of Life. In: *HCI at the End of Life - Workshop at CHI 2010*. Atlanta, GA, USA, April 2010

[SEL00] SAVAKIS, A.E. ; ETZ, S.P. ; LOUI, A.C.: Evaluation of image appeal in consumer photography. In: *Proc. SPIE Human Vision and Electronic Imaging*, 2000. – ISSN 0361–0748, P. 111–121

[SGJ01] SANTINI, Simone ; GUPTA, Amarnath ; JAIN, Ramesh: Emergent Semantics through Interaction in Image Databases. In: *IEEE Trans. on Knowl. and Data Eng.* 13 (2001), Nr. 3, P. 337–351. – DOI 10.1109/69.929893. – ISSN 1041–4347

[SGM02] SZYPERSKI, C. ; GRUNTZ, D. ; MURER, S.: *Component Software : Beyond Object-Oriented Programming*. 2nd. Addison Wesley, 2002. – ISBN 0–201–74572–0

[SJ07] SCHERP, Ansgar ; JAIN, Ramesh: Towards an ecosystem for semantics. In: *MS '07: Workshop on multimedia information retrieval on The many faces of multimedia semantics*. New York, NY, USA : ACM Press, 2007. – ISBN 978–1–59593–782–7, P. 3–12

[SJ08a] SINHA, Pinaki ; JAIN, Ramesh: Classification and annotation of digital photos using optical context data. In: *CIVR*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–070–8, P. 309–318

[SJ08b] SINHA, Pinaki ; JAIN, Ramesh: Semantics In Digital Photos: A Contenxtual Analysis. In: *ICSC '08: Proceedings of the 2008 IEEE International Conference on Semantic Computing*. Washington, DC, USA : IEEE Computer Society, 2008. – ISBN 978–0–7695–3279–0, P. 58–65

[SLC+02] SEBE, N. ; LEW, M.S. ; COHEN, I. ; GARG, A. ; HUANG, T.S.: Emotion recognition using a Cauchy Naive Bayes classifier. In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on* 1 (2002), P. 17–20 vol.1. – DOI 10.1109/ICPR.2002.1044578. – ISSN 1051–4651

[SLCS99] SCLAROFF, Stan ; LA CASCIA, Marco ; SETHI, Saratendu: Unifying textual and visual cues for content-based image retrieval on the World Wide Web. In: *Comput. Vis. Image Underst.* 75 (1999), Nr. 1-2, P. 86–98. – DOI 10.1006/cviu.1999.0765. – ISSN 1077–3142

[SLH06] SHADBOLT, N. ; LEE, Tim B. ; HALL, W.: The Semantic Web Revisited. In: *IEEE Intelligent Systems* Bd. 21, 2006, P. 96–101

[SNN+08] SCHERP, Ansgar ; NACK, Frank ; NAHRSTEDT, Klara ; INOUE, Masashi ; GIRGENSOHN, Andreas ; HENRICH, Andreas ; SANDHAUS, Philipp ; THIEME, Sabine ; ZHOU, Michelle: Interaction and user experiences with multimedia technologies: challenges and future topics. In: *HCC '08: Proceeding of the 3rd ACM international workshop on Human-centered computing*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–320–4, P. 1–6

[SRB10] SANDHAUS, Philipp ; RABBATH, Mohammad ; BOLL, Susanne: Blog2Book - Transforming Blogs into Photo Books Employing Aesthetic Principles. In: *ACM Multimedia 2010 - Technical Demonstrations*. Florenze, Italy, Oct 2010

[SREB11] SANDHAUS, Philipp ; RABBATH, Mohamad ; ERBIS, Ilja ; BOLL, Susanne: Employing Aesthetic Principles for Automatic Photo Book Layout. In: *Proc. of International Conference on Multimedia Modelling*. Taipei, Taiwan, January 2011

[STB08] SANDHAUS, Philipp ; THIEME, Sabine ; BOLL, Susanne: Processes of photo book production. In: *Special Issue of Multimedia Systems Journal on Canonical Processes of Media Production* 14 (2008), Nr. 16, P. 351–357. – DOI 10.1007/s00530–008–0136–y

[SWS+00] SMEULDERS, Arnold W. ; WORRING, Marcel ; SANTINI, Simone ; GUPTA, Amarnath ; JAIN, Ramesh: Content-Based Image Retrieval at the End of the Early Years. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000), Nr. 12, P. 1349–1380. – DOI 10.1109/34.895972. – ISSN 0162–8828

[SYJL09] SUN, Xiaoshuai ; YAO, Hongxun ; JI, Rongrong ; LIU, Shaohui: Photo Assessment Based on Computational Visual Attention Model. In: *Proceedings of MM'09*, 2009, P. 541

