Universitat Politècnica de Catalunya
Barcelona

Media Aesthetics Based Multimedia Storytelling

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Signal Theory and Communications

by

Pere Obrador

2011

ACTA DE QUALIFICACIÓ DE LA TESI DOCTORAL

Reunit el tribunal integrat pels sota signants per jutjar la tesi doctoral:

Títol de la tesi: ....
Autor de la tesi: ....

Acorda atorgar la qualificació de:

□ No apte
□ Aprovat
□ Notable
□ Excel·lent
□ Excel·lent Cum Laude

Barcelona, …………… de/d'…..................…………….. de ....….

El President

....
(nom i cognoms)

El Secretari

....
(nom i cognoms)

El vocal

....
(nom i cognoms)

El vocal

....
(nom i cognoms)

El vocal

....
(nom i cognoms)

To my parents . . .
who—among so many other things—
saw to it that I mastered both English and touch-typing
while I was still in elementary school.
. . .
. . .
To my wife, Malén, and children, Jan Pol and Aidan Tau . . .
for their infinite patience.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to thank Josep Ramon Casas, my PhD. advisor, for allowing me to keep on working on my PhD. for such a long time, while I have been enrolled in many different professional endeavors.

I would like to acknowledge the *"Sa Nostra" Caixa de Balears*, for their scholarship which helped me while I was a Visiting Scholar at the *University of Southern California* in 1994. I would also like to thank Llorenç Huguet at *Universitat de les Illes Balears* for his mentorship during this time.

I would like to acknowledge C.-C. Jay Kuo at the *University of Southern California* for his direction in 1994 and 1995, and for his financial support in the form of a Research Assistantship in 1995. I would like to thank Ferran Marquès at *Universitat Politècnica de Catalunya* for helping me contact C.-C. Jay Kuo in the first place.

I would like to thank all my co-authors, in both academic publications, as well as in patents and patent applications for their help in pushing these technologies to a new level. Specially, I would like to thank Nathan Moroney for sharing my vision and allowing me to use some of his great technology, as well as providing some initial *aesthetics* ground truth data obtained via his *photo harvesting* system. Special thanks, also, to David Berfanger for helping me run early user studies.

Thanks to all the *Hewlett-Packard* employees that volunteered their photo collections for the photo book user study presented in Chapter 13. Also thanks to all the people that volunteered their photo collections for the user study presented in Chapter 14.

I would like to thank my former *Hewlett-Packard* management for allowing me to work on these topics, and to present them as part of this dissertation.

This dissertation includes work which has been carried out while I have been employed by *Hewlett-Packard Company.* Intellectual property rights associated with such work are owned by *Hewlett-Packard Development Company, L.P.*

Hewlett-Packard Development Company, L.P. has licensed the use of such intellectual property rights for the submission and academic publication of this dissertation, but not for any other purpose.

I would also like to thank my current *Telefonica* management for allowing me to work on these topics, and to present them as part of this dissertation.

This dissertation also includes work which has been carried out during my current employment engagement at *Telefonica I+D Company*, part of the *Tele-*

fonica Corporation. Intellectual property rights associated with such work are owned by *Telefonica I+D Company.*

Telefonica I+D Company has licensed the use of such intellectual property rights for the submission and academic publication of this dissertation, but not for any other purpose.

Vita

1967 Born, Arta, Illes Balears, Spain.

1985-1991 Graduated from a B.S.+M.S. in Telecommunications Engineering, UPC, Barcelona, Spain.

1991-1993 R&D Engineer at DIMAT, working on Power Line Digital Communications, Barcelona, Spain.

1993 Recipient of the *Caixa ”Sa Nostra”* scholarship.

1994 Visiting Research Scholar working on Image and Video Compression at the Electrical Engineering Department, USC, California.

1995 Research Assistant working on Image and Video Compression at the Electrical Engineering Department, USC, California.

1994-1995 Graduated from a M.S.E.E.-Systems, USC, California.

1995-2000 Image Compression Responsible R&D Engineer for Large Format Printers, Hewlett-Packard, Barcelona, Spain.

2000-2008 Research Scientist working on Image Processing and Multimedia Indexing and Retrieval, Hewlett-Packard Laboratories, Palo Alto, California.

2002-2006 Project Manager, Photo-Video Research Project, Hewlett-Packard Laboratories, Palo Alto, California.

2008-present Research Scientist working on Multimedia Analysis, Indexing and Retrieval, Telefonica Research, Telefonica I+D, Barcelona, Spain.

- 24 Granted US patents

- 27 Filed and published (not granted yet) patents

- 4 Filed (not published yet) patents

- 4 trade secrets

- 3 defensive publications

- 10 HP-Labs tech reports

- 26 Peer reviewed conference and journal papers

Abstract of the Dissertation

Media Aesthetics Based Multimedia Storytelling

by

Pere Obrador
Doctor of Philosophy in Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain, 2011

Since the earliest of times, humans have been interested in recording their life experiences, for future reference and for storytelling purposes. This task of recording experiences –*i.e.*, both image and video capture– has never before in history been as easy as it is today. This is creating a digital information overload, that is becoming a great concern for the people that are trying to preserve their life experiences. As high-resolution digital still and video cameras become increasingly pervasive, unprecedented amounts of multimedia, are being downloaded to personal hard drives, and also uploaded to online social networks on a daily basis. The work presented in this dissertation is a contribution in the area of multimedia organization, as well as automatic selection of media for storytelling purposes, which eases the human task of summarizing a collection of images or videos in order to be shared with other people. As opposed to some prior art in this area, we have taken an approach in which neither user generated tags nor comments –that describe the photographs, either in their local or on-line repositories– are taken into account, and also no user interaction with the algorithms is expected. We take an image analysis approach where both the context images –e.g. images from online social networks to which the image stories are going to be uploaded–, and the collection images –*i.e.*, the collection of images or videos that needs to be summarized into a story–, are analyzed using image processing algorithms. This allows us to extract relevant metadata that can be used in the summarization process. Multimedia-storytellers usually follow three main steps when preparing their stories: first they choose the main story characters, the main events to describe, and finally from these media sub-groups, they choose the media based on their relevance to the story as well as based on their aesthetic value. Therefore, one of the main contributions of our work has been the design of computational models –both regression based, as well as classification based– that correlate well with human perception of the aesthetic value of images and videos. These computational aesthetics models have been integrated into automatic selection algorithms for multimedia storytelling, which are another important contribution of our work. A human centric approach has been used in all experiments where it was feasible, and also in order to assess the final summarization results, *i.e.*, humans are always the final judges of our algorithms, either by inspecting the aesthetic quality of the media, or by inspecting the final story generated by our algorithms. We are aware that a perfect automatically generated story summary is very hard to obtain, given the many subjective factors that play a role in such a creative process; rather, the presented approach should be seen as a first step in the storytelling creative process which removes some of the ground work that would be tedious and time consuming for the user. Overall, the main contributions of this work can be capitalized in three: (1) new media aesthetics models for both images and videos that correlate with human perception, (2) new scalable multimedia collection structures that ease the process of media summarization, and finally, (3) new media selection algorithms that are optimized for multimedia storytelling purposes.

CHAPTER 1

Introduction

Over the last ten years, the advent of digital cameras, pervasive camera phones, and digital video cameras, has created a flood of consumer created media, as well as professionally created media. This digital information overload has created true data management problems in which end users can hardly find the media that they want to use or consume at a particular point in time [125].

In this dissertation we focus our efforts on consumer generated content, since professionally generated content has usually been edited thoroughly by professional editors and curators, and the media has been tagged with relevant tags so that it can be easily found in the future. There is, therefore, less need for automatic image/video analysis of professional media, in order to help in managing it.

The advent of on-line image and video sharing sites has allowed for manual tagging of media by a large number of users. This has helped in creating new tag based search user interfaces very popular among users –*e.g.* Flickr, PicasaWeb, YouTube. One of the main problems with these tag based search interfaces is that not all tags describe the actual content of the media [65], and therefore tags can only help find the content that has been correctly tagged. Actually, Kuchinsky et al. [71] found that most of the times full collections or folders are assigned the same tags, making it hard to find specific images in the collection. This fact allows for image/video analysis techniques to enhance the search and retrieval user experience, research field that has shown great activity recently [107, 36].

The work presented in this dissertation is in this latter research area –*i.e.*, image and video analysis applied to media search, retrieval and organization. We hypothesize that the aesthetic appeal of either images or videos, has a set of universal features that correlate well with the aesthetic perception of human end users. Therefore, we develop a series of computational models for media aesthetics, which are presented in Part I of this dissertation.

We present two applications of these media aesthetics models to media organization, search and retrieval, namely:

1. Aesthetics based image search re-ranking –see Chapter 9 in Part I of this dissertation.

2. Aesthetics based photo storytelling of personal collections –see Part II of this dissertation.

End users have increasingly large collections of digital photos as well as digital videos, and less and less time to devote to organizing them or selecting the media that they want to share with family and friends [47]. We strongly feel that the contribution that we can make to solving this problem is a greater one, and therefore we concentrate on the second application listed above –*i.e.*, photo storytelling, which is an essential, intimate and non-trivial feature of photograph use, and a vehicle by which people communicate experiences

Both these topics –*i.e.*, media aesthetics and media storytelling– are inherently subjective and therefore human centric design is essential in order to test these algorithms. In other words, the results of the algorithms are either tested against a ground truth generated by users/viewers, as we do in the case of image and video aesthetics; or the users provide their own image collections, and judge how well the algorithm performed on them, as we do in the case of photo storytelling.

We need to note that the intellectual property of the work presented in this dissertation belongs in part to the *Hewlett-Packard Company*, and in part to the *Telefonica I+D Company*, and have been protected by a series of patents, listed in Appendix D. We hereby thank both corporations for allowing us to use this material as part of the dissertation[1] [2]. Due to the fact that all this work has been done in industrial research laboratories, the presented contributions have always been developed with a close target for technological transfer to business units.

Since the work for this dissertation has been going on for a long period of time, multiple publications, including academic papers, as well as patents, have been written by the author on a variety of topics. For the sake of clarity, we have listed the author's publications, *i.e.*, each contribution, in Appendix D in the form of [C id], *e.g.* [C1,C2,C3]. The rest of the cited bibliographical references are listed in the Bibliography at the end of the dissertation, and these are listed in the traditional numbered way, *e.g.* [1,2,3].

1.1 Definitions

Here we define some of the main terms used in this dissertation. Other definitions are included in the Glossary section, at the end of this dissertation, where the

[1] *Hewlett-Packard Development Company, L.P.* has licensed the use of such intellectual property rights for the submission and academic publication of this dissertation, but not for any other purpose.

[2] *Telefonica R&D Company* has licensed the use of such intellectual property rights for the submission and academic publication of this dissertation, but not for any other purpose.

reader can also find an Acronyms section.

A media object is an object of any of the possible multimedia types, like image, video or audio. In this dissertation we focus on images and videos.

The aesthetics field deals with the human appreciation of beauty, in which they study the psychological responses to beauty and artistic experiences. Therefore, a media aesthetics model tries to automatically predict a media object's aesthetic value, *i.e.*, how beautiful would it be perceived by humans.

An event can be defined as a significant occurrence or happening, or as a social gathering or activity. In both cases, if the user is capturing memories in the form of either photos or videos, there will be a collection of media that will represent that event.

Photo storytelling is the activity of telling stories using photos as part of the resources used by the storyteller to accomplish his/her goal. In the more traditional setting, photo storytelling was performed using a photo album around which a group of people would listen to one or more storytellers explain past experiences or events [32].

1.2 Motivation

As mentioned above, information overload is one of today's major concerns. Users are capturing and generating increasing amounts of digital media due to its perceived zero cost, creating huge image and video collections. At the same time most cell phones ship with a built-in camera –see Fig. 1.1– some of them taking really high quality photos and even videos. We illustrate this assessment with some facts:

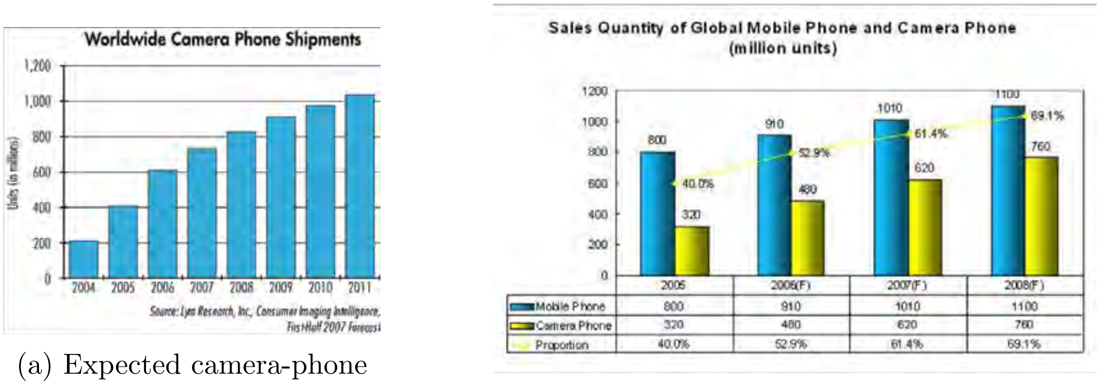
1. billions of pieces of media content per day are uploaded to Facebook in average every day (as of September 2010) .

2. 100 million photos per day are uploaded to Facebook in average every day (as of September 2010) .

3. an average user has 130 friends in Facebook (as of September 2010) .

4. about 20 hours of video footage, most of it user-generated, are uploaded on the YouTube every minute (as of May 2009).

5. the number of user-generated video creators is expected to grow by 77% from 2008 to 2013, in the US (as of May 2009).



	2005	2006(F)	2007(F)	2008(F)
Mobile Phone	800	910	1010	1100
Camera Phone	320	480	620	760
Proportion	40.0%	52.9%	61.4%	69.1%

(a) Expected camera-phone shipments

(b) Phone vs. camera-phone sales

Figure 1.1: Expected camera-phone shipments world wide (a); and phone vs. camera-phone sales world wide (b). Source: Lyra Research, Inc., Consumer Imaging Intelligence, First-Half 2007 forecast.

In order to share these collections with family and friends, they need to be parsed manually by end-users that generate a summary to be shared, imposing a big burden on them. In a series of user studies, researchers have shown the need for certain automatic or semi-automatic algorithms that should alleviate end users of such demanding and time consuming tasks [68, 47, 32]. For instance, the main implications for design in a photo organization user study presented in [68], were to design tools to help users sort, cluster poor quality –*i.e.*, poor aesthetics, cluster similar images, and keeping representative images –*i.e.*, both semantically relevant, and highly aesthetic images. At the same time Kirk et al. [68] mentions that users typically enjoy the creative process involved in photo story creation and they rely heavily on emotional and contextual information –*i.e.*, who is in the story, or who is the story going to be told to– in order to select images, also mentioned in [80]. In another user study in which users were asked to create a methodological photo story, Landry et al. [72] noticed that participants excluded portions of their experience when they did not have media to visually represent them; they also noted that users listed a sequence of events that took place in their story, as opposed to documenting the dramatic arc, as the literary theory suggests –*i.e.*, exposition, rising action, climax, falling action, and dénouement. These implications have been great motivators in our research.

Another motivation for our research has been the photo book use case, in which a fixed set of images needs to be selected, *i.e.*, the photo book has space for a fixed number of images. This motivated us to devise an algorithm that would target that specific image count, and at the same time it would select both highly aesthetic and highly representative images of the story at hand.

1.3 Goals and contributions

From the motivations stated above, we draw a set of goals for this dissertation, as follows:

1. Build media aesthetics computational models that correlate well with human perception in order to help users manage their media.

   A series of media aesthetic models will be investigated. We will look at both regression based models, as well as classification based models. We will build models for both image aesthetics as well as for video aesthetics. Since photos of people are so important, from the storytelling standpoint, we will also investigate regression based face aesthetic models. These algorithms have been protected by patents, see [C43,C49,C82,C83].

   Regression models for image aesthetics have been published in:

   [C9] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver. The role of tags and image aesthetics in social image search. In WSM 09: Proc. of the 1st SIGMM workshop on Social media, pages 65–72. ACM, 2009.

   [C11] P. Obrador and N. Moroney. Low level features for image appeal measurement. In SPIE, Electronic Imaging, Image Quality and System Performance VI, volume 7242, pages 72420T–1–72420T–12. IS&T/SPIE, 2009.

   [C13] P. Obrador. Region based image appeal metric for consumer photos. In 2008 IEEE 10th Workshop on Multimedia Signal Processing, pages 696–701, 2008.

   A regression based face aesthetics model has been published in:

   [C11] P. Obrador and N. Moroney. Low level features for image appeal measurement. In SPIE, Electronic Imaging, Image Quality and System Performance VI, volume 7242, pages 72420T–1–72420T–12. IS&T/SPIE, 2009.

   An image re-ranking application using a regression based image aesthetics model has been published in:

   [C9] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver. The role of tags and image aesthetics in social image search. In WSM 09: Proc. of the 1st SIGMM workshop on Social media, pages 65–72. ACM, 2009.

   A classification based image aesthetics model has been published in:

   [C4] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In International Conference on Image Processing, pages 3185–3188. IEEE, 2010.

A classification based video aesthetics model has been published in:

[C5] A. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. Computer Vision, ECCV 2010, 6315:1–14, 2010.

2. Create new image collection structures that allow to target specific summarization counts.

   We will create hierarchical scalable structures that allow each photo collection to be relevance ordered, *i.e.*, the first image in the list is the most relevant from a storytelling perspective; the second one, along with the first provide the best story recollection when using only two images, and so on and so forth.

   Photo collection hierarchical scalable structures have been published in:

   [C10] P. Obrador and N. Moroney. Automatic image selection by means of a hierarchical scalable collection representation. In Proc. SPIE, Vol. 7257. IS&T/SPIE, 2009.

3. Create new photo selection algorithms that sample the original photo collection in a meaningful way from a multimedia storytelling point of view. Test these algorithms in user studies in which the users provide their photo collections, and judge how well the proposed algorithms perform.

   As suggested in the implications for design in [68], we will make use of media clustering in many different dimensions including time, similarity, near-duplicates, and face recognition, in order to identify relevant story passages and characters. Media aesthetics will be used, along with certain measures of semantic relevance, in order to select representative images for each cluster. Social context will also taken into consideration, *i.e.*, who is the story going to be shared with. These algorithms have been protected by patents, see [C71,C79,C80]. Even though we do not have any academic publication in the area of video storytelling, a few patents have been filed to protect these approaches, see [C56,C75].

   Automatic photo collection summarization algorithms analyzing only the images in the collection to be summarized have been published in:

   [C10] P. Obrador and N. Moroney. Automatic image selection by means of a hierarchical scalable collection representation. In Proc. SPIE, Vol. 7257. IS&T/SPIE, 2009.

   Automatic photo collection summarization algorithms that, on top of analyzing the images of the collection to summarize, also analyze the context images that the users have in their on-line social network, have been published in:

[C2] P. Obrador, R. de Oliveira, and N. Oliver. Supporting personal photo storytelling for social albums. In Proceedings of the international conference on Multimedia, MM 10, pages 561–570, New York, NY, USA, 2010. ACM.

[C3] P. Obrador, R. de Oliveira, and N. Oliver. Audience dependent photo collection summarization. In Proceedings of the international conference on Multimedia, Grand Challenge 10, New York, NY, USA, 2010. ACM.

4. Build an automatic image selection system that helps users reduce the overall workload to generate a story to share with friends and family.

   We will build an actual software system, that will automatically ingest the user's photos, and generate an initial story summary for him/her. This proposed system should be seen as the first component of an iterative, incremental loop based on a construct, examine and improve cycle [56], which leads to the final story to be shared.

   The implementation of a system for photo collection summarization has been published in:

   [C14] P. Obrador, N. Moroney, I. MacDowell, and E. OBrien-Strain. Softbacks: single click photo selection for photo-book creation. In Symposium on document engineering. ACM, 2008.

1.4 Organization of the dissertation

This thesis is organized in three main parts. Part I presents the work we did in the area of media aesthetics assessment, where both regression and classification models are created in order to predict the aesthetic value of media objects, both images and video clips. Part II presents the multimedia storytelling algorithms we developed, in which media aesthetics play a critical role. Part III presents a practical implementation of one of the photo storytelling system described in Part II, both the back-end as well as the front end of the application. Finally, the conclusions of the dissertation are presented.

A glossary and a list of the acronyms used in this dissertation are also listed at the end of the dissertation, right after the conclusions in Chapter 19.

We have also added four appendices, in order to fully describe our contributions, namely:

1. In Appendix A we describe an image highlight detection approach, which improves the perceptual image sharpness map presented in Chapter 5 in

the presence of non-linearities caused by scene illumination on the camera light sensor.

2. In Appendix B we describe the sharpness density function, which is used in Chapter 5.

3. Appendix C is a compendium of the conference papers published by the author in the area of context based image selection for documents. In this type of approach the document into which the image needs to be inserted is used as "query" into an image database, and the image that best matches the document, from a color harmony and visual balance standpoint –*visual aesthetic* parameters– is selected. This methodology –*i.e.*, analyzing the document context with image processing techniques– inspired, in a way, the approach to social network context based image selection applied to storytelling presented in Chapter 14. In this approach we analyze the context where the collection is going to be shared –*i.e.*, the images that the user has already shared on that social network– in order to learn storytelling traits of the user, and consequently, perform a better image selection.

4. In Appendix D, we have included a list all the published documents by the author, including conference and journal papers, granted patents, as well as published patents –*i.e.*, not granted yet–, and patent applications –*i.e.*, not published yet. As mentioned above, this list is included for clarity, and is segregated from the rest of the bibliographical references at the end of this dissertation.

Part I

Media Aesthetics

CHAPTER 2

Media aesthetics introduction

In this Part I of the dissertation we present our contributions to the area of media aesthetics, which is a critical building block of our media storytelling system described in Part II. Even though media aesthetics is an important type of metadata extracted through image and video analysis, it is not enough in order to accomplish a good automatic media storytelling result on its own. Other metadata is therefore needed in order to accomplish the best possible results in media storytelling, such as capture time, color histogram, etc., as it will become clear in Part II. These other algorithms are briefly described in Chapter 3, for the sake of completeness, even though we have made few contributions to them.

Media appeal may be defined as the interest that a media object generates when viewed by human observers. This incorporates three main factors:

1. the usefulness of that media object for the task at hand that the user is willing to accomplish;
2. how the user feels emotionally about the subject –*i.e.,* content– of the media object; and, finally,
3. how aesthetic the media object is.

The first two are subjective, and therefore very hard to measure objectively. In this dissertation we are therefore going to focus on the latter, which we will refer to as *media aesthetic appeal* from now on –either *image aesthetic appeal*, or, *video aesthetic appeal*. We hypothesize that the aesthetic appeal of either images or videos, has a set of universal features that correlate well with the aesthetic perception of human end users.

The field of image and video aesthetics deals with the creation and appreciation of beauty in images. A wide range of psychological and perceptual factors play a role in our appreciation of the aesthetic value of an image [46, 101]), including the presence of people and their facial expressions, image sharpness, contrast, colorfulness, color harmony and composition. It, therefore, incorporates subjective factors on top of the more traditional objective quality measures.

As mentioned above, we face the challenge of developing efficient multimedia data management tools that enable users to organize and search multimedia content from growing repositories of digital media.

It is not uncommon to browse these large media collections in order to select aesthetically appealing media [99] to accomplish a certain task, like creating a slideshow or a photo book. There have been efforts towards building fully manual software applications that will allow for such triage in an easy and fast way, but unfortunately, when the collection size grows, any manual solution may prove tiring and time consuming. In recent years, there has been a proliferation of consumer digital media taken and stored in both personal and online repositories (*e.g.* Flickr, Picasa Web Album, Facebook, YouTube). Capturing and storing digital photos and videos is cheap and easy. Therefore, users tend to keep most of the pictures taken. As the amount of user-generated digital multimedia content increases, there is a growing need for efficient tools to search not only for relevant but also aesthetically appealing content to be shared with friends and family [99, 68].

Query-based image search approaches rely heavily on the similarity between the input textual query and the tags or other related text added to the images by users, and, therefore, they might include a large number of relevant photos, all of them containing similar tags, but with varying levels of image quality and aesthetic appeal. This approach has been somewhat successful in Web image search, where tags, comments and ratings are typically added on an image-by-image basis by large numbers of users (*e.g.* Flickr), or the image relevance is inferred from its surrounding text (*e.g.* Google Image search). However, image search in personal repositories is still a challenging task: users do not typically label each image individually, but tend to annotate their pictures in batch or bulk mode [71], assigning the same tags to groups of images that belong to the same event or photographic session. Moreover, the tags do not necessarily describe the content of the image [65]. Finally, the photos taken by the average consumer vary a lot in their photographic quality and aesthetic appeal [101].

Quantifying the aesthetic value of a media object is a hard problem [35], which explains why the simpler problem of classifying media objects into high *vs.* low aesthetic appeal has been prevalent in the research community [83, 35, 64]. In the interest of accomplishing our goals in Part II of this dissertation, though, it is of upmost interest the design of a regression based model –*i.e.*, in a photo storytelling scenario it is of critical importance being able to rank the images in a specific set based on their aesthetic value.

2.0.1 Image aesthetics introduction

User studies were conducted in order to identify the right features to use in an image appeal measure; these studies also revealed that a photograph may be appealing even if only a region/area of the photograph is actually appealing. Extensive experimentation helped identify a good set of low level features. These features were optimized using extensive ground truth generated from sets of consumer photos covering all possible appeal levels. The approach presented in Chapter 5 starts from the premise that any photograph will be looked at by a human observer, who will eventually focus his attention on the region that is more appealing within that image –this psycho-visual reasoning was introduced in [14]– and then the user may decide whether the content of such region is useful for the specific task he is considering [22]. This is a hard problem indeed, since image aesthetic appeal is mostly subjective. Savakis et al. [101] present a detailed list of relevant attributes listed by users when considering image appeal. Certain traditional quality measures appear –*e.g.* sharpness, contrast– but colorfulness, and mainly visual composition and the presence of people are prominent in the list. In Chapter 5 we build an image aesthetic appeal regression model which uses these features along with others we found useful through our own user study, and present the regression results. We also analyze the role that query relevance and image aesthetics play in the decisions that users make when selecting images in a consumer image search task; this is presented in Chapter 5, as an application case of the presented regression model.

Due to the importance of *people faces* when considering image appeal [101], a detailed study of a set of face features is also presented in Chapter 6, including face size, sharpness, contrast and smile detection, which allows us to build a face aesthetics regression model.

In Chapter 7 we focus solely on features related to image composition in order to build a classification model for image aesthetics. Visual composition is one of photography's most important features when it comes to assessing the aesthetic appeal of an image, it actually ranked as the top image attribute in an image aesthetic appeal user study [101]. Since visual composition is quite related to art, we introduce some of those concepts below in some detail, so that the algorithms presented in Chapter 7 are easier to understand.

2.0.2 Visual composition introduction

Photographers face the problem of how to compose their subjects in aesthetically pleasing ways within a rectangular frame. Over the centuries, many artists have used so called *universal* visual composition rules (*i.e.*, design principles) in order to create aesthetically pleasing images [46]. Some of the most important rules in visual composition have to do with the *simplicity of the scene*, the *balance among visual elements* and *geometry* [46]. In our work, we focus on the *simplicity of the scene* and the *visual balance* guidelines.

1. *Simplicity of the scene*: the simpler the scene, the more pleasing the image usually is, *i.e.*, the image should display the main relevant subjects, or objects, within the frame, so that distracting objects are kept to a minimum [46]. The size of an object, along with its *relative brightness* (see fig. 7.1 [46, 41]) are correlated with the object's dominance in the scene. A smaller object can also become relevant if its color contrasts with the rest of the image (these are called *accents* in the art literature). For instance, the bird's blue wing in Fig. 7.2a is an *accent* region.

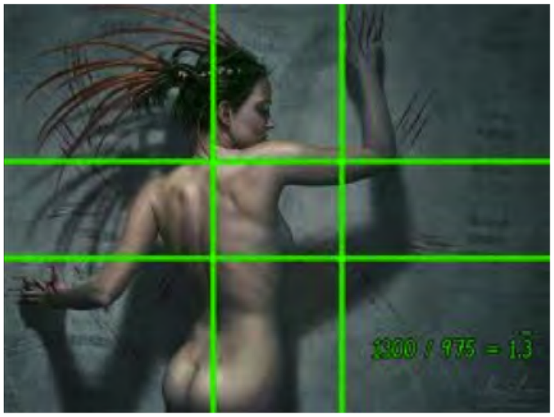


Figure 2.1: Simplicity: The simpler the scene, the more pleasing the image usually is, *i.e.*, the image should display the main relevant subjects, or objects, within the frame, so that distracting objects are kept to a minimum [46].

2. *Visual balance*: Certain layouts of the relevant objects in a scene are more aesthetically pleasing than others. The *rule of thirds* is probably the most well known *visual balance* rule. This rule states that important compositional elements of the photograph should be situated in one of the four possible intersections –also called *power points*– of the lines that divide the image into nine equal rectangles, as seen in Fig. 2.2, or along those dividing lines, potentially creating visual tension, and therefore a more interesting composition.

   The *rule of thirds* is a simplified version of the *golden mean* or *golden section* [26], which has been widely used in paintings and architecture since the Renaissance. At the time, artists realized that proportions based on simple numbers (*i.e.*, 2:1, 3:2) produced a static composition, whereas a dynamic –*i.e.*, more interesting– composition was accomplished by utilizing ratios such as the *golden ratio* $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.618$, which was known to the ancient Greeks as the best known *harmonious* division [46]. This rule also divides the frame with two horizontal and two vertical lines –see Fig. 2.3. The location of these lines is calculated using the *golden ratio*, so that each line divides the frame into two rectangles[46], the small one with side a and the large one with side b, where: $\frac{a}{b} = \frac{b}{a+b} = \varphi^{-1}$.

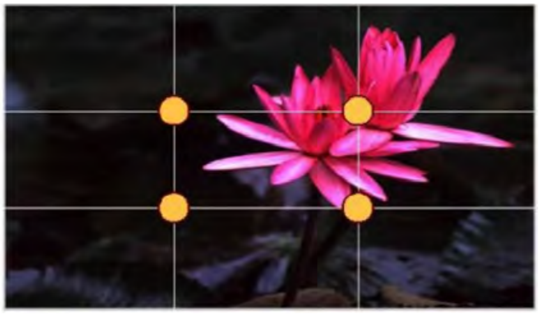


Figure 2.2: Rule of thirds: the flower is placed in the top-right power point, increasing the visual tension.

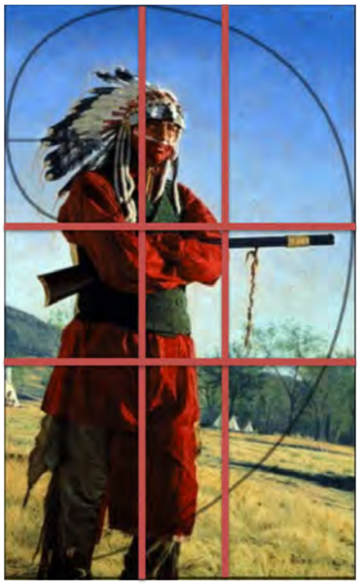


Figure 2.3: Golden mean: the person is placed on one of the golden mean dividing vertical lines.

The *golden triangle rule* is a special case of the *golden mean*, where a diagonal line divides the image, corner to corner, and a second line is drawn from one of the other corners of the image towards the diagonal intersecting it at a right angle [96], see Fig. 2.4, creating one *power point*.

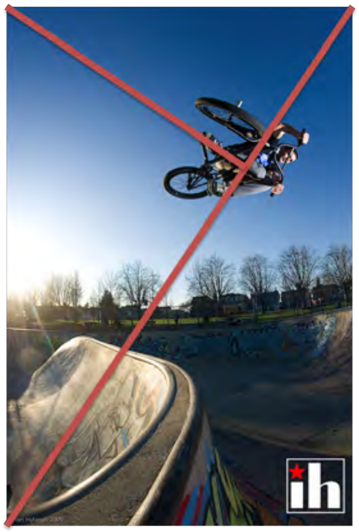


Figure 2.4: Golden triangles: the person and bicycle are located on the power point.

In addition, visual composition guidelines indicate that if a subject is not exactly along a *dividing line*, or exactly centered on a *power point*, the image may still be well composed within certain margins. The quality of the composition will degrade as the subject moves away from the *power point* or *dividing line* [46].

3. *Layout pleasantness*: Note that these are *visual balance* guideline design rules (*i.e.*, that may help at the time of picture taking). However, these rules may be broken [46], *i.e.*, an image can still be aesthetically appealing as long as its relevant objects are visually balanced, generating a *pleasant layout* in such a way – not necessarily following any of the rules stated above.

2.0.3 Video aesthetics introduction

The second type of media objects that we consider in this dissertation from an aesthetics perspective are consumer videos. In a similar way to images, text query-based image and video search approaches rely heavily on the similarity between the input textual query and the textual metadata (*e.g.* tags, comments, etc.) that has previously been added to the content by users. Relevance is certainly critical to the satisfaction of users with their search results, yet not sufficient. For example, any visitor of YouTube will attest to the fact that the most *relevant* search results today include a large amount of user generated data of *varying aesthetic appeal.* Hence, filtering and re-ranking the videos with a measure of their aesthetic value would probably improve the user experience and satisfaction with the search results. In addition to improving search results, another challenge faced by video sharing sites is being able to attract advertisement to the user generated content, particularly given that some of it is deemed to be "unwatchable" [123], and advertisers are typically reluctant to place their clients' brands next to any material that may damage their clients' reputations [85]. We believe that the analysis of the aesthetic value of videos may be one of the tools used to automatically identify the material that is "advertisement worthy" *vs.* not. Last, but not least, video management tools that include models of aesthetic appeal may prove quite useful to help users navigate and enjoy their ever increasing – yet rarely seen – personal video collections. A classification based video aesthetics model is presented in Chapter 8.

2.0.4 Organization of Part I of the dissertation

The rest of Part I of this dissertation is organized as follows: in Chapter 3, we first describe the other media analysis and indexing tools, apart from media aesthetics, which will be of great use in Part II of this dissertation – *i.e.*, for media storytelling; we then describe the prior art in the area of media aesthetics in Chapter 4, followed, in Chapter 5, by the description of the regression based model for image aesthetics –originally published in [C9 ,C11, C13]; in Chapter 6 we describe a regression based model for the aesthetic appeal of faces in photographs –originally published in [C11]; Chapter 7 describes a classification model that classifies images into aesthetically appealing, and not aesthetically appealing using solely visual composition low level features –originally published in [C4] ; in Chapter 8 we describe a classification model that classifies videos into aesthetically appealing, and not aesthetically appealing –originally published in [C5]; and we finalize with Chapter 9 with an example application of image search re-ranking using the regression based image aesthetic appeal model described in Chapter 5, which we originally published in [C9].

CHAPTER 3

Useful media indexing tools

As we mentioned above, media aesthetics alone are not enough in order to create an automatic selection for a media story. Therefore, in order to accomplish the best possible automatic summarization, other types of metadata may be needed. In this chapter, we briefly describe the metadata that will be used in Part II of this dissertation, along with their corresponding clustering algorithms. Keep in mind that we are not comparing these algorithms to the state of the art, but, rather, we describe them herein for the sake of completeness of our overall system.

3.1 Time Clustering

Time clustering groups the images that are close in time together, which is an effective way to detect the temporal events in a photo collection. Time clustering has, therefore, been used in the past in order to detect events in photo collections by analyzing the differences between capture time of adjacent photographs in the collection, either with fixed thresholds [94], or with adaptive thresholds [95, 50, 80, 51]. Also, multi-resolution temporal approaches have been used to detect the event boundaries, as described in [31].

Some very simple, and fast, algorithms have been used in this dissertation, which give a good result once they are coupled with other clustering approaches, like similarity and near-duplicate clustering –see Section 3.2. In this dissertation, due to historical reasons, two different time clustering algorithms have been used, both of which are now briefly described.

3.1.1 Time clustering algorithm #1

We implemented a hierarchical time clustering based on the algorithm presented in [51], which subdivides the collection into increasingly smaller event clusters. The algorithm looks at the time difference between consecutive photos, and draws a relationship weight (RW), see Eq. 3.1. This algorithm will be used in the system presented in Chapter 13.

$$RW(t) = -log(t_i - t_{i-1} + 1) \tag{3.1}$$

By finding the global minimum of RW(t), the collection can be split into 2 temporal events, which in turn can be split recursively by finding their corresponding local minima; this process is repeated until reaching a minimum predetermined cluster size.

3.1.2 Time clustering algorithm #2

In this second algorithm, time clusters are detected by an algorithm similar to that in [94], where a photo is included into a new time cluster if it was captured more than a certain amount of time T_t since the previous photo was captured. This allows us to target a specific number of time clusters just by varying T_t, which is an important feature in the photo storytelling system described in Chapter 14.

3.2 Image similarity

Content based retrieval has used different types of similarity metrics in order to retrieve images from a database that were close to an example image presented by the user. Color, texture, image regions, local features, like the scale-invariant feature transform (or SIFT), etc., have been used in the past in order to measure similarity [107, 36, 81].

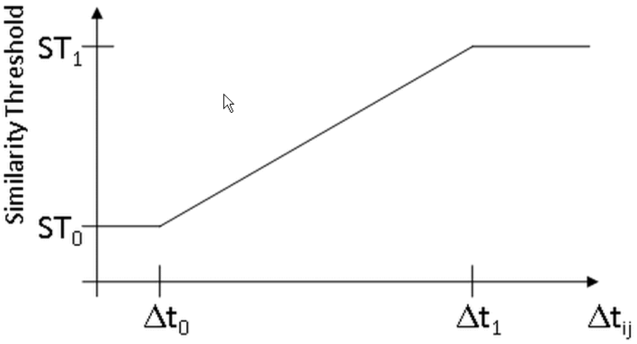
As it will become clear in Part II of this dissertation, one of the important aspects of photo storytelling is to avoid repetition (*i.e.*, avoid redundant images that are taken one right after the other of the same subject/scenery), or select a subset of images of one specific setting –*i.e.*, which may be quite similar in color content.

In this dissertation, three similarity algorithms have been used at different points in time: a global color similarity algorithm, a region-based color similarity algorithm, and a local feature based near-duplicate detection.

We made a contribution to each of these algorithms by modifying the similarity threshold depending on the difference between the capture time of the two photos being considered. Cooper et al. [31] already combined both time and content in order to improve event clustering with a similarity measure that linearly relies less on content-based similarity as the inter-photo capture time difference grows. Inspired by this approach, we propose a similarity threshold that varies linearly with the difference between the photo's capture time, ΔT, see Fig. 3.1. This is quite useful in a storytelling setting, where, even though two images might be similar, if they were taken a long time apart, the emotional content of each of those photos might be different for the user.

Images belong to the same cluster if they are close in time and/or they are similar, *i.e.*, in many collections we noticed specific groups of similar images, to be interleaved with other photographs, where the photographer might have turned around to capture a different object or view, and then turned around again, to capture the main event that was being captured previously. In order to be able to capture these intermittent groups of similar images, we make the similarity threshold in all our image similarity algorithms, to be dependent on the capture time difference between pictures, *i.e.*, a group of photos belongs to a similarity cluster, if any pair of photos' similarity measure is above a scene similarity threshold $ST(\Delta T)$ which varies with the difference between pictures' capture time ΔT, see fig. 3.1.

Figure 3.1: Capture time difference dependent similarity threshold.



The similarity threshold is low for $\Delta T \ll \Delta T_1$, *i.e.*, images taken close in time and similar to each other will be clustered together –with $\Delta T_0 = 5 seconds$, whereas for $\Delta T \geq \Delta T_1$, the similarity threshold is $ST_1 = 1$, *i.e.*, similarity does not apply. As opposed to [31], where the time difference cap ΔT_1 is set to 48 hours, we make it adaptive to the time duration of the time cluster at hand: $\Delta T_1 = \frac{1}{3} timeClusterDuration$. Gargi [48] showed that these thresholds are actually user dependent, and therefore could potentially be personalized.

The parameters presented in this section were optimized by using 4 collections from 3 different users which we used as our ground truth.

3.2.1 Global color similarity

In order to detect similar images that belong to the same setting, or *scene*, as defined in Chapter 14, we use a relatively simple global color similarity metric, which is the normalized histogram intersection in HSV – hue, saturation, value

[52] – color space [108], which proves to be fast yet quite effective for our storytelling application.

3.2.2 Region based color similarity

In this algorithm, similarity is calculated by a region based approach. First the image is quantized to a set (usually 25) of lexical colors (*i.e.*, color names used most often by humans in on-line experiments), and the image is segmented into regions of constant lexical color [C12]. These lexical color regions are quite resilient to illumination changes, which helps in the case of slight exposure changes common in consumer level digital cameras. The similarity calculation is directly proportional to the sizes of the color patches, and inversely proportional to the region distance –in pixels– and inversely proportional to the color distance –in CIE-Lab space [39], for details please refer to the patent [C73], which, at the same time, is a similar approach to the one we presented in [C17]. By varying the similarity threshold, we can obtain a hierarchy of similarity clusters, which is used in Chapter 13.

3.2.3 Near duplicate images

Since the introduction of the SIFT local features by Lowe [81], these, and variations of similar conceptual features, have been used in the literature for detecting objects, buildings, etc. in photographs, due to their much better accuracy than any of the prior algorithms attempting to do the same task. We, therefore, use Lowe's implementation of the SIFT features, along with the normalized SIFT [81] feature similarity function described in [106]. This allows us to detect near-duplicate images in the photo-storytelling system proposed in Chapter 14. In that chapter, a set of images that are detected as near-duplicates are referred to as a *shot*.

3.3 Face processing

The *characters* in a story are probably its most important element [78]. Users tend to be quite sensitive to the selection of images with people in their photo stories. At the same time, the aesthetic appeal of those faces, as described in Chapter 6, is an important element of the final appeal of the multimedia story. Therefore, face detection, clustering and expression recognition, turn out to be of great importance in order to come up with a high quality multimedia story.

No new contributions were made to the algorithms presented below, which are shortly described for the sake of completeness.

3.3.1 Face detection

In this dissertation we use the algorithm presented in [118, 62], using the implementation included in the OpenCV package [19]. This algorithm is quite effective at finding faces in images, and at the same time it gives an accurate positioning and size of each face.

3.3.2 Face clustering

In order to detect the main *characters* of a story, it is necessary to cluster the images based on *who* appears in the photos. In this way, a *character* histogram can be built, and more relevance can be given to the largest face clusters.

In this dissertation we have used two main face clustering algorithms:

1. The face clustering algorithm described in [129], was used in the system presented in Chapter 13.

2. The face clustering algorithm described in [62] , was used in the system presented in Chapter 14.

3.3.3 Smile detection

Two smile detection algorithms have been used in our work. The first one is presented in[116], which was used as part of our face aesthetic appeal metric in Chapter 13. The second one is a variation of the algorithm presented in [124], which was used as part of our face aesthetic appeal metric in Chapter 14. Both return a smile probability.

CHAPTER 4

Media aesthetics prior art

The field of media aesthetics has recently generated a lot of interest in the image and video analysis research community. The main goal behind such research is to accurately characterize and measure aesthetic appeal of both images and videos. Given the human-centric nature of the concept of media aesthetics, two approaches have been proposed in the literature to collect aesthetics ground truth:

1. Visual Experiments: In these approaches, ground truth data is gathered through user experiments where aesthetic appeal ratings are collected from each inspected photo or video. The images are usually chosen from unedited personal collections so that they represent the full spectrum of image aesthetics in a typical consumer image management environment (see Fig. 9.2). This ground truth is then used to train an aesthetics model [82, 113, 127]. The main drawback of this approach is that the training data set tends to be small due to the high cost of manually labeling/rating all the media objects.

2. Photo Forum Data: Currently, there are a number of Web communities where people share vast amounts of high-quality images on virtually any topic (*e.g.* Flickr.com, photoblogs.org, photosig.com, photo.net). The photos that are published in these forums tend of have higher quality than the average consumer photo [131]. In addition, these photos are typically labeled with rich metadata and ratings –actively provided by volunteer users– such that these ratings can be used to rank photos based on their voted quality (*e.g.* rank by most interesting in Flickr). For instance, San Pedro et al. [100], use the *favorite* assignments on Flickr to draw the required ground truth.

In this dissertation we focus mainly on the former –*i.e.*, visual experiments– even though our image composition analysis in Chapter 7 was done using photo forum data from *photo.net*.

In the literature, what we are referring to as *image aesthetics*, has been named in a variety of ways: image aesthetics [35], image attractiveness [100], image appeal [101], or photo quality [64, 83].

One of the earliest works in the domain of image aesthetics is that by Savakis *et al.* [101] where they performed a large scale study of the possible features that might have an influence on the aesthetic rating of an image; these were listed by users when considering image aesthetic appeal in their personal image collections. While traditional quality measures appear in the list –*e.g.* sharpness, contrast– other more subjective features like colorfulness and specially image composition and the presence of people are cited as two of the most important features. However, no algorithm was proposed to evaluate the aesthetic appeal of the images.

We now describe the prior art in the areas of *image and video quality*, which the study by Savakis *et al.* [101] showed were important when assessing the aesthetic value of a media object. We then describe the specific prior art in the areas *image aesthetics*, *visual composition based aesthetics* and *video aesthetics*. We finalize with a brief description of the prior art in the area of *image search re-ranking*, which is relevant to the application described in Chapter 9.

4.0.4 No reference image and video quality prior art

In the area of the more traditional no-reference image and video quality there has actually been quite a lot of work during the past decade, of which the work by Wang et al. [122] is the most referenced, in which they presented a new framework for the design of image quality measures based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field. It follows that a measure of structural information change can provide a good approximation to perceived image distortion. Reviews on image and video quality assessment algorithms can be found in [126, 122, 90, 38]. We have also done some work in the area of no-reference quality measures, more particularly in the area of sharpness assessment; this work has been published in [C6] and in [C1].

4.0.5 Image aesthetics prior art

We now list the prior work in proper media aesthetic appeal field. We first describe algorithms that try to estimate the aesthetics by analyzing the whole image; next we describe algorithms that divide the image into two main areas –*i.e.*, the region of interest and the background– and analyze each of them independently; and finally we look at one algorithm that performs no image analysis whatsoever, and estimates media appeal through a different set of metadata.

Most of the earlier work in assessing image aesthetics tried to estimate it by processing the image as a whole:

1. In [80] low quality screening is performed in an albuming application. In [82, 127] image appeal is measured by calculating sharpness and colorfulness.

2. In [113], Tong *et al.* extracted features – including measures of color, energy, texture and shape – from images and a two-class classifier (high *vs.* low aesthetic appeal) was proposed and evaluated using a large image database with photos from COREL and Microsoft Office Online (high aesthetic appeal) and from staff at Microsoft Research Asia (low aesthetic appeal). One drawback with this approach is that some of the selected features lacked photographic/perceptual justification. Furthermore, their dataset assumed that home users are poorer photographers than professionals, which may not always be true.

3. Datta *et al.* [35] extracted a large set of features based on photographic rules. Using a dataset from an online image sharing community, *i.e.*, photo.net, the authors discovered the top 15 features in terms of their cross validation performance with respect to the image ratings. The authors reported a classification (high *vs.* low aesthetic appeal) accuracy of 70.12%. This approach incorporates a low depth-of-field indicator, a shape convexity score and a familiarity measure. This familiarity measure yields higher aesthetic appeal for uncommon images checked against a large image library.

4. Ke *et al.* [64] utilized a top-down approach, where a small set of features based on photographic rules were extracted. The spatial distribution of edges, color distribution and hue count are incorporated in order to classify between high quality photos and low quality photos. A dataset obtained by crawling DPChallenge.com was used and the photo's average rating was utilized as ground truth.

5. San Pedro *et al.* [100], use textual –*i.e.*, from the Flickr web site– as well as various visual features for constructing a vector representation of photos and for building classification and regression models. They show that both textual and visual features complement each other and provide better results once combined.

More recently there have been other approaches that analize different regions in the image in order to come up with the final aesthetic measure:

1. In [83], Luo and Tang furthered the approach proposed in [64] by extracting the main subject region (using a sharpness map) in the photograph. A small set of features were tested on the same database as in [64], and their

approach was shown to perform better than that of Datta *et al.* [35] and Ke *et al.* [64]. In this method, contrast and simplicity features are introduced in order to measure isolation of the main region from distractions in the background; lighting, composition geometry and color harmony are also analyzed.

2. In [128] certain features are analyzed for the whole image, and other features are analyzed for the salient region – see [59] for a detailed description on image saliency.

Other approaches do not perform any image analysis, and use other sources of information in order to assess the aesthetic appeal of an image:

1. In [131], a vertical image search engine was introduced that: (1) indexes images from multiple photo forums; and (2) ranks the images based on a quality index that is obtained from the rating scores across forums, without any content image analysis.

4.0.6 Visual composition based aesthetics prior art

Since image composition is regarded as one of the most important features regarding image aesthetic appeal [101], we look at the prior art regarding this feature independently. Few image aesthetics algorithms have taken image composition features into consideration. *Simplicity* has been accounted for in various ways: as the number of colors, quantized to 4096 bins, in the background of the region of interest (L_1) [83]; the number, up to 5, of segmented regions larger than 1% of the image size (D_1) [35]; or the overall number of segmented regions (F_1) [40]. Low depth of field photography (*i.e.*, having the region of interest in focus, and the background out of focus as has been considered in [83, 35]), and having the main subject be *salient* (considered in [128]) can also help reduce complexity. *Visual balance* compliance with the *rule of thirds* is measured in [83] by calculating the minimum distance of the centroid of the region of interest to the four *power points* (L_2). In [35, 128], the Hue, Saturation and Value (*i.e.*, HSV color space) averages within the inner *rule of thirds* rectangle are computed (D_2, D_3, D_4). The features that we have presented, $D_1, D_2, D_3, D_4, L_1, L_2$, and F_1, were implemented and will be used in Chapter 7 in our competitive study.

4.0.7 Video aesthetics prior art

To the best of our knowledge, only the work in [83] has tackled the challenge of modeling video aesthetics. In that work their goal was to automatically distinguish between low quality (*i.e.*, amateurish) and high quality (*i.e.*, professional) videos. They applied image aesthetic measures, where each feature was calculated on a subset of the video frames at a rate of 1 frame per second (fps), coupled with two video-specific features – length of the motion of the main subject region and motion stability. The mean value of each feature across the whole video was utilized as the video representation. They evaluated their approach on a large database of YouTube videos and achieved good classification performance of professional *vs.* amateur videos (≈ 95 % accuracy).

4.0.8 Image search re-ranking prior art

One of the applications of an image aesthetics regression model is image search re-ranking – see Section 9, using the image aesthetic appeal measure for each of the photos as an extra parameter for the final rank assigned to the photos. We briefly review the most relevant prior work in the area of image search re-ranking.

The retrieved images in current image search engines may be disorganized or irrelevant for a particular user [33]. In order to solve this problem, two approaches have been proposed in the literature:

1. semantic clustering of the results [119] –through either visual, text or link analysis (or a combination of them, [21]); and
2. image search re-ranking, where the results of a baseline search engine are re-ranked according to some criteria.

One of the most popular criteria in image search re-ranking is the analysis of the visual information contained in the images, in order to re-rank and improve the baselines search engine results [33, 60, 43]. Most of the prior work in this area assumes that there is one dominant cluster of images within each image set returned by a keyword query, and treats images inside this cluster as the desired ones [43, 58]. Schroff et al. [103] first remove irrelevant images from the retrieved set by using a multimodal approach (*i.e.*, text, metadata and visual features), and then re-rank based on visual similarity. In [60], a visual similarity based graph is created among all retrieved images in the set, and they apply an iterative procedure based on the PageRank computation in order to re-rank the images. In order to close the semantic gap, a human interaction loop is proposed by Cui et al.[33]: the user picks one of the images from the retrieved set –which is categorized into one of the predefined categories– and a category-specific similarity clustering generates the final re-ranking results. Alternative non content-based approaches assign a relevance metric to the images based, for example, on the relevance of the HTML document linking to the image [77]. In the area of consumer images, [66] presents a re-ranking method for Flickr images, that fuses tag relevance with location annotation information and visual cues, producing a ranked list of clusters representing different views of a certain location.

In a similar way to [66], in this dissertation we evaluate the fusion of the tag relevance with a visual cue: the aesthetic appeal of the image, so that images that are aesthetically more appealing would be ranked higher. Note that while Luo et al. [83] have recently proposed an image aesthetics re-ranking algorithm for Web images (queried form the MSN Live Search engine), we tackle the problem of consumer image search (Picasa Web Album images [1]). In addition, we evaluate the role that query relevance and aesthetics play both individually and in combination, by means of a user study – see Chapter 9.

Finally, closely related to our re-ranking work is that of Choi and Rasmussen [27], who explored which factors –beyond tag semantics– play a role in an image search task. In a user study of an image retrieval system (the American memory photo archives of the Library of Congress), they observed that users valued image quality and clarity (*i.e.*, parameters of image aesthetics) in addition to the image semantics. While Choi and Rasmussens work also adopts a human centric perspective in the context of an image search task, the work presented in Chapter 9 focuses on consumer image repositories (*i.e.*, consumer image management), and analyzes both quantitative and qualitative user feedback on the role that query relevance and aesthetics play in a personal image search task.

CHAPTER 5

Regression based image aesthetics

In this chapter we present an image aesthetics regression model that has been fine-tuned to have high precision at the top of the ranking, as shown in the results presented in Section 5.6. Having good precision at the top of the ranking is essential in the media storytelling algorithms presented in Part II of this dissertation, as we will see. We originally published this work in [C9,C11,C13], and was also protected by two patents [C43,C49].

Another application of such a model is the ranking of a set of images based on their aesthetic appeal, which we test in Section 5.6. And, finally, a quite relevant application, given the popularity of web based image search engines, is the possibility to combine the relevance given by the search engine, with the aesthetic appeal of the image. In Chapter 9 we describe such a system from a human centric perspective.

The current chapter is organized as follows: in Section 5.1 we present the learnings from a user study, which provides some insights for the features that should be implemented in order to build our regression model; these low level features are described in detail in Section 5.2; the approach to extracting the relevant region of the image is presented in Section 5.3; in Section 5.4 a set of multiplicative factors, dependent on other low level features, are presented, which modify the aesthetic appeal of the photograph in a multiplicative way, as presented in Section 5.5; finally, the regression results for this model are presented in Section 5.6.

5.1 User Studies

A set of user studies were conducted [C13] in order to understand what the relevant image aesthetic appeal features are. Six photographers, with different expertise level ranging from point and shoot expertise all the way up to semi-professional, were interviewed about the main attributes they look for in images within each of the 6 image aesthetic appeal categories used, namely, Excellent, Very Good, Good, Fair, Poor and Very Poor. The following list summarizes the findings.

Excellent photos, typically have good focus, contrast, colorfulness, and excellent composition. A categorization of Excellent requires higher sharpness, since these images may be blown up on the album cover or in a frame.

Good/Very Good photos need NOT be in perfect focus (corroborated by [101]). There is a sharpness threshold (different for each human observer), above which an image will make it into the Good aesthetic appeal category, given that it has reasonable contrast and it is reasonably colorful, *i.e.*, it is important to preserve such memory at the expense of some blur. Having a region or a few regions with high image aesthetic appeal is enough to categorize such image as Good aesthetic appeal, or above.

Fair and **Poor** images will only be used if they depict a relevant, semantically important, event.

Very Poor images will rarely be used.

- One of the main findings from the user studies conducted for this research was that observers may find an image aesthetically appealing even if only one relevant region in the image is aesthetically appealing. In other words, there is no need to have all the regions in an image be aesthetically appealing for the image to be considered aesthetically appealing.

- The interesting regions of an image have more aesthetic appeal if the tone reproduction of that region is better. Tone reproduction is the process of mapping scene luminance to image lightness; Ansel Adams, the photographer, described in [7] the *zone system*, a system to obtain the proper exposure in order to maximize the quality of the image by determining where detail would appear in an image.

- Photos with a single relatively large object easy to isolate from the background (size and homogeneity parameters introduced in [84]) are favored. These parameters applied to the relevant region as we define it in Section 5.3 correlates well with professional photography, where the main subject or subjects are usually isolated from the background, either by placing the subject before a non-distracting background, or by using low depth of field (blurred background), allowing the viewers' attention to focus on the relevant object [127].

- People (medium shot and portraits) are usually very important, more so for certain users than others. People smiling are usually favored in consumer photography. Incorrect color rendering of people faces is very unappealing.

- Photos from a certain aesthetic appeal category may move up into a more aesthetically appealing category after cropping for better composition.

These user studies allowed us to focus on the main features to be extracted from images in order to quantify their image aesthetic appeal, which are described in the following sections.

Based on the findings from our user study described above, we devise a set of low level features, that allows us to measure the aesthetic appeal of consumer photos –see Section 5.6. Both for training and testing purposes, consumer images, ranging from very bad to excellent, were used in this research, and also the observers that tagged the images based on their aesthetic appeal ranged from point-and-shooters to photography enthusiasts.

Bajcsy [14] stated that when humans look at a photograph they tend to focus their attention on the region that is most aesthetically appealing within that photograph, and then, based on the properties of this region they decide whether the image is useful for their task at hand, *i.e.*, whether the image is aesthetically appealing to them in order to accomplish that specific task or not. The aesthetics regression model proposed in this chapter takes advantage of these findings by identifying the region in the photograph that is most aesthetically appealing – from now on, referred to as the relevant region.

We hypothesize that an image is aesthetically appealing if it has a large relevant region that is sharp, colorful, well illuminated and well contrasted with the background. The overall aesthetic measure is composed of a main term, a combination of sharpness, contrast and chroma within the relevant region, and a set of penalty and reward factors, where each of these factors either increase or decrease the overall aesthetic measure of the photograph.

Note that each image is first downsampled, so that the longest side of the photograph will be 1024 pixels, in order to normalize the sharpness and noise measures across images.

The relevant region is calculated in an image region by region basis, by (see Fig. 5.1):

1. first performing an image segmentation[1]; and within each of the image regions,

2. calculating the region's sharpness, described below in Section 5.2.1,

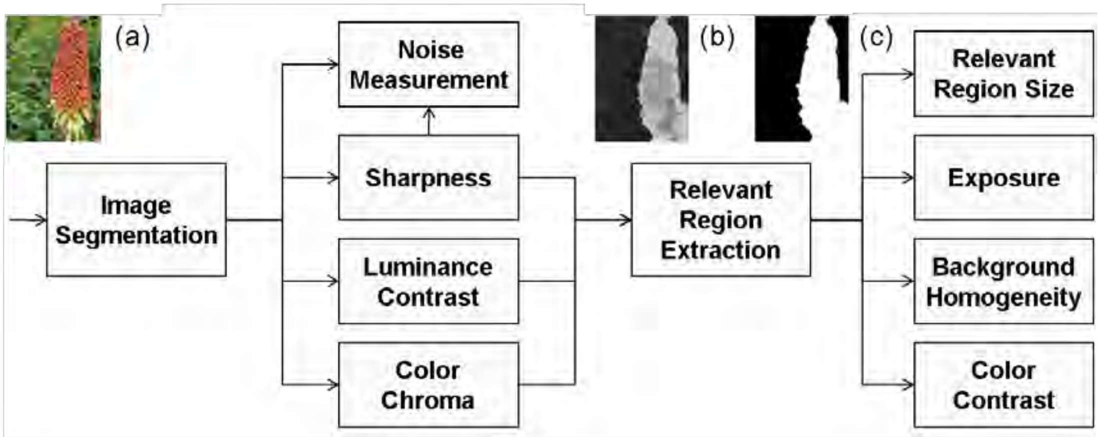
3. calculating the region's colorfulness, described in section 5.2.3,

4. and, calculating the region's average luminance contrast, as described in Section 5.2.4;

[1] In this dissertation, a variety of image segmentation algorithms have been used: (a) in Chapter 13 we used a segmentation algorithm proposed by the author in [C12, C18]; (b) in Chapter 14 we used a fast and efficient image segmentation algorithm presented in [42]

5. finally, these region based features are combined in order to obtain the *relevant region* of the image being analyzed, as described in Section 5.3.

Figure 5.1: Proposed region based image aesthetic appeal framework. Left (a): original image; Center (b): Image Aesthetic Appeal Map; Right (c): Relevant Region.



5.2 Low level features

As described above, a set of low level features were identified as being relevant to quantifying the aesthetic value of a digital photograph. In this section we will describe each and every one of these features in detail.

5.2.1 Sharpness

Sharpness describes the clarity of detail in a photo and it is a fundamental block in our image aesthetics model. In other words, the sharpness feature measures how well-focused each region in the photograph is. Each region in the image is assigned a representative sharpness value. The actual sharpness measure for each pixel (j,k) is based on a multi-resolution laplacian approach calculated on the luminance channel of the image (see Fig. 5.2), in which all 4 levels of the laplacian pyramid are combined in order to be resilient to image noise, see Eq. 5.1. This is done in such a way due to the known fact that sharp edges have energy content at all frequency bands, and therefore each of the presented filters would generate high energy, *i.e.*, if any of them does not output a strong signal at a specific pixel, it means that there is no sharp edge at that particular pixel.

The algorithm we have implemented is the following. The original luminance

Figure 5.2: Region based sharpness calculation workflow. Starting with a 4-level laplacian filter-bank, followed by the combination of these outputs on a pixel by pixel basis (Eq. 5.1), and modifying its result by the local contrast (Eq. 5.2). Each region in the segmented image is then assigned one single sharpness value.

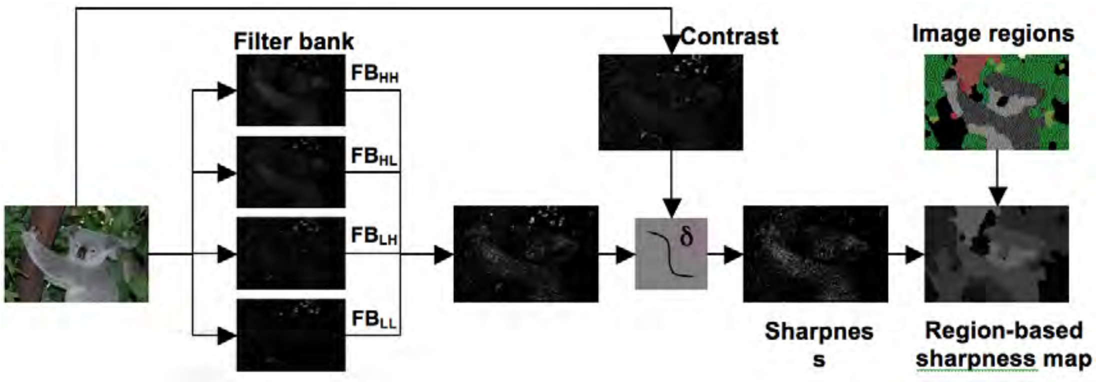


image is first low pass filtered by a 3×3 gaussian kernel $-\sigma = 0.5-$, FB_{HH} is obtained by taking the absolute value of the difference between the original image and its filtered version, resulting in a high pass version of the image –therefore HH as subscript. That low pass filtered version is downsampled $-2 \times 2-$ and filtered again with the same filter, repeating the process and generating FB_{HL}. This process is repeated twice more, generating FB_{LH}, and finally FB_{LL}. After downsampling each filter output to the smallest output size –*i.e.*, FB_{LL} –the four filter outputs are combined in Eq. 5.1.

$$S(j,k) = \delta(contrast(j,k)) \cdot FB_{LL}(j,k) \cdot FB_{LH}(j,k) \cdot FB_{HL}(j,k) \cdot FB_{HH}(j,k) \quad (5.1)$$

High contrast edges are well known to generate a much higher sharpness measure than the one perceived by the Human Visual System (HVS); in order to counteract this problem a local contrast correction function, δ, has been implemented based on [44] (see Eq. 5.2).

$$\delta(contrast(j,k)) = \begin{cases} -0.0042 \cdot contrast(j,k) + 1 & \text{if } 0 \leq contrast(j,k) \leq 50 \\ 0.8 \cdot e^{-0.024 \cdot (contrast(j,k) - 50)} & \text{if } 50 < contrast(j,k) \leq 200 \end{cases} \quad (5.2)$$

In order to solve the *blooming* problem that exists with certain camera sensors, when electrons of a clipped[2] pixel flow into the neighboring cells –this can be seen as halos around objects in backlit images– the contrast function $contrast(j,k)$ has been implemented in a multi-resolution approach, in order to capture both high and low resolution contrasts. The contrast at each resolution is measured using the root-mean-square contrast, as presented by Peli in [92].

$$contrast(j,k) = max\Bigg(\left[\frac{1}{8}\sum_{3\times3}(x-\bar{x})^2\right]^{1/2}, \left[\frac{1}{48}\sum_{7\times7}(x-\bar{x})^2\right]^{1/2}, \left[\frac{1}{120}\sum_{11\times11}(x-\bar{x})^2\right]^{1/2}\Bigg) \tag{5.3}$$

To this effect, the contrast is calculated over three windows of different sizes (3×3, 7×7 and 11×11, all centered on the inspection pixel (j,k)), and the maximum of the three is selected to be the contrast measure at that point (see Eq. 5.3).

Fig. 5.2 shows the whole sharpness metric calculation diagram, which was originally presented in [C13]. An important limitation of this model is that strong highlights may cause the luminance in pixels to be clipped on the sensor itself. This produces non-linearities –*i.e.*, aliases– which are incorrectly detected in our filter bank, generating an incorrect sharpness measure. Appendix A presents a simple, yet effective approach to solve this problem.

Finally, our regression based aesthetic appeal model is a region based one, and therefore we will assign one single sharpness value to each of the regions the image has been segmented into.

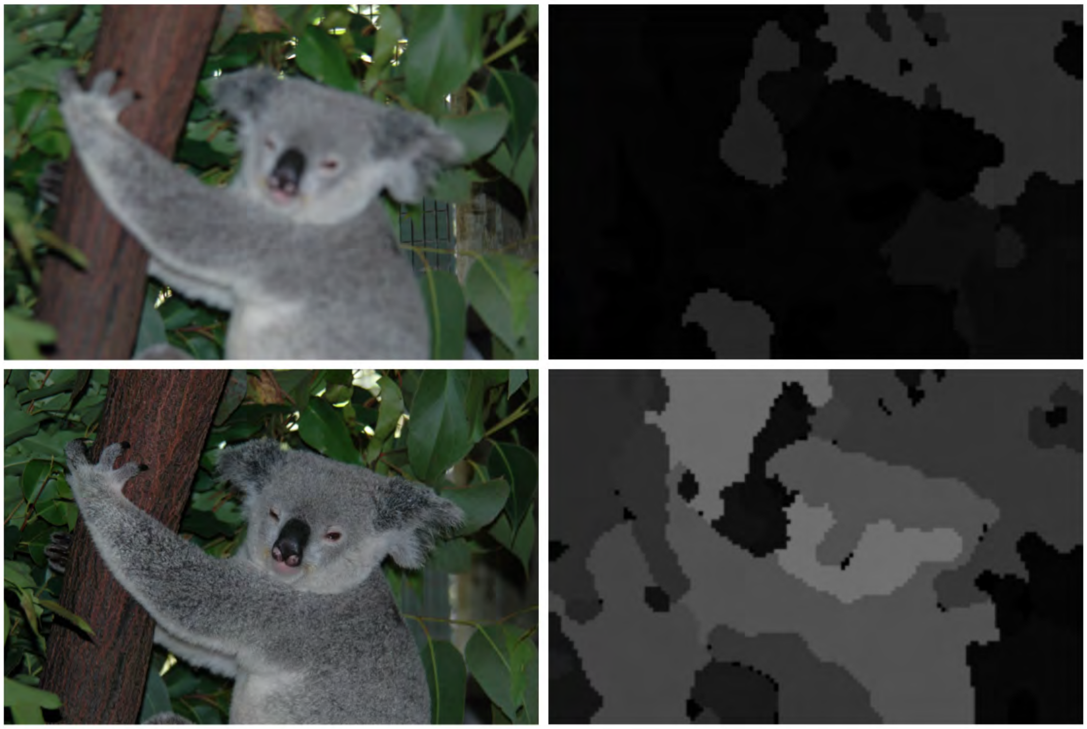
In order to avoid the *border effect* –*i.e.*, the border of an object offers an ambiguous sharpness measure since it is not clear whether it belongs to the object or the background– all the regions in the segmented image are eroded, so that only the sharpness measures inside each of the regions are taken into account in order to assign the sharpness to each of those regions.

For each eroded region, the maximum value of sharpness is assigned to it as the region sharpness. Fig. 5.3 shows an example of two photos of the same

[2] A *clipped* pixel is equivalent to the pixel with its photo detector saturated, *i.e.*, completely full. Instead, *saturation* is used in this dissertation to describe the purity of colors in a variety of color spaces

subject, a koala, one in perfect focus, and another one completely out of focus, and their respective region based sharpness maps.

Figure 5.3: Region based sharpness maps. In the top example, the koala is completely out of focus; in the bottom example, the koala is in perfect focus.



5.2.2 Sharpness density

Sharpness density (SD) is defined as the percentage of an image region that has energy content in mid and high frequencies –*i.e.*, not only low frequencies. This is a very convenient measure in order to identify which regions have a reliable sharpness measure, since regions with low SD may actually be composed of multiple regions that have been merged in the segmentation process, some of them flat; or the whole region is too flat to have any reliable sharpness reading. Section 5.3.1 will show how the sharpness density is very useful to fuse the sharpness, contrast and colorfulness for each of the image regions. See Appendix B for a detailed description of this algorithm.

5.2.3 Color based features

Colorfulness of a color is usually defined as the difference between that color against gray. For a photograph, measuring its colorfulness it is a bit harder since the definition of a colorful photograph is not clear, but experts seem to agree that it is correlated with the color variety and their purity in the photograph. In [54], Hasler et al. carried out an experiment with observers in order to come up with a formula that would correlate with human perception of photograph colorfulness. In that work Hasler et al. combine both the chroma standard deviation ($\sigma_{a_i b_i}$) as well as the mean chroma magnitude ($\mu_{a_i b_i}$) in the CIE-Lab color space [39] for a particular region i, see Eq. 5.4.

$$cf_i = \sigma_{a_i b_i} + 0.37\mu_{a_i b_i} \tag{5.4}$$

Where $\sigma_{a_i b_i}$ is the trigonometric length of the standard deviation in CIE-Lab space, which we are going to refer to as *color contrast* –ccn_i. And where $\mu_{a_i b_i}$ is the distance of the centre of gravity in CIE-Lab space to the neutral color axis, for all pixels in region i.

$$ccn_i = \sigma_{a_i b_i} = \sqrt{{\sigma_{a_i}}^2 + {\sigma_{b_i}}^2} \tag{5.5}$$

Unfortunately, when trying to rate the aesthetic appeal with respect to colorfulness on a region by region basis Eq. 5.4 the color contrast term is quite insignificant, since most image segmentation algorithms [42] [C18] generate segments that have roughly homogeneous color, and therefore the first term in Eq. 5.4, $\sigma_{a_i b_i}$ (which we will referring to as *color contrast*, see Eq. 5.5), is quite insignificant. Therefore, and in the interest of speeding up the computation, we have dropped this term in the image aesthetic appeal map calculation, *i.e.*, we calculate only the chroma mean ($0.37\mu_{a_i b_i}$) within each region i, see section 5.3.1.

As shown in [54], color contrast is also very important in order to judge the aesthetic appeal of images, and we will show how we incorporate this measure –calculated on the overall image– into our system in section 5.4.2.

5.2.4 Luminance contrast

Contrast is a measure of the difference in brightness between light and dark areas in a scene. The contrast, CN, is measured in each region using the root-mean-square contrast, which has been proven useful when comparing contrast of different images [92]. See Eq. 5.6 and 5.7.

$$CN_i = \left[\frac{1}{n_i - 1} \sum_{j,k \, \epsilon \, region_i} (x(j,k) - \bar{x_i})^2 \right]^{1/2} \tag{5.6}$$

$$\bar{x_i} = \frac{1}{n_i} \sum_{j,k \, \epsilon \, region_i} x(j,k) \tag{5.7}$$

5.2.5 Chroma noise

High image noise (often found in camera-phone images or in very low end cameras), also degrades the aesthetic appeal of images. In this section we describe a method to tackle this problem.

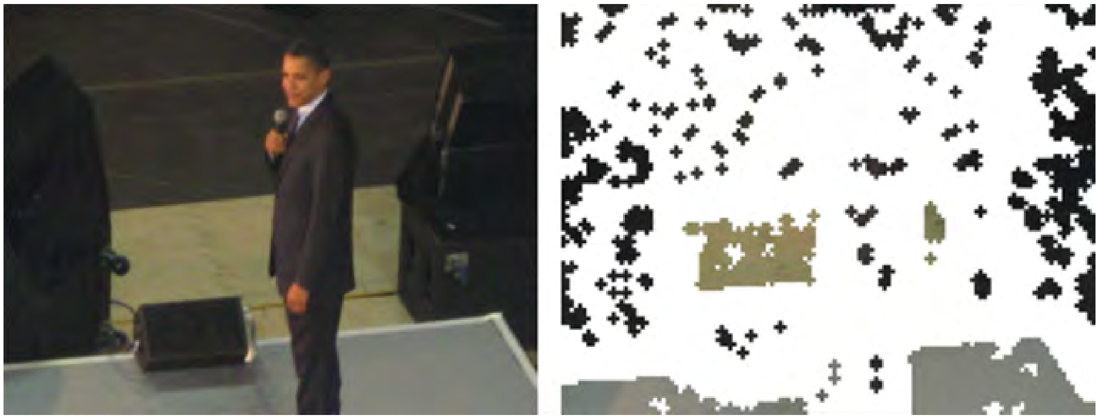
Noise in digital pictures has become a serious problem as the pixel size in the image sensors has been shrinking to allow for high megapixel counts. This fact is exacerbated by low quality optics and low quality sensors in camera-phones. Accurate modeling of the image noise at the device output is hard, due to various image processing steps such as local and global contrast enhancements, various non-linear transformations and compression [69].

Measuring noise in digital images is usually done as a two step process [69, 11]: first, the intensity-homogeneous blocks (*i.e.*, blocks with the lowest structure variation) of the image are identified; second, the noise is measured within each of these intensity-homogeneous blocks. In [11] they try to estimate the noise variance for the whole image. However, as [69] showed, in many cases of practical interest the noise is not spatially stationary over the image. The proposed approach performs an image matting step in order to identify the areas of the image with a certain degree of structure variation. This result is used to isolate the regions that have very little structure variation (see Fig. 5.4), where the pixels are assumed to be independent and identically-distributed (iid). Note that the signal in these intensity-homogeneous regions should be nearly constant such that the variation is mainly due to noise.

As local noise estimates should not be influenced by distant data samples, we calculate the noise variance within each intensity homogeneous region as the average of all variances calculated on a 3 × 3 support. In addition and in order to have a reliable estimate, we need to use sufficient data samples. Therefore, we only calculate the noise variance in intensity-homogeneous regions larger than a threshold size (*i.e.*, 320 pixels in the current implementation).

Since the SNR in the chrominance (C_b and C_r, in the YC_bC_r color space [52]) channels, *i.e.*, *chroma noise*, is typically lower than that of the luminance channel [76], we compute the noise variance in the C_b and C_r channels only, and keep the

Figure 5.4: Original noisy image (left), and luminosity-homogeneous regions, larger than 320 pixels each, (right).



maximum of the two per intensity-homogeneous region. The final noise measure –*lNoise*– is calculated by ordering the noise measures, based on their value, of all intensity homogeneous regions that were large enough for the task, and selecting the median value of the top 5, in order to avoid outliers.

5.3 Relevant region extraction

We define the *relevant region* of an image, as the region in the image that presents either the texture detail –sharpness–, saliency –contrast– and color purity –colorfulness–, or a combination of them. From an aesthetic standpoint, a consumer photograph should have at least one region with either of these three features –*i.e.*, at least a certain amount of texture in an object, or a salient object, or a colorful object– otherwise the image has no features on which to judge it, again, in the consumer photography arena –*i.e.*, artistic photographs may be judged aesthetically excellent based on very different grounds.

As noted above, our model is composed of an image segmentation first step, and for each of the segmented regions one single value of sharpness, one of contrast and one of colorfulness are kept. In this section we describe how these low level features can be combined in an additive way in order to calculate the main term of our aesthetic model. For this reason, first an *aesthetic appeal map* is created, which assigns an aesthetic appeal value to each of the pixels in the image, on a region by region basis. From this map, the *relevant region* is extracted via thresholding with an image adaptive threshold. Our model will measure on this *relevant region* the main term of our regression based image aesthetics model –see

Eq. 5.11.

5.3.1 Aesthetic appeal map

The sharpness, contrast and colorfulness[3] maps, along with the sharpness density of each region are combined in order to render an aesthetic appeal map (center image in Fig. 5.1), introduced by the author in [C15,C16]. In areas where there is texture/edge content, the sharpness metric is complemented in an additive way with a contribution from both contrast and colorfulness; in areas that are mostly soft –*i.e.*, they contain energy only in the low frequencies– the contribution of contrast and colorfulness is much larger, *i.e.*, this will increase the image aesthetic appeal measure of highly salient regions [59] or highly colourful regions with little texture/edge content. Note, though, that the presented aesthetic appeal map is different from the saliency map presented in [59] in which sharpness does not play any role, and where color contrast is taken into account as opposed to chroma magnitude. The aesthetic appeal map at each pixel (j,k) location, belonging to a particular *region*, is given by Eq. 5.8, 5.9 and 5.10.

$$AMap_{j,k} = S_{j,k} + \alpha(S_{j,k}) \cdot CN_{j,k} + \beta(S_{j,k}) \cdot cf_{j,k} \tag{5.8}$$

$$\alpha(S_{j,k}) = \begin{cases} \frac{1}{A+B \cdot SD_{region \supset (j,k)}} & \text{if } SD_{region \supset (j,k)} < SDThres \\ \frac{1}{E} & \text{if } SD_{region \supset (j,k)} \geq SDThres \end{cases} \tag{5.9}$$

$$\beta(S_{j,k}) = \begin{cases} \frac{1}{C+D \cdot SD_{region \supset (j,k)}} & \text{if } SD_{region \supset (j,k)} < SDThres \\ \frac{1}{F} & \text{if } SD_{region \supset (j,k)} \geq SDThres \end{cases} \tag{5.10}$$

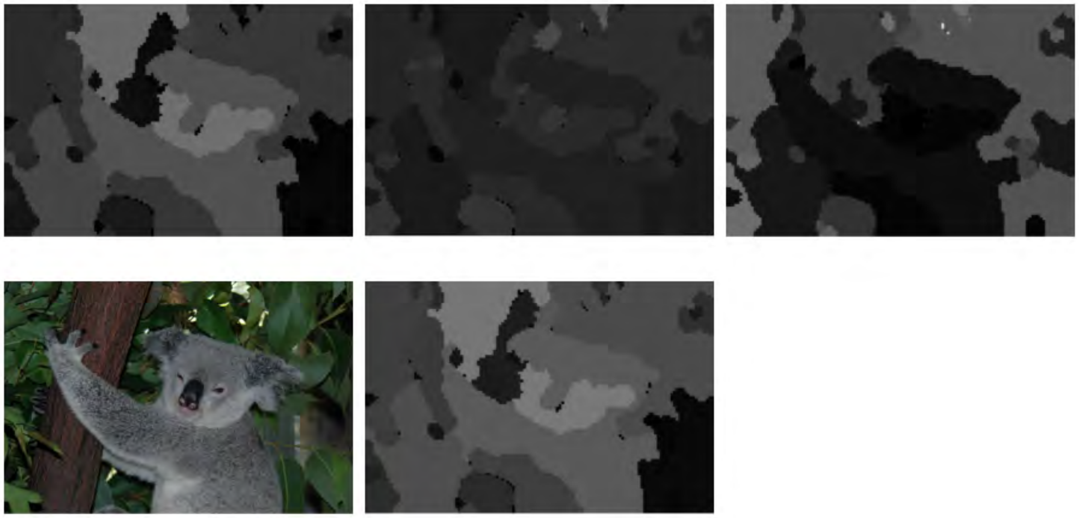
Where SD_{region} is the Sharpness Density for that particular region to which the pixel (j,k) belongs to; $SDThres$ is the sharpness density threshold that has been experimentally set to 0.33, *i.e.* regions with lower SD, will have increasing contribution from contrast and chroma in the region's aesthetic appeal. The other parameters are set to: $A = 2$, $B = 57$, $C = 2$, $D = 21$, $E = 2$ and $F = 9$.

By building this map in an additive way, regions with no sharpness content can still measure high in aesthetic appeal due to contrast or colorfulness contribution, and vice-versa.

In Fig. 5.5 we show an example of the original image, its region based sharpness map, its contrast and colorfulness map, and their combination into the aesthetic appeal map. Note that the out-of-focus foliage in the background increases its aesthetic appeal thanks to its colorfulness.

[3] As described in Section 5.2.3, we calculate the colorfulness of each region as $cf_i = 0.37\mu_{a_i b_i}$.

Figure 5.5: Aesthetic appeal map generation (bottom), from the region based sharpness map (top-left), the contrast map (top-center) and the colorfulness map (top-right).

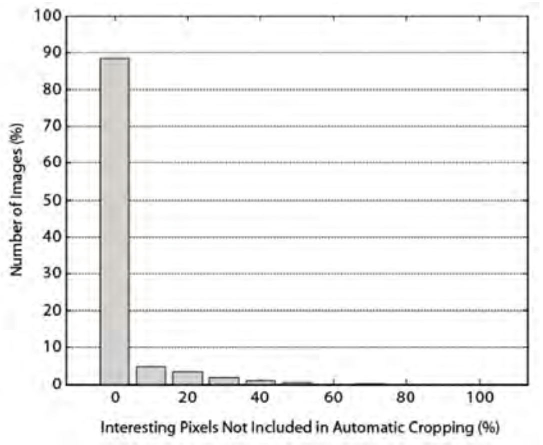


5.3.2 Relevant region calculation

Since different images may have very different aesthetic appeal value distributions, an image dependent threshold is calculated. This threshold is set to one half the maximum value in the aesthetic appeal map introduced above. All regions with the aesthetic appeal map value above that threshold will be set to aesthetically appealing, which generates a binary map –see right image, *c*, in Fig. 5.1. From now on, we define this aesthetically appealing region, as the *relevant region*. The most important term of our aesthetic appeal metric will be measured within this region, see Eq. 5.11.

An experiment with over 2000 consumer images and with 3 observers per image was carried out, where the observers, with diverse photography skills, were asked to delimit the aesthetically appealing regions of an image (using up to 3 rectangles). These regions were then intersected with the smallest rectangle that included the automatically generated aesthetically appealing region binary map. The results show (Fig. 5.6) that the automatically generated aesthetically appealing region binary map is efficient at including the user labelled aesthetically appealing areas of the photograph.

Figure 5.6: Images in which the automatically generated aesthetically appealing region does not include a certain number of pixels from the manually selected. Experiment run on 2000 images each rated by 3 users. Thanks to D. Berfanger for creating this figure.



5.3.3 Relevant region aesthetic contribution

Once the relevant region has been extracted, the aesthetic appeal map introduced in Section 5.3.1 is averaged over the whole relevant region. This provides the main aesthetic term of our approach, or Relevant Region Aesthetic Contribution (RRAC), as presented in Eq. 5.11.

$$RRAC = \frac{1}{N} \sum_{j,k \in relevantRegion}^{N} AMap_{j,k} \tag{5.11}$$

5.4 Multiplicative aesthetic factors

As mentioned above, from an aesthetic standpoint, a consumer photograph should have at least one region with either of these three features –*i.e.*, at least a certain amount of texture in an object, or a salient object, or a colorful object– otherwise the image has no features on which to judge it. This is why we define the main aesthetic term to be *RRAC*, presented in Eq. 5.11. In our model, this main term provides the aesthetic base measure for the image, which, in certain occasions, may have to be modified based on other low level features, *i.e.*, in the light exposure on the relevant region is either too dark –underexposure– or too bright –overexposure– a *multiplicative factor* will penalize the aesthetic value of the image accordingly, leaving it unchanged if the light exposure is within an allowed range of values; on the other hand, if the simplicity (see Section 2.0.2) of the image is high, then a *multiplicative factor* will increase –reward factor– the aesthetic value of the image.

From the individual low level features presented above in Section 5.2, a set of multiplicative factors, both reward factors, as well as penalty factors, are calculated. Each of these will multiply the main aesthetic appeal term, or *relevant region aesthetic contribution*, or *RRAC*, introduced Eq. 5.11 above. These factors include two exposure factors, a color contrast factor, a background homogeneity factor, and a noise factor, which will be described below.

Each multiplicative factor was fine tuned on a specific set of over 200 ground truth images. Each of these sets of images was built for training purposes only, and within each of them the specific low level feature taken into consideration varied through the whole possible spectrum of values in order to find the optimal multiplicative factor that correlated with human perception.

5.4.1 Exposure factors

Exposure is the total amount of light allowed to fall on the photographic medium (photographic film or image sensor) during the process of taking a photograph. Ansel Adams, the photographer, described in [7] the *zone system*, a system to obtain the proper exposure in order to maximize the quality of the image by determining where detail will appear in an image. In this work, we propose the use of histogram based low level features in order to come up with the aesthetic exposure factors. The histogram, which shows the concentration of tones, running from dark to light, can be used to extract overexposure and underexposure metrics.

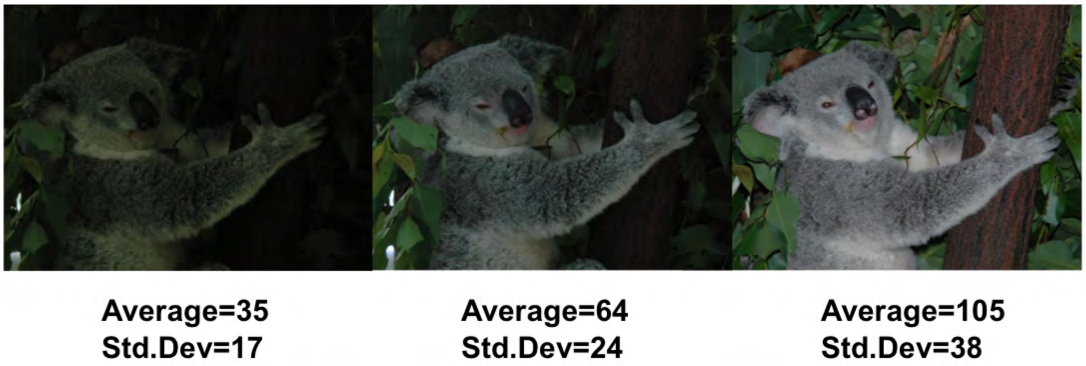
Overexposure and underexposure metrics have been implemented in order to penalize the overall image aesthetic appeal based on the luminance histogram distributions. No penalization exists if the histogram is not too skewed or clipped, and if its standard deviation is not too large or too small. A simple, yet effective model has been implemented, based on the average of the luminance histogram, $\bar{L}$, and its standard deviation σ_L, see Eq. 5.12 and 5.13. See Fig. 5.7 for three examples of the same subject photographed under three different lighting conditions.

$$luminanceFactor = \begin{cases} A_L + (1 - A_L) \cdot \frac{\bar{L}}{LL_{Thres}} & \text{if } \bar{L} < LL_{Thres} \\ 1 - B_L \cdot \frac{\bar{L} - LH_{Thres}}{255 - LH_{Thres}} & \text{if } \bar{L} > LH_{Thres} \\ 1 & else \end{cases} \tag{5.12}$$

$$contrastFactor = \begin{cases} A_{sL} + (1 - A_{sL}) \cdot \frac{\sigma_L}{sLL_{Thres}} & \text{if } \sigma_L < sLL_{Thres} \\ 1 - B_{sL} \cdot \frac{\sigma_L - sLH_{Thres}}{255 - sLH_{Thres}} & \text{if } \sigma_L > sLH_{Thres} \\ 1 & else \end{cases} \tag{5.13}$$

where $B_L = B_{sL} = 0.2$, and $A_L = A_{sL} = 0.8$. The thresholds were derived from our sets of 200 overexposed images, and 200 underexposed images: $LL_{Thres} = 70$, $LH_{Thres} = 160$, $sLL_{Thres} = 35$ and $sLH_{Thres} = 60$.

Figure 5.7: Exposure examples, with corresponding mean and standard deviation of luminance in the relevant region. Left: very bad aesthetic appeal; center: bad aesthetic appeal; right: good aesthetic appeal



The exposure factor, EF, is the product of the luminance factor and the contrast factor, see Eq. 5.14.

$$EF = luminanceFactor \cdot contrastFactor \tag{5.14}$$

5.4.2 Color factors

Color contrast –in a slightly modified formulation from the one described in Section 5.2.3, ccn_i– is computed on the entire relevant region ($CCN_{relevantR}$) and on the whole image ($CCN_{overall}$). The color contrast measure for a specific region i is presented in Eq. 5.15.

$$CCN_i = 0.75 + 0.5 * \sqrt{\frac{{\sigma_a}^2 + {\sigma_b}^2}{1400}} \tag{5.15}$$

The final color contrast measure is given by Eq. 5.16, which is the maximum between the color contrast calculated in the relevant region, and the color contrast calculated in the overall image. This allows to take into account both highly colorful backgrounds, as well as highly colorful aesthetically appealing regions.

$$CCNM = max(CCN_{relevantR}, CCN_{overall}) \tag{5.16}$$

Accordingly, a Color Contrast reward Factor (CCNF) is defined in Eq. 5.17.

$$CCNF = \begin{cases} 1 & \text{if } CCNM < 14 \\ \frac{33}{1000} \cdot CCNM - 0.53 & \text{if } 14 \leq CCNM < 26 \\ 1.4 & \text{if } CCNM \geq 26 \end{cases} \tag{5.17}$$

This factor rewards regions with high color contrast, but it does not penalize regions with low color contrast, as could be the case in black and white photography.

Colorfulness –also in a slightly modified formulation from the one presented in Section 5.2.3, cf_i– is defined below in Eq. 5.18. It is calculated on the overall image.

$$CF = 0.75 + 0.5 * \left[\sqrt{\frac{{\sigma_a}^2 + {\sigma_b}^2}{1400}} + 0.37 \sqrt{\frac{{\mu_a}^2 + {\mu_b}^2}{1400}} \right] \tag{5.18}$$

The colorfulness factor is defined in a similar way to CCNF, see Eq. 5.19.

$$CFF = \begin{cases} 1 & \text{if } CF < 14 \\ \frac{33}{1000} \cdot CF - 0.53 & \text{if } 14 \leq CF < 26 \\ 1.4 & \text{if } CF \geq 26 \end{cases} \tag{5.19}$$

5.4.3 Background homogeneity factor

Photo composition determines how objects should be arranged and balanced within the boundaries of the image. This regression based model avoids detecting compositional rules –we will be looking at compositional rules in our classification based scheme in Chapter 7. This regression based model does, on the other hand, look at one *simplicity* feature: how homogeneous the background of the photograph is, which is defined as how easy it is for a human subject to identify the subject of the photograph –*i.e.*, how isolated the relevant region is from the background– without distractions from other objects that may draw the viewer's attention. In [84] a visual test confirmed the importance of homogeneity. The way this measure is implemented is as follows: the aesthetic appeal map is thresholded twice, once with the image dependent threshold introduced above

($\frac{1}{2}$ of the maximum aesthetic appeal value), and once with one half of the first threshold ($\frac{1}{4}$ of the maximum aesthetic appeal value). With a lower threshold the aesthetically appealing region will expand to other regions of some intermediate aesthetic appeal value.

The more similar those two binary maps are the higher the background homogeneity measure, *i.e.*, if the generated binary map changes significantly, it means that there are distractions in the background, and the relevant region is not as easy to segment out from the background, therefore less homogeneous. Eq. 5.20 presents the euclidean distance between the centroids of each of the relevant regions –*i.e.*, thresholded at $\frac{1}{4}$ and at $\frac{1}{2}$ of the maximum value of the aesthetic appeal map.

The size of the relevant region is also important composition-wise. Region sizes either below $\frac{1}{9}$, or above $\frac{2}{3}$ of the image area are less aesthetically appealing. Therefore, only relevant regions –both the one thresholded at $\frac{1}{4}$, as well as the one thresholded at $\frac{1}{2}$– that are between these sizes are taken into account for the background homogeneity measure –*i.e.*, otherwise $HF = 1$.

$$CD = \sqrt{(C_{x_{1/2}} - C_{x_{1/4}})^2 + (C_{y_{1/2}} - C_{y_{1/4}})^2} \tag{5.20}$$

The reward background homogeneity factor is presented in Eq. 5.21, which only applies to distances between centroids below 15% of the image's width.

$$HF = \begin{cases} 1 & \text{if } CD > 0.15 \cdot ImWidth \\ 1.25 - 0.25 \cdot \frac{CD}{15} & \text{else} \end{cases} \tag{5.21}$$

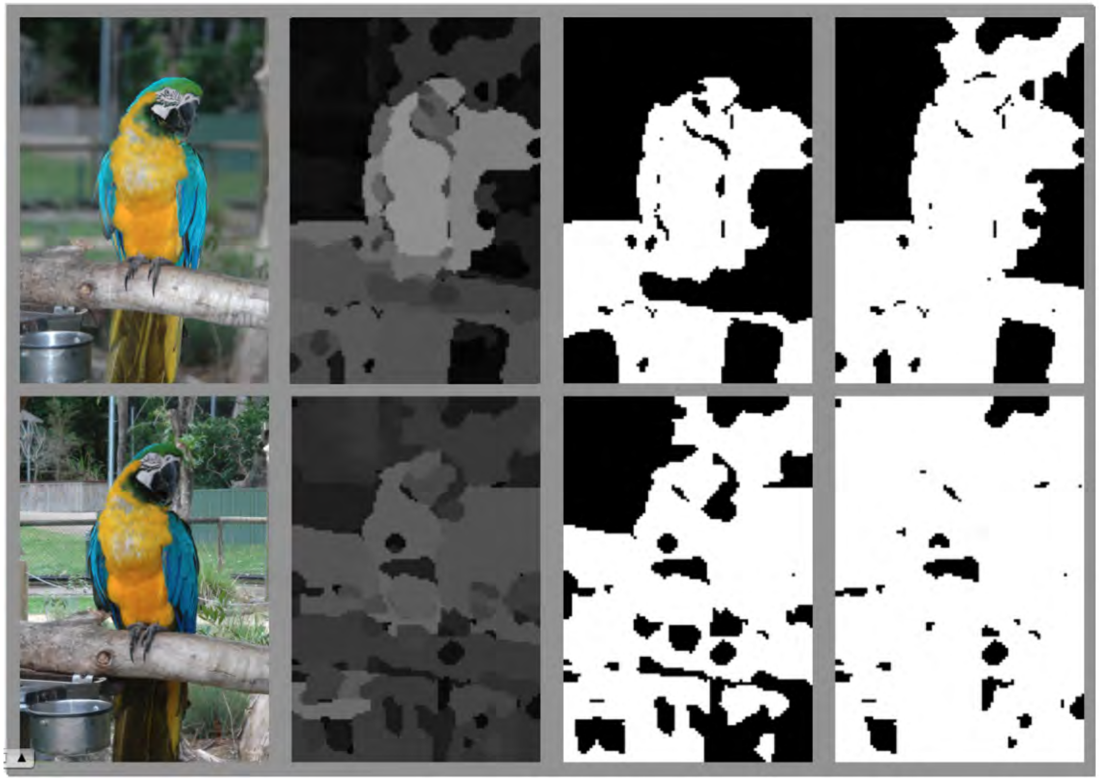
5.4.4 Noise factor

The noise penalty factor (NOF) is defined in Eq. 5.22.

$$NOF = \begin{cases} 1 & \text{if } lNoise < 1 \\ 1.235 - 0.235 \cdot lNoise & \text{if } 1 \leq lNoise < 3.125 \\ 0.5 & \text{if } lNoise \geq 3.125 \end{cases} \tag{5.22}$$

The NOF determines in what measure the overall aesthetic appeal measure will be decreased, with $lNoise$ as defined in Section 5.2.5. All thresholds and constants have been fine tuned over a 200 image training set of images with a wide range of noise levels.

Figure 5.8: Background homogeneity. In the top image, the background is quite homogeneously out of focus and low contrast, the foreground is in focus and high contrast, therefore the aesthetic appeal region thresholded at 1/2 and 1/4 do not change much, as opposed to the example in the bottom, which is less aesthetically appealing. Also the aesthetic appeal region of the one at the bottom, when thresholded at 1/4 is much larger than the size limit of 2/3 of the size mentioned above.



5.5 Image aesthetics models

The final image aesthetics model is comprised of a main term, *i.e.*, the *relevant region aesthetic contribution*, introduced in Section 5.3.3, $RRAC$, and a set of multiplicative factors, introduced in Section 5.4.

Two main models have been used throughout this dissertation, namely:

1. Model taking into account the colorfulness factor CFF, and no noise factor, see Eq. 5.23.

2. Model taking into account the color contrast factor $CCNF$ and the noise factor NOF, see Eq. 5.24;

$$AM_4 = RRAC \cdot EF \cdot HF \cdot CFF \tag{5.23}$$

$$AM_5 = RRAC \cdot EF \cdot HF \cdot CCNF \cdot NOF \tag{5.24}$$

See Table 5.3 to see the formulas for the first three models that were tested –*i.e.*, AM_1 throuth AM_3. In the models presented in Eq. 5.23 and 5.24, the first term is the most important one, *i.e.*, the average of the aesthetic appeal map over the relevant region, or $RRAC$. This factor is modified by penalty and reward factors [C11], where EF is the exposure factor, NOF is the noise factor, HF is the background homogeneity factor, $CCNF$ is the color contrast factor, and CFF is the colorfulness factor. Note that EF and NOF are penalty factors, *i.e.*, $EF < 1$ if relevant region is badly exposed –either overexposed or underexposed, $NOF < 1$ if noise is visible in the image; conversely HF, $CCNF$ and CFF are reward factors, *i.e.*, $HF > 1$ if background is homogeneous, $CCNF > 1$ if the relevant region or the overall image present high color contrast, and $CFF > 1$ if the overall image is colorful. Note that all factors have a Max and a Min cap –see Eq. 5.12, 5.13, 5.17 and 5.22 for examples.

These models were tested, see Section 5.6, and they were used in a set of applications that go from image search and retrieval re-ranking, see Chapter 9, to image aesthetics based storytelling, see Chapters 13 and 14.

5.6 Results

The features presented in the previous sections, and their parameters, have been combined in a few different ways in order to find the best model for different applications.

The test set for these experiments was obtained by evaluating consumer type digital images into broad image aesthetic appeal categories. 2000 images randomly selected from internet photo albums were graded, none of which had been used in the training/fine-tuning stages. Image noise was not an issue in this data set, and therefore the noise factor was not used at this point in time, *i.e.*, $NOF = 1$.

Observers graded the images using a categorical scale ranging from Very Poor to Excellent (rating them as integer values 1 through 6). These observers were instructed to evaluate the Overall Image Aesthetic Appeal of the images' main subject area, and not to consider neither the subject content, nor the image composition in their judgment. Each image was evaluated by three different observers, and the final ground truth value was set to the average of those 3 grades.

	$\frac{1}{N}\cdot\sum[S]$	$\frac{1}{N}\cdot\sum[\alpha(0)\cdot cf+\beta(0)\cdot cf]$	$\frac{1}{N}\cdot\sum[S+\alpha(S)\cdot CN+\beta(S)\cdot cf]$
$\cdot 1$	31.2/15.7	6.2/11.3	33.2/16.0
$\cdot EF$	33.5/16.7	13.3/9.0	35.3/15.6
$\cdot EF\cdot CFF$	34.0/14.8	8.4/9.2	33.2/13.1
$\cdot EF\cdot CFF\cdot HF$	29.6/15.0	X/X	28.0/13.5

Table 5.1: Linear correlation value / standard deviation of the linear regression residuals. Image aesthetic appeal measured on the whole image. Calculated on 2000 test images, each rated by 3 observers. "X" is a value not statistically significant.

	$\frac{1}{N}\cdot\sum[S]$	$\frac{1}{N}\cdot\sum[S+\alpha(S)\cdot CN+\beta(S)\cdot cf]$
$\cdot 1$	30.8/15.5	33.1/15.8
$\cdot EF$	33.1/16.5	35.3/15.3
$\cdot EF\cdot CFF$	33.6/14.6	33.0/12.9
$\cdot EF\cdot CFF\cdot HF$	29.2/14.8	27.8/13.4

Table 5.2: Linear correlation value / standard deviation of the linear regression residuals. Image aesthetic appeal measured on the relevant aesthetically appealing region, without taking into account the size of the relevant region in the calculation of the background homogeneity factor. Calculated on 2000 test images, each rated by 3 observers. "X" is a value not statistically significant.

A small number of outliers –mainly due to disagreements between observers as to what was the region of interest– were removed from the dataset. The inter-human observations linear correlations averaged to 56%, which shows how hard it is to agree on how to grade the image aesthetic appeal of a photograph.

The 2000 test images were analyzed by a thorough set of image aesthetic appeal measures. These aesthetic appeal measures can be subdivided into:

1. Whole image analysis (Table 5.1) vs. relevant region analysis (Table 5.2 and Table 5.3;

2. Size of the relevant region not taken into consideration in Table 5.2., vs. size of the relevant region used in Table 5.3.

For each experiment, the image aesthetic appeal results were normalized, and the linear correlation between these and the averaged ground truth data was calculated (first value in each table cell, in Table 5.1, Table 5.2 and Table 5.3). This correlation value indicates how well the image aesthetic appeal measure predicts the average ground truth (all correlation coefficients in the tables were

	$\frac{1}{N} \cdot \sum [S]$	$\frac{1}{N} \cdot \sum [S + \alpha(S) \cdot CN + \beta(S) \cdot cf]$
$\cdot 1$	33.4/13.8	AM_1=35.1/14.3
$\cdot EF$	36.1/14.5	AM_2=37.5/13.5
$\cdot EF \cdot CFF$	35.8/13.9	AM_3=33.7/11.8
$\cdot EF \cdot CFF \cdot HF$	29.9/13.1	AM_4=27.9/12.1

Table 5.3: Linear correlation value / standard deviation of the linear regression residuals. Image aesthetic appeal measured on the relevant aesthetically appealing region, taking into account the size of the relevant region in the calculation of the background homogeneity factor. Calculated on 2000 test images, each rated by 3 observers. "X" is a value not statistically significant.

statistically significant, except where noted). In order to measure the effectiveness of the presented metrics, a linear regression between the aesthetic appeal measure and the averaged ground truth was calculated, and the standard deviation of the residuals from this regression were measured (second value in each table cell).

The way to read the tables is the following –take for example the cell tagged as AM_4 in Table 5.3:

Take the formula in the 3^{rd} column's top cell:

$\left[\frac{1}{N} \cdot \sum [S + \alpha(S) \cdot CN + \beta(S) \cdot cf]\right]$ (from Table 5.3, column #2), and then multiply it by the left-most cell formula in row #4, $[\cdot EF \cdot CFF \cdot HF]$. This results in the following aesthetic appeal measure:

$$AM_4 = \frac{1}{N} \sum_{j,k \in relevantRegion}^{N} [S_{j,k} + \alpha(S_{j,k}) \cdot CN_{j,k} + \beta(S_{j,k}) \cdot cf_{j,k}] \cdot EF \cdot CFF \cdot HF \quad (5.25)$$

These results show the following preferences:

1. Always measure the features on the aesthetically appealing region –RRAC– only.

2. The best metrics all take into account the relevant region size when calculating the background homogeneity factor.

Inspection of these results shows the best algorithms in Table 5.3 –*i.e.*, highest correlation and lowest spread of residuals. In order to narrow down the best algorithm for selecting the highest aesthetic appeal images, another experiment was

Figure 5.9: (a) Average precision-recall (bottom 10% values retrieved), when querying for the lowest aesthetically appealing images of a collection. (b) Average precision-recall (top 10% values retrieved), when querying for the highest aesthetically appealing images of a collection. Averaged over 3 collections (overall 1200 images), none of them used in the training stage.

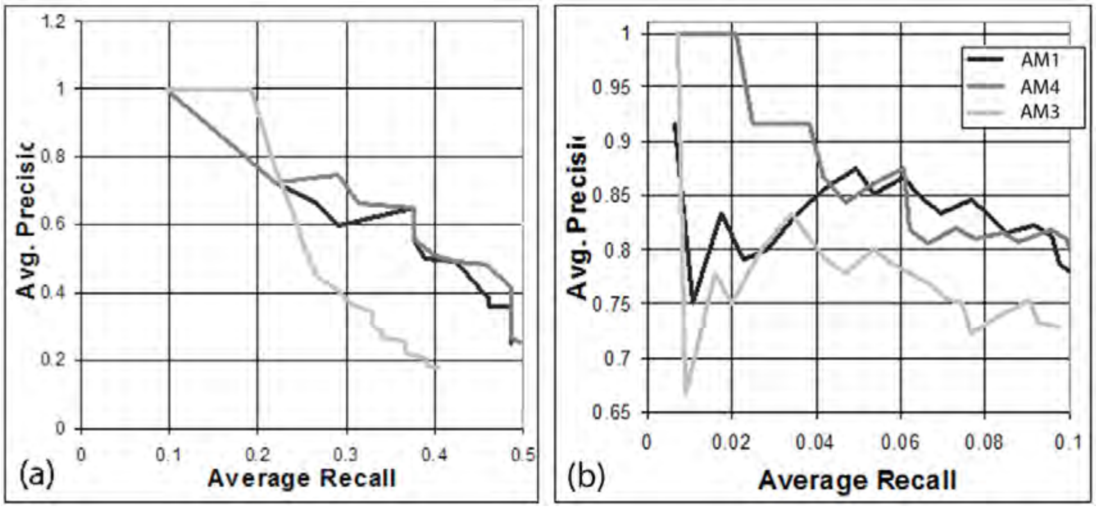
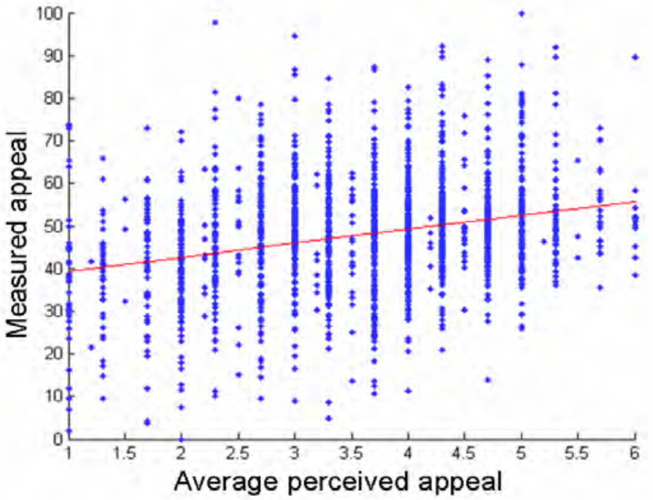
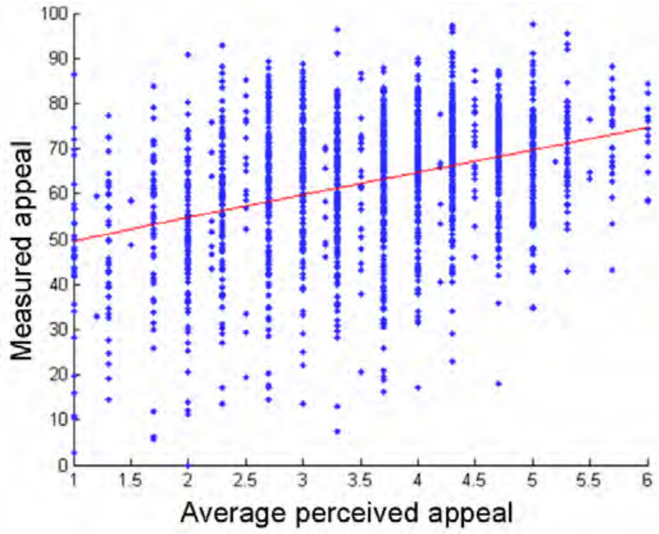


Figure 5.10: Correlation plot between Aesthetic Appeal Metric #4 (AM_4) and the averaged perceived aesthetic appeal for the 3 observers over 2000 images. The overall correlation coefficient is lower than in Fig. 5.11, but this metric is more precise (see Fig. 5.9b) when querying for highly aesthetically appealing images.



run where the images were ranked from worst to best measured aesthetic appeal, and the average recall and average precision were measured. The algorithms la-

Figure 5.11: Correlation plot between asthetic Appeal Metric #2 (AM_2) and the averaged perceived aesthetic appeal for the 3 observers over 2000 images.



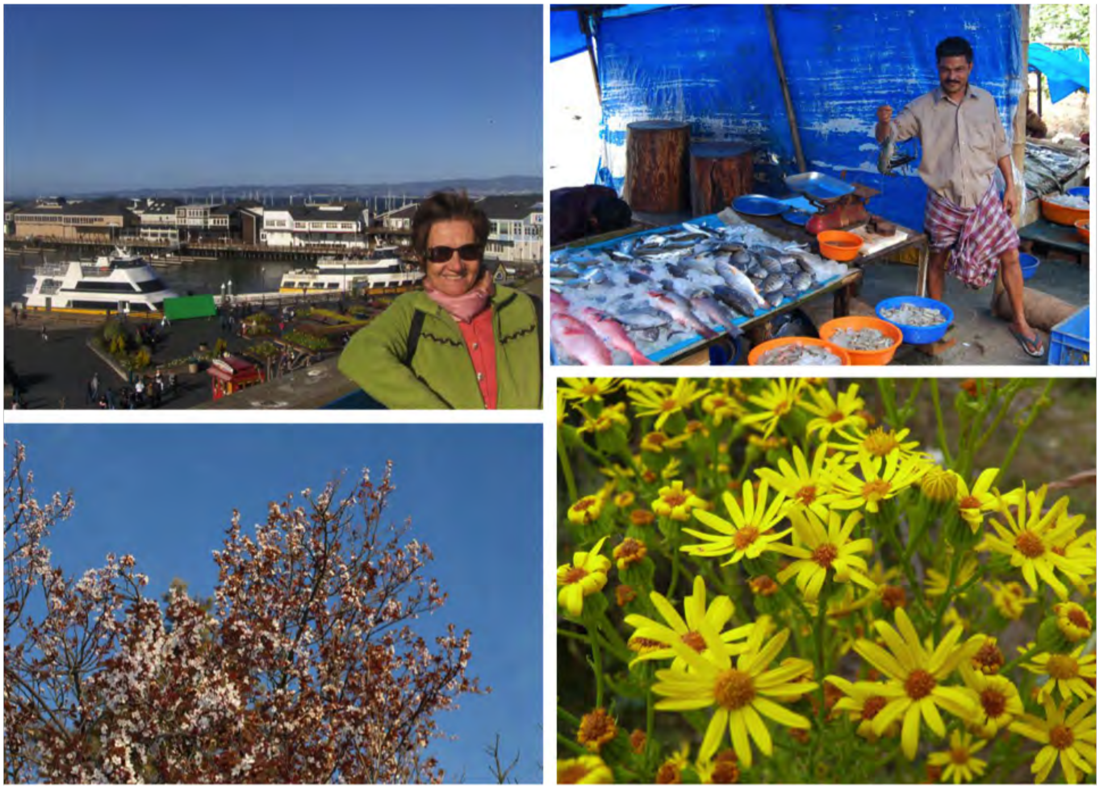
belled AM_i in Table 5.3 outperformed all others, more specifically, Fig. 5.9b shows AM_4 outperforming the other metrics in an average precision-recall graph when querying for the most aesthetically appealing images in a collection. Fig. 5.10 shows the correlation plot for AM_4.

Therefore, AM_4 has been successfully implemented as the image aesthetic appeal metric in the storytelling scenario –see Part II of this dissertation– where a good ranking of the most aesthetically appealing images in a collection is paramount. See Fig. 5.12 and 18.4 for a few examples of the most aesthetically appealing images from a sub-set of the image collections –each above 800 photos– that were used in the user study in Chapter 13.

AM_2, has the actual highest correlation value of 37.5% (compare to 56% for the inter-human correlation from our experiment), see Fig. 5.11. This would be the ideal aesthetic model to use when the whole range of aesthetic values –both high and low– are of interest.

Finally, note that the Noise factor has not been used in these experiments, as explained above. On the other hand, the Noise factor is quite useful in Chapter 9, in which all types of images, including those captured by cell-phones in very low light conditions, are used in the experiments.

Figure 5.12: Top aesthetic images in each of 4 collections used in the photo-book user study in Chapter 13. Since these were used as the photo-book wrap around cover, only landscape oriented images –*i.e.*, not portrait–were considered.

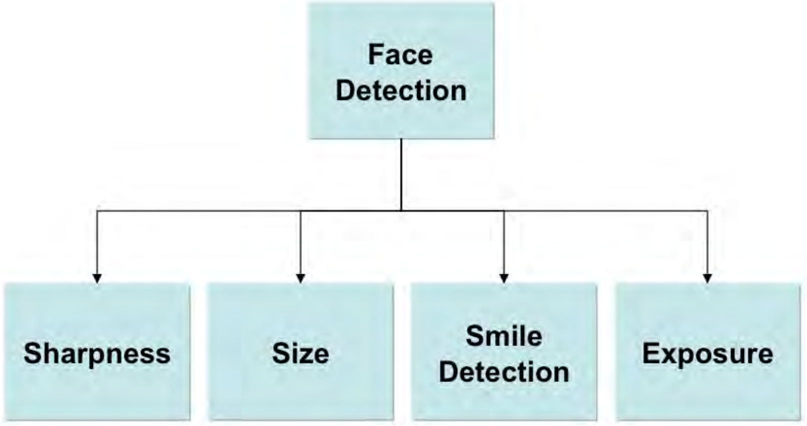


CHAPTER 6

Regression based face aesthetics

Photo based storytelling for consumer end users is, mostly, about events, vacations, etc., that the users enjoyed with family and/or friends. These experiences are then retold to those very same people, *i.e.*, remembrance, or to other people, *i.e.*, storytelling [C2]. Savakis et al. [101] showed in an in depth user study how important face aesthetics is, which includes both face image aesthetics as well as facial expression.

Figure 6.1: Face aesthetics framework.



There has been some research in trying to understand facial attractiveness [111] using face features including symmetry. Unfortunately, these type of approaches would favor a character over another based on their looks, which would go against the storytelling principles –see Part II of this dissertation. In order to avoid this kind of bias, in this chapter we present a regression based model in order to rank faces of the same person based on their aesthetic appeal.

We first describe the low level features we use in Section 6.1, followed by the description of the regression model in Section 6.2. This work was originally published in [C2 ,C11].

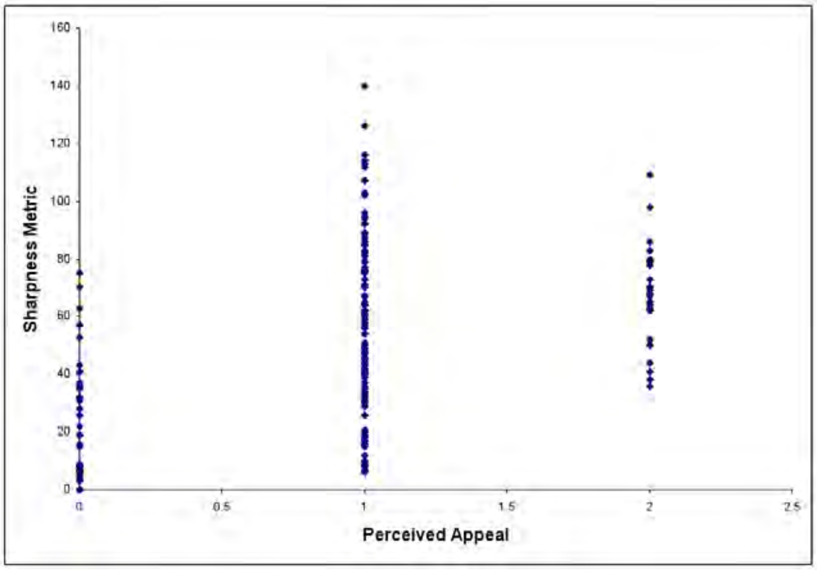
6.1 Face aesthetics features

Savakis et. al [101] already showed how important faces are when considering image aesthetics, specially in consumer photography. At the same time, people photographs are critical for consumer photo storytelling, as it will become clear in Part II of this dissertation. To this effect a series of face specific features have been implemented. In this implementation a face detection based on [118] has been used.

A set of 214 face images of the same person, covering all possible levels of face aesthetic appeal, have been used as a training set to derive the features below. These 214 images were tagged by two people –his parents– with different photography expertise level, in which they were asked to label the face aesthetic appeal (bad, 0, ok, 1, good, 2), average luminance (under-exposed, *ue*, slight under-exposed, *slightue*, good exposure, *ce*, slight over-exposed, *slightoe*, over-exposed, *oe*), and contrast (flat, lower contrast, good, over-contrasty), based on that persons face alone. We are aware of the low number of observers per image in this case, but the fact that they were the parents of the subject, resulted in a high engagement during the experiment.

Fig. 6.1 presents the different low level features that have been used in order to estimate the face aesthetic appeal metric.

Figure 6.2: Face sharpness –$fSharpness$– metric plotted against the face aesthetic appeal ground truth (bad, 0, ok, 1, good, 2). The linear correlation is 0.532.



6.1.1 Face sharpness

In the case of the face sharpness, we re-used Eq. 5.1 exactly as defined in Section 5.2.1. In the face aesthetic appeal case, sharpness is solely calculated within the face region. We are going to refer to this face sharpness measure as *fSharpness* for the rest of this chapter. See Fig. 6.2 to see how this measure correlates with the aesthetic appeal of faces in our ground truth set. The linear correlation between $fSharpness$ and the ground truth is 0.532.

6.1.2 Face light exposure

In a similar way to Eq. 5.12 and 5.13, both the average luminance, $\bar{L}$, and the luminance standard deviation, σ_L, are calculated in the face region. The ground truth gathered in our experiments –see Fig. 6.3 and Fig. 6.4– was used to optimize the constants. The resulting luminance and contrast metrics are presented in Eq. 6.1 and in Eq. 6.2 respectively.

Figure 6.3: Measured average luminance within the face region, plotted against the ground truth values from our 214 training images for under-exposure (ue), slight under-exposure (slight-ue), correct exposure (ce), slight over-exposure (slight-oe) and over-exposure (oe); the bars represent +/- one standard deviation.

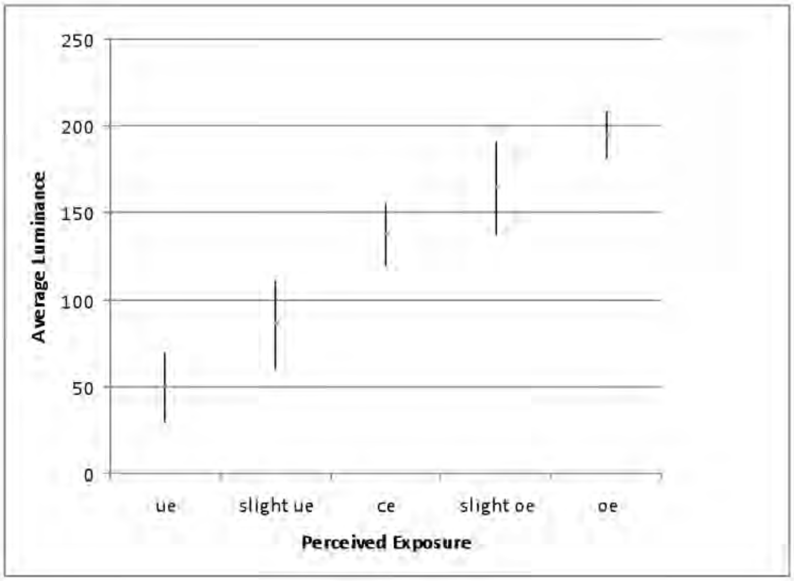
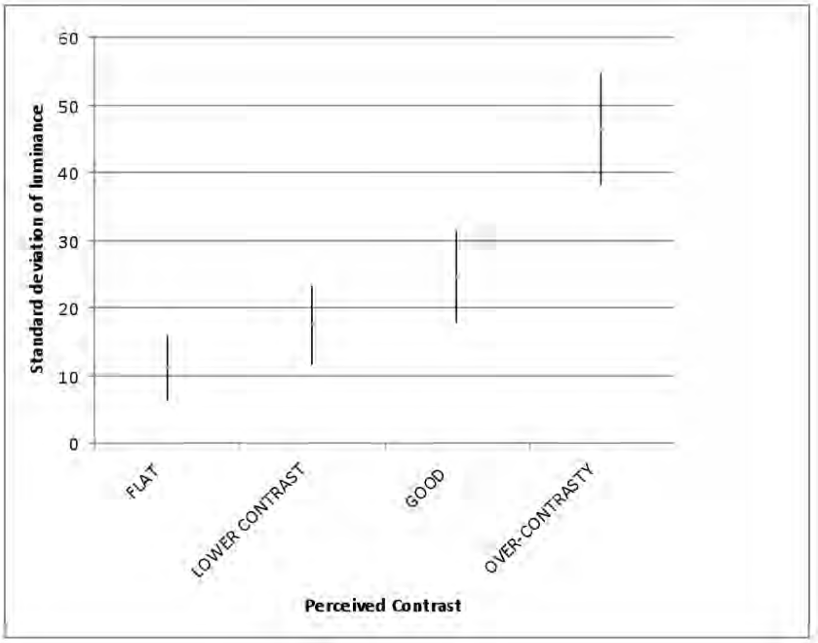


Fig. 6.3 and 6.4 plot the average luminance and the standard deviation of the luminance against the perceived face aesthetic appeal, respectively.

Fig. 6.5 plots the average exposure metric ($fLuminance$ in Eq. 6.1) against

Figure 6.4: Measured standard deviation of the luminance within the face region, plotted against the ground truth values from our 214 training images for very low contrast (FLAT), low contrast, good contrast and too much contrast (OVER-CONTRASTY); the bars represent +/− one standard deviation.



the perceived face aesthetic appeal.

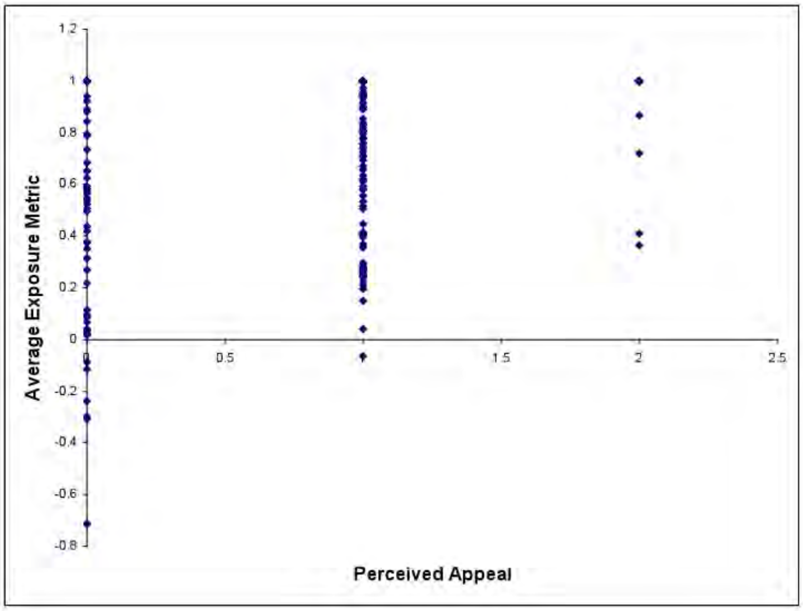
$$fLuminamce = \begin{cases} A_{fL} + (1 - A_{fL}) \cdot \frac{\bar{L}}{fLL_{Thres}} & \text{if } \bar{L} < fLL_{Thres} \\ 1 - B_{fL} \cdot \frac{\bar{L} - fLH_{Thres}}{255 - fLH_{Thres}} & \text{if } \bar{L} > fLH_{Thres} \\ 1 & else \end{cases} \tag{6.1}$$

$$fContrast = \begin{cases} A_{sfL} + (1 - A_{sfL}) \cdot \frac{\sigma_L}{sfLL_{Thres}} & \text{if } \sigma_L < sfLL_{Thres} \\ 1 - B_{sfL} \cdot \frac{\sigma_L - sfLH_{Thres}}{255 - sfLH_{Thres}} & \text{if } \sigma_L > sfLH_{Thres} \\ 1 & else \end{cases} \tag{6.2}$$

$B_{fL} = B_{sfL} = 1$, and $A_{fL} = fA_{sfL} = 0$, $fLL_{Thres} = 110$, $fLH_{Thres} = 155$, $sfLL_{Thres} = 10$ and $sfLH_{Thres} = 40$.

The linear correlation between $f`Luminance$ and the ground truth is 0.418, and the linear correlation between $fContrast$ and the ground truth is 0.193.

Figure 6.5: Face average luminance metric –*fLuminance*– metric plotted against the face aesthetic appeal ground truth. The linear correlation is 0.418.



6.1.3 Face size

Face size –*fSize*– is defined as the fraction of pixels that belong to the face region, detected as described above, with respect to the overall size of the image. This feature turns out to have a strong correlation with the perceived face aesthetic appeal of 0.379 (see Fig. 6.6).

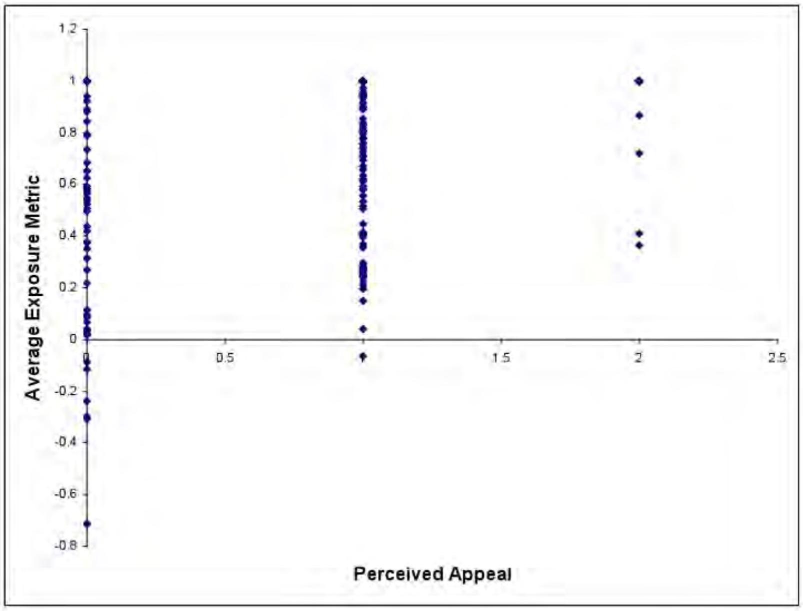
6.1.4 Smile detection

We use the smile detection algorithms presented in 3.3.3, which return a smile probability –the smile metric *fSmile*. Surprisingly, there is a very low correlation between smile probability and perceived face aesthetic appeal (see Fig. 6.7).

A detailed inspection of the data shows that faces that are quite small are usually discarded by the observers as not aesthetically appealing, regardless of whether the person is smiling or not. This allows for optimization of a cut-off size to allow for smile probability to be used effectively for face aesthetic appeal measurement. Faces larger than 3.6% of the image are used in the final implementation (see Fig. 6.8).

The linear correlation between the *fSmile* –when measured only on the faces larger than 3.6% of the image size– and the ground truth is 0.142.

Figure 6.6: Face size metric –$fSize$– metric plotted against the face aesthetic appeal ground truth. The linear correlation is 0.379.



6.2 Face aesthetics model

Our experiments in [C11] have shown that a linear combination of the metrics presented above is a good approach to calculate a face aesthetic appeal metric that will correlate highly with the perceived face aesthetic appeal. Inspection of Table 6.1 shows that the face sharpness metric –$fSharpness$– is the feature with highest correlation (see Fig. 6.2), therefore this has been used as the anchor feature in the results below. Table 6.2 shows the correlations of the different face appeal metrics that have been tested so far. Each linear combination coefficient has been optimized for highest correlation with the face aesthetic appeal ground truth.

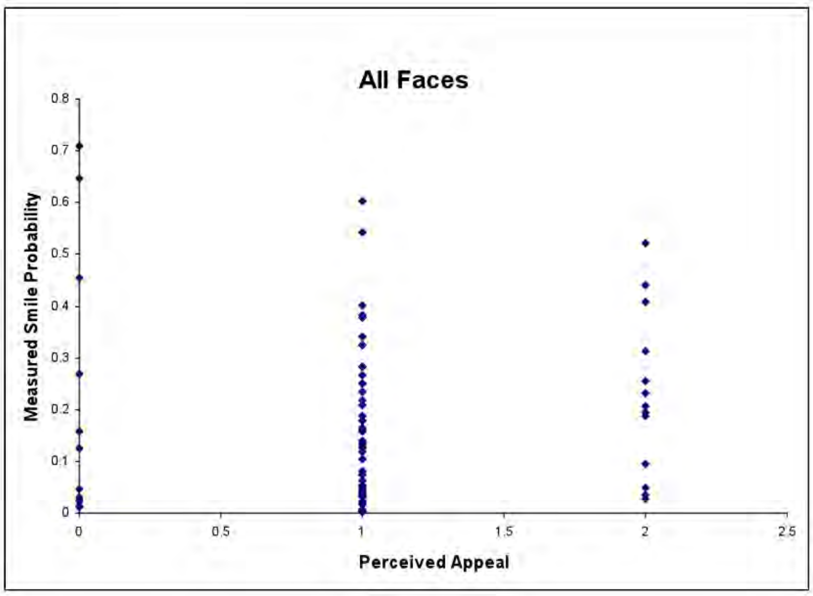
These results are in accordance with Table 6.1, where the correlation of each of these features with $fSharpness$ are shown. $FaceAppealMetric3$ ($FAMetric3$) is the best metric of the set, due mainly to the low correlation between the face size metric ($fSize$) and the face sharpness metric ($fSharpness$).

The general face aesthetics model, for a specific face j, is presented in Eq. 6.3.

$$FA(j) = \frac{1}{2}(fSharpness(j) + \psi fSize(j)) \tag{6.3}$$

where $\psi = 2.4$ for $fSize(j) < 0.42$, and $\psi = \frac{1}{fSize(j)}$ for $fSize(j) \geq 0.42$

Figure 6.7: Measured smile probability, $fSmile$, plotted against the perceived face aesthetic appeal.



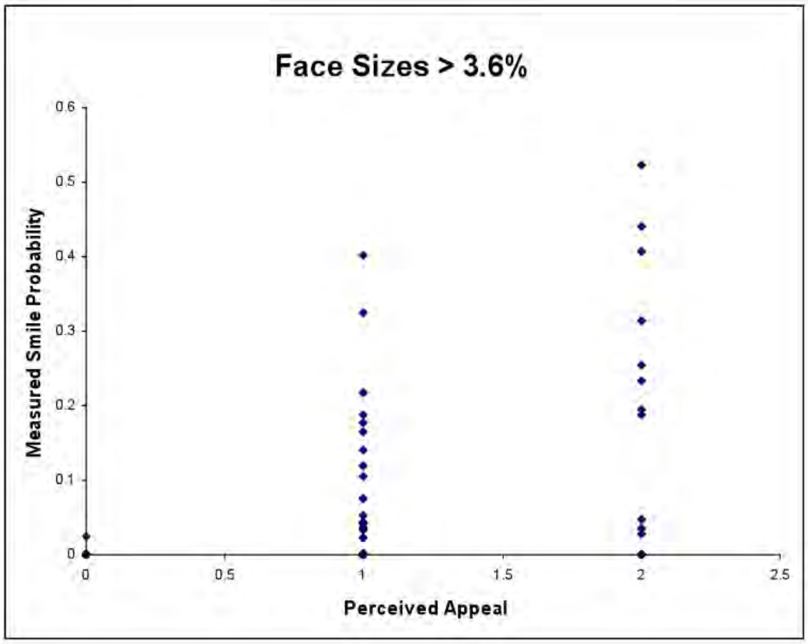
These results are corroborated in [88], where they use a sharpness measure of the face, and the ratio of the size of the detected face using [118] with respect to the overall size of the image, for face quality assessment .

In order to apply a face aesthetics regression model to the storytelling context, we carried out a small experiment in which we asked users to select certain images of themselves, as well as from friends and family, that could be used in a storytelling scenario. The findings were that face expression was a far more important feature than we had found in the aesthetics experiments presented above –where a single image was shown to the users and they had to rate the face aesthetic appeal without any task in mind. We found that faces with smiles have tremendous importance for storytelling purposes –*i.e.*, we hypothesize that users want to relay their happy memories.

Faces with very low face appeal ($FA(j) < 0.08$) are set to $fSmile(j) = 0$, whether it is because of low sharpness or very small size. This is due to our findings in storytelling experiments, where very unappealing faces are never selected, even if the person is smiling.

The final face aesthetics measure for storytelling purposes combines $fSmile$ with the face aesthetic appeal model from Eq. 6.3, as presented in Eq.6.4 for a specific face j.

Figure 6.8: Measured smile probability, $fSmile$, plotted against the perceived face aesthetic appeal, but only for faces larger than 3.6% of the image size, with a correlation of 0.142.



$$A_f(j) = \rho\, fSmile(j) + \tau\, \frac{1}{2}(fSharpness(j) + \psi fSize(j)) \qquad (6.4)$$

Where $\rho = 0.8$, $\tau = 0.2$, and ψ as defined above.

This model is used in Chapter 14. In Chapter 13, on the other hand, a simplification is used where only $fSmile$ is used, *i.e.*, $\rho = 1$ and $\tau = 0$.

	fSharpness	fLuminance	fContrast	fSize	fSmile, fSize>3.6%
Face aesthetic appeal ground truth	0.532	0.418	0.193	0.379	0.142
fSharpness	1.000	0.443	0.191	-0.017	0.177

Table 6.1: Linear correlation factors between the face based features that have been investigated and the perceived ground truth, in row #1. In row #2 the linear correlation between each face aesthetic appeal metric, and the face sharpness metric are listed.

	$FAMetric_1$	$FAMetric_2$	$FAMetric_3$	$FAMetric_4$
Face appeal formula	$fSharpness + 50 \cdot fLuminance$	$fSharpness + 30 \cdot fContrast$	$fSharpness + 305 \cdot fSize$	$fSharpness + 210 \cdot fSmile$, $(fSize > 3.6)$
Correlation	0.570	0.540	0.659	0.589

Table 6.2: Linear correlation factors between the face aesthetic appeal metrics ($FAMetric_i$) and the face aesthetic appeal ground truth.

CHAPTER 7

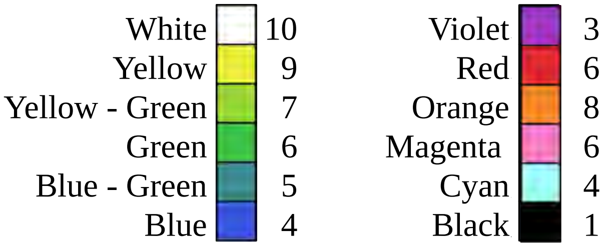
Classification based image aesthetics using visual composition features

In this chapter, we explore the role that visual composition plays in image aesthetic appeal classification. We propose low-level visual composition features that approximate traditional photography composition guidelines, such as *simplicity* and *visual balance* (*e.g. golden mean, golden triangles*). We then use these features to build an image aesthetics classifier that we test with a baseline dataset. Interestingly, our approach, that only takes into account visual composition features, yields close to state-of-the-art image aesthetic-based classification accuracies, which highlights the importance of image composition in image aesthetic appeal assessment. We originally published this work in [C4], and it was also protected by a patent application [C83].

Even though Savakis *et al.* [101] found visual composition to be the most important attribute when assessing image appeal, current computational approaches to image aesthetics have not analyzed features related to image composition in depth. In this chapter, we focus on understanding the impact that composition has on aesthetics by taking a close look at visual composition theory [46, 96, 41, 26], and proposing and computing low level features that relate only to the so called visual composition guidelines or rules. We then build a classifier that uses these composition based features to automatically classify images from a baseline image dataset [35] into high *vs.* low aesthetic appeal obtaining close to state-of-the-art image aesthetics classification performance –where the state of the art features are mostly non-composition based, confirming the importance of image composition in determining the aesthetics value of an image [101].

Our findings also suggest that visual composition plays an important role on determining the aesthetic value of an image. Even though aesthetics are highly subjective, some of the visual composition features presented in this chapter must be general enough to play a key role in automatically classifying high *vs.* low aesthetic appeal images.

Please refer to Chapter 2, in which we summarize the main rules from the photographic composition literature that are incorporated into our aesthetics model.



White	⬜	10	Violet	🟪	3
Yellow	🟨	9	Red	🟥	6
Yellow - Green	🟩	7	Orange	🟧	8
Green	🟩	6	Magenta	🟪	6
Blue - Green	🟦	5	Cyan	🟦	4
Blue	🟦	4	Black	⬛	1

Figure 7.1: *Relative brightness* color weights.

The chapter is structured as follows: The description of our approach is presented in Section 7.1. Our experimental results are presented in Section 7.2. Finally, Section 7.3 summarizes our conclusions and lines of future research.

7.1 Algorithm description

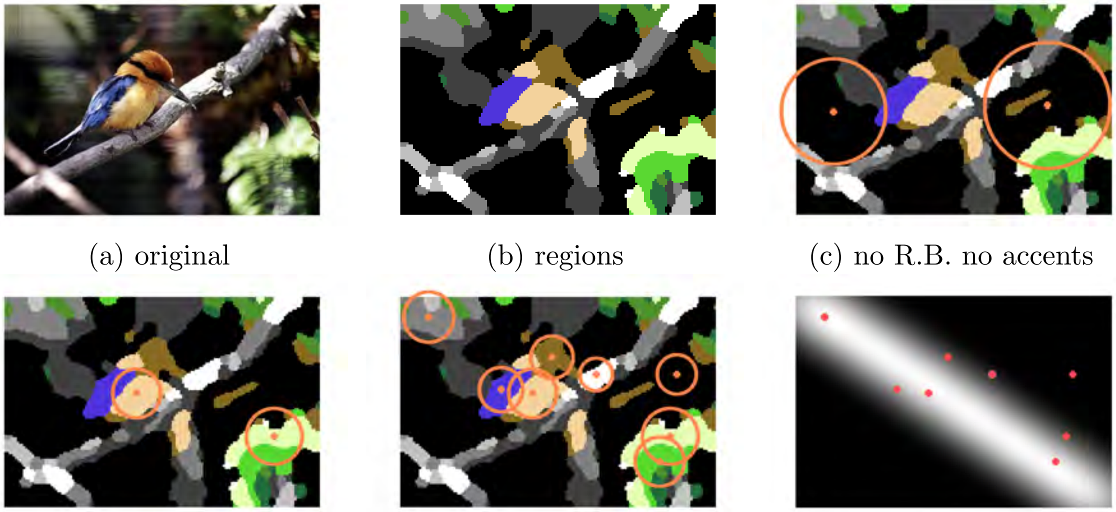
As previously described, identifying the position of the relevant subjects in the image is of paramount importance. Since the *relative brightness* of an image object, or region, is an important factor in order to determine its dominance within the image frame, we use a color image segmentation algorithm [C18]. In order to account for each region's dominance within the frame, we use the *relative brightness* table presented in [41, 46] (see Fig. 7.1) and interpolate the remaining weights using the color's brightness value (V, in Hue, Saturation, Halue (HSV) space [108]). The *relevance* of a region R_i is defined as the product of its size, in number of pixels, by its *relative brightness*:

$$relevance(R_i) = size(R_i) \times relativeBrightness(R_i)$$

There are other ways of detecting relevant subjects in an image, such as via face, people and other object detection algorithms. However, we decided to focus our work on the more general non-semantic case and hence we use color. Only relevant enough regions are taken into consideration by our approach, as we hypothesize these contribute to visual composition the most. We detect relevant regions in two complementary ways:

a) A region is selected if its *relevance* is above the threshold T_1, where T_1 is a percentage of the *relevance* of the region with highest *relevance* in the image. This ensures an image dependent selection of relevant regions, see Fig. 7.2d.

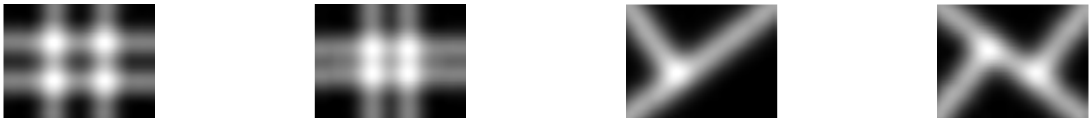
b) *Accent* regions are selected by inspecting the color bins from which no regions were selected above (*i.e.*, contrasting colors). We select the largest region of such a color bin if its size is above a threshold T_2, where T_2 is a percentage of the sum of all region's sizes within this color bin (Fig. 7.2e).



(a) original (b) regions (c) no R.B. no accents

(d) R.B., no accents (e) R.B. and accents (f) centroids layout

Figure 7.2: Relevant region extraction, taking size, *relative brightness* (R.B.) and *accents* into consideration (T_1=75, T_2=10). (f) Shows relevant centroids on template α_{29}, α_{51}. Photo by C. Pandino.



(a) α_{12}, α_{34} (b) α_{17}, α_{39} (c) α_{22}, α_{44} (d) α_{32}, α_{54}

Figure 7.3: Templates: (a) *rule of thirds*, (b) *golden mean*, (c) one of the *golden triangles*, (d) one of the *golden triangle* combinations.

We shall describe next the 55 composition-related features ($F_i, i = 1..55$) that we compute for each image in order to characterize its aesthetic value.

1. *Simplicity,* measured by 3 low-level features: the overall number of regions (F_1), as in [40], the number of relevant regions (F_2) and the number of *accent* regions (F_3). In this work we do not do any sharpness/frequency analysis to account for low depth of field situations, nor detect salient regions, and we defer these improvements to future work.

2. *Layout's Pleasantness,* or the overall *visual balance* without following a specific rule. It is modeled by measuring the homogeneity in the layout of the relevant regions in the scene, given by the average distance between centroids of the relevant regions, normalized by the image diagonal (F_4), and without normalization (F_6), together with their standard deviation

(F_5 and F_7). We also compute the normalized (F_8) and absolute (F_{10}) average distance between the centroids of the relevant regions minus the radii of the relevant regions, assuming a circular region of area equal to its size (*i.e.*, correlated with distances between regions borders), and their standard deviation (F_9 and F_{11}).

3. *Visual Balance*. In order to generate low level features that will correlate with the position of the relevant regions within the frame, we have devised a set of image dependent templates (*i.e.*, they adapt to different image aspect ratios). These templates are used to measure the relevant regions' compliance with the composition rules described in Chapter 2. We create a template for each specific *rule n*, by generating each of the rule's dividing lines individually, where l_n^i is the i^{th} dividing line for *rule n*, and convolving them with a 2D gaussian kernel with standard deviation σ. The dividing lines are combined – by adding all of them – creating thus the specific *rule*'s template α_n:

$$\alpha_n(x, y) = K \sum_{i=1}^{D} e^{-\frac{x^2+y^2}{2\sigma^2}} * l_n^i(x, y)$$

being D the number of dividing lines in the template, and K a normalization factor. After early experimentation, we found that $\sigma = L_{max}/20$, where L_{max} is the length of the image's longer side, generated an appropriate margin around the dividing lines and yielded satisfactory results. Fig. 7.3 depicts a few examples of the rules' templates –see Fig. 7.4, 7.5 and 7.6, for the full set of templates used in this research. Note that the templates are designed so that if a region's centroid lies close to a *power point*, it will have a much larger contribution than if it just lies close to a dividing line.

These rule-based *visual balance* features are calculated by adding up all the template contributions at each of the relevant regions' centroids; the gaussian introduced above allows for the degradation of the centroid contribution when it deviates from the *dividing lines* or *power points*:

$$F_n = \sum_{j=1}^{M} \alpha_n(Cx_j, Cy_j)$$

where F_n is the feature being considered, C_j are the coordinates of the j^{th} relevant region centroid and M is the number of relevant regions in the image.

For each of the rules, we extract features with the entire template, and also with each of the individual dividing lines in the templates –see Fig. 7.2f for an example of individual dividing line template. We do this in order to determine if any of them might have a stronger influence than the others, yielding 5 features for the *rule of thirds* (F_{12}-F_{16}, see Fig. 7.4), and also 5 features for the *golden mean* rule (F_{17}-F_{21}, see Fig. 7.5). In the case of the *golden triangle*'s rule, we generate templates for all rotations and symmetries, and also add the combination of the two golden triangle templates given a diagonal dividing line (see Fig. 7.3d), making up 12 features altogether (F_{22}-F_{33}, see Fig. 7.6). Hence, in total, we compute 22 template-based features and their normalized counterparts – where we normalize by dividing the feature values by the overall number of relevant regions (F_{34}-F_{55}).

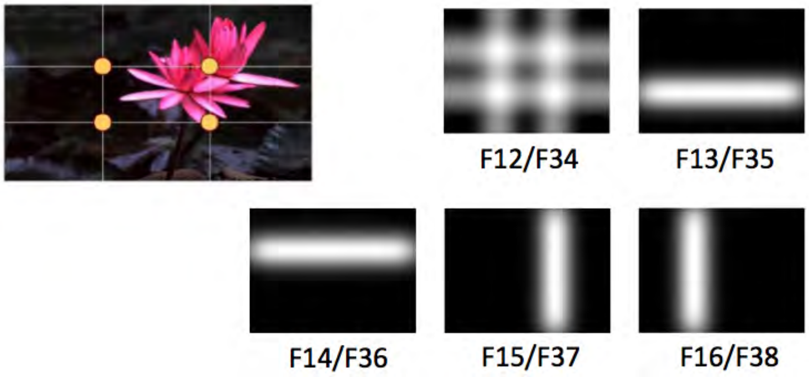


Figure 7.4: All templates used to represent the rule of thirds.

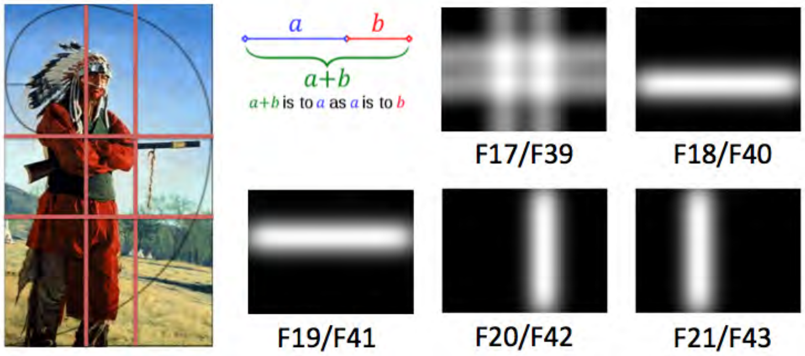


Figure 7.5: All templates used to represent the golden mean.

Note that from the proposed 55 features, only F_1 has been used in the literature prior to this work.

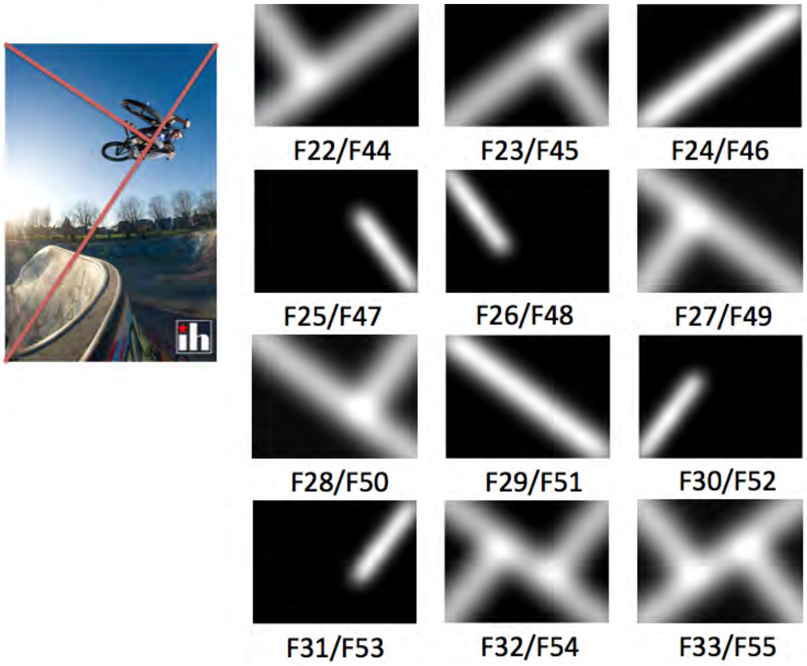


Figure 7.6: All templates used to represent the golden triangles.

Set	ProposedSet	DattaCompSet	LuoCompSet
8%	66.5%, (85, 25, 0, -1)	63%, (–, –, -4.5, 2.5)	61.1% (–, –, 1, 11.5)
16.5%	62.8%, (50, 40, -3.5, -1.5)	59.1%, (– ,– ,-3.5, -1.5)	57.8% (–, –, -1, 9)

Table 7.1: Classification accuracy of the compared feature sets. Table cells include: 5-CV accuracy, and the optimized parameters (T_1, T_2, SVM γ and $Cost$)..

7.2 Experimental results

We downloaded the same set of photos used in [35, 128] from Photo.net. A total of 3141 images were downloaded –*i.e.*, 20 less than in [128] since some users removed their images from the site. We perform two sets of experiments, one in which we train an SVM classifier with the top/bottom 8% of the images (538 images), and a second one with the top/bottom 16.5% (1040 images). In order to find the optimal combination of the proposed 55 composition features, we use a hybrid of filter-based and wrapper-based approach –similar to [35]. In all experiments reported below, we carry out a five-fold cross-validation with a standard RBF kernel for the SVM classifier, optimizing γ and Cost using the LibSVM package [24]), where the SVM is run 200 times for each of the low level features and their combinations.

In order to maximize the classification accuracy based on visual composition,

Set	AllCompSet	[35]	[128]
8%	69.3% (50, 10, -4, 0.5)	69.8%	78.4%
16.5%	65.8% (50, 50, -2, -2.5)	–	–

Table 7.2: Classification accuracy of the compared feature sets. Table cells include: 5-CV accuracy, and the optimized parameters (T_1, T_2, SVM γ and $Cost$). The results of the aesthetics algorithms presented in [35] and [128] are approximate readings from a graph in [128].

we experimented with all the possible components of our feature set (see section 7.1). All experiments were run with and without considering (1) the *relative brightness*, and (2) the *accent* component. In addition, for each of these four component combinations, a full grid search for both thresholds T_1 and T_2 was performed, in 5% steps. The best results were obtained by taking both the relative *brightness* and the *accents* into consideration.

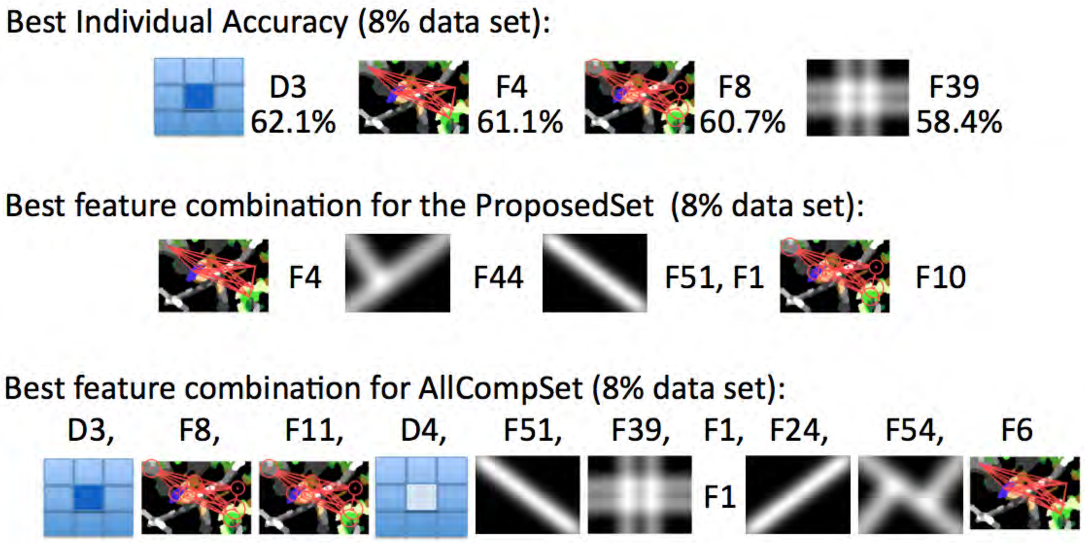

Best Individual Accuracy (8% data set):
D3 62.1%
F4 61.1%
F8 60.7%
F39 58.4%
Best feature combination for the ProposedSet (8% data set):
F4
F44
F51, F1
F10
Best feature combination for AllCompSet (8% data set):
D3,
F8,
F11,
D4,
F51,
F39,
F1,
F24,
F54,
F6
F1

Figure 7.7: Results of the proposed schemes. Features listed in order of importance, *i.e.*, left most important. Note that $D3$, and $D4$ are the Saturation and Value (HSV color space) averages within the inner rule of thirds rectangle respectively, as described in Chapter 4.

In order to compare with previous work, we have implemented two competing sets of composition features, see Chapter 4: *LuoCompSet*, with features L_1 and L_2 [83]; and *DattaCompSet* with D_1-D_4 [35]. As explained above, we leave the low depth of field features for future work, and therefore they were not implemented for this competitive study. Finally, we combined our proposed features with L_1-

L_2 and D_1-D_4 to generate the results presented under *AllCompSet* in Table 7.2. The best results are obtained in the 8% set with a 5-CV accuracy of 69.3%. The top 10 features in order of importance are: $D_3, F_8, F_{11}, D_4, F_{51}, F_{39}, F_1, F_{24}$, F_{54}, and F_6. Note that F_{24}'s template is the symmetrical of that of F_{29}, see Fig. 7.2f.

On the same 8% set, our top 6 *ProposedSet* features (see Table 7.1) in order of importance were: F_4, F_{44}, F_{51}, F_1, F_{10}, F_{54}, yielding an accuracy of 66.5%, higher than the state-of-the-art (*LuoCompSet* and *DattaCompSet*). When considering individual classification accuracy in the 8% set, with T_1=50, T_2=10, the four top features turned out to be: D_3 (62.1%), F_4 (61.1%), F_8 (60.7%) and F_{39} (58.4%).

See Fig. 7.7 for a visualization of the features that perform the best in the configurations described above.

7.3 Discussion

From the proposed feature set, we observe that the *pleasant layout* features (*e.g.* $F4$ and $F8$) are more important than the template-based ones. This might be due to the fact that we only consider the relevant region's centroid in our feature calculation, which might be inaccurate for larger regions. Coming up with a more sophisticated region descriptor is part of our future work. We also notice that the less relevant regions (below T_1=50% of the most relevant region) do not have a big impact in our composition features, confirming our hypothesis. Finally note that $D3$ and $D4$ are not strictly composition features, but they rather detect the saturation and the luminance of the central part of the image, which could explain how well they complement the proposed feature set.

As seen in Table 7.1 and 7.2, our proposed features perform better than the other composition features proposed in the literature. In addition, when combined together with these other features, the classification accuracy is similar to the results obtained by an aesthetics algorithm presented in [35], in which a very diverse set of aesthetics-related features were used, including exposure, colorfulness, saturation, hue, familiarity, texture, composition, color harmony and low depth of field.

CHAPTER 8

Video aesthetics

In this chapter, we tackle the problem of characterizing the aesthetic appeal of consumer videos and automatically classifying them into high or low aesthetic appeal. To the best of our knowledge, the work presented in this chapter, which we first published in [C5] and protected by a patent application [C82], represented the first effort to automatically characterize the aesthetic appeal of *consumer* videos and classify them into high or low aesthetic appeal. For this purpose, we first carry out a controlled user study (Section 8.1) to collect unbiased estimates of the aesthetic appeal of 160 consumer videos and thus generate ground truth. Next, we propose low-level features calculated on a per-frame basis, that are correlated with visual aesthetics (Section 8.2.1), followed