[TCMK02] TAN, Tele ; CHEN, Jiayi ; MULHEM, Philippe ; KANKANHALLI, Mohan: SmartAlbum: a multi-modal photo annotation system. In: *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*. New York, NY, USA : ACM, 2002. – ISBN 1–58113–620–X, P. 87–88

[TGO+06] TUFFIELD, Mischa M. ; GIBBINS, Nicholas ; O'HARA, Kieron ; BREWSTER, Christopher ; DUPPLAW, David P. ; WILKS, Yorick ; SLEEMAN, Derek ; SHADBOLT, Nigel R.: Image annotation with Photocopain. In: *International Workshop on Semantic Web Annotations for Multimedia*, 2006

[TLZ+04] TONG, Hanghang ; LI, Mingjing ; ZHANG, HongJiang ; HE, Jingrui ; ZHANG, Changshui: Classification of Digital Photos Taken by Photographers or Home Users. In: AIZAWA, Kiyoharu (Hrsg.) ; NAKAMURA, Yuichi (Hrsg.) ; SATOH, Shin'ichi (Hrsg.): *PCM (1)* Bd. 3331, Springer, 2004 (Lecture Notes in Computer Science). – ISBN 3–540–23974–X, P. 198–205

[TM01] TSYMBALENKO, Yelena ; MUNSON, Ethan V.: Using HTML Metadata to Find Relevant Images on the World Wide Web. In: *Proceedings of Internet Computing 2001* Bd. 2. Las Vegas : CSREA Press, 2001, P. 842–848

[TOPS07] TRONCY, Raphael ; OSSENBRUGGEN, Jacco van ; PAN, Jeff Z. ; STAMOU, Giorgos: Image Annotation on the Semantic Web / W3C. Version: 2007. `http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/`. 2007. – W3C Note

[TP91] TURK, Matthew ; PENTLAND, Alex: Eigenfaces for recognition. In: *J. Cognitive Neuroscience* 3 (1991), Nr. 1, P. 71–86. – DOI 10.1162/jocn.1991.3.1.71. – ISSN 0898–929X

[Tru00] TRUST, G.: *Union list of artists names online.* 2000

[VFG$^+$07] VIANA, Windson ; FILHO, José B. ; GENSEL, Jérôme ; OLIVER, Marlène V. ; MARTIN, Hervé: PhotoMap - automatic spatiotemporal annotation for mobile photos. In: *W2GIS'07: Proceedings of the 7th international conference on Web and wireless geographical information systems.* Berlin, Heidelberg : Springer-Verlag, 2007. – ISBN 3–540–76923–4, 978–3–540–76923–1, P. 187–201

[VFJZ99] VAILAYA, Aditya ; FIGUEIREDO, M. ; JAIN, Anil K. ; ZHANG, Ji: Content-Based Hierarchical Classification of Vacation Images / Department of Computer Science, Michigan State University. East Lansing, Michigan, February 1999 (MSU-CPS-99-9). – Technical Report. – 16 P.

[VHDA$^+$05] VAN HOUSE, Nancy ; DAVIS, Marc ; AMES, Morgan ; FINN, Megan ; VISWANATHAN, Vijay: The uses of personal networked digital imaging: an empirical study of cameraphone photos and sharing. In: *CHI '05: CHI '05 extended abstracts on Human factors in computing systems.* New York, NY, USA : ACM, 2005. – ISBN 1595930027, P. 1853–1856

[VHDT$^+$04] VAN HOUSE, N.A. ; DAVIS, M. ; TAKHTEYEV, Y. ; AMES, M. ; FINN, M.: The social uses of personal photography: Methods for projecting future imaging applications. In: *University of California, Berkeley, Working Papers* 3 (2004), P. 2005

[VJ01] VIOLA, Paul ; JONES, Michael: Rapid Object Detection using a Boosted Cascade of Simple Features. (2001). `citeseer.ist.psu.edu/article/viola01rapid.html`

[Wag86] WAGENAAR, Willem A.: My memory: A study of autobiographical memory over six years. In: *Cognitive Psychology* 18 (1986), Nr. 2, P. 225 – 252. – DOI 10.1016/0010–0285(86)90013–7. – ISSN 0010–0285

[WBL02] WANG, Z. ; BOVIK, A.C. ; LU, L.: Why is image quality assessment so difficult? In: *IEEE International Conference on Acoustics Speech and Signal Processing* Bd. 4 IEEE, 2002, P. 3313–3316

[WCTV85] WAAL, H. ; COUPRIE, LD ; THOLEN, E. ; VELLEKOOP, G.: *Iconclass: an iconographic classification system.* North-Holland Pub. Co., 1985

[WQS$^+$06] WANG, Jingdong ; QUAN, Long ; SUN, Jian ; TANG, Xiaoou ; SHUM, Heung-Yeung: Picture Collage. In: *cvpr* 1 (2006), P. 347–354. – DOI 10.1109/CVPR.2006.224. – ISSN 1063–6919

[WSZ00] WENYIN, Liu ; SUN, Yanfeng ; ZHANG, Hongjiang: MiAlbum - a system for home photo managemet using the semi-automatic image annotation approach. In: *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*. New York, NY, USA : ACM, 2000. – ISBN 1–58113–198–4, P. 479–480

[WZ04] WANG, Y. M. ; ZHANG, H.: Detecting image orientation based on low-level visual content. In: *Computer Vision and Image Understanding (CVIU)* 3 (2004), Nr. 3, P. 328–346

[XZC+08] XIAO, Jun ; ZHANG, Xuemei ; CHEATLE, Phil ; GAO, Yuli ; ATKINS, C. B.: Mixed-initiative photo collage authoring. In: *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–303–7, P. 509–518

[YFM+09] YAN, Rong ; FLEURY, Marc-Olivier ; MERLER, Michele ; NATSEV, Apostol ; SMITH, John R.: Large-scale multimedia semantic concept modeling using robust subspace bagging and MapReduce. In: *LS-MMRM '09: Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining*. New York, NY, USA : ACM, 2009. – ISBN 978–1–60558–756–1, P. 35–42

[YHBO10] YEH, Che-Hua ; HO, Yuan-Chen ; BARSKY, Brian A. ; OUHYOUNG, Ming: Personalized photograph ranking and selection system. In: *Proc. of the 18th International Conference on Multimedia*, ACM, Oct. 2010, P. 211–220

[YKA02] YANG, Ming-Hsuan ; KRIEGMAN, David J. ; AHUJA, Narendra: Detecting Faces in Images: A Survey. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (2002), Nr. 1, P. 34–58. – DOI 10.1109/34.982883. – ISSN 0162–8828

[YPHG09] YOU, Junyong ; PERKIS, Andrew ; HANNUKSELA, Miska M. ; GABBOUJ, Moncef: Perceptual Quality Assessment Based on Visual Attention Analysis. In: *Proceedings of MM'09*, 2009, P. 561

[YSR05] YAVLINSKY, A. ; SCHOFIELD, E. ; RÜGER, S.: Automated image annotation using global features and robust nonparametric density estimation. In: *Image and Video Retrieval* (2005), P. 507–517

[ZCPB09] ZAGORIS, Konstantinos ; CHATZICHRISTOFIS, Savvas A. ; PAPAMARKOS, Nikos ; BOUTALIS, Yiannis S.: img(Anaktisi): A Web Content Based Image Retrieval System. In: *SISAP '09: Proceedings of the 2009 Second International Workshop on Similarity Search and Applications*. Washington, DC, USA : IEEE Computer Society, 2009. – ISBN 978–0–7695–3765–8, P. 154–155

[ZCPR03] ZHAO, W. ; CHELLAPPA, R. ; PHILLIPS, P. J. ; ROSENFELD, A.: Face recognition: A literature survey. In: *ACM Comput. Surv.* 35 (2003), Nr. 4, P. 399–458. – DOI 10.1145/954339.954342. – ISSN 0360–0300

[ZCWT10] ZHANG, T. ; CHAO, H. ; WILLIS, C. ; TRETTER, D.: Consumer Image Retrieval by Estimating Relation Tree From Family Photo Collections / HP Laboratories. 2010. – Technical Report

[ZKS08] ZHAO, J. ; KLYNE, G. ; SHOTTON, D.: Building a Semantic Web image repository for biological research images. In: *The Semantic Web: Research and Applications* (2008), P. 154–169

[ZLL+09] ZHU, Changyun ; LI, Kun ; LV, Qin ; SHANG, Li ; DICK, Robert P.: iScope: personalized multi-modality image search for mobile devices. In: *MobiSys '09: Proceedings of the 7th international conference on Mobile systems, applications, and services*. New York, NY, USA : ACM, 2009. – ISBN 978–1–60558–566–6, P. 277–290

[ZTL+06] ZHAO, Ming ; TEO, Yong ; LIU, Siliang ; CHUA, Tat-Seng ; JAIN, Ramesh: Automatic Person Annotation of Family Photo Album. In: *CIVR*, 2006, P. 163–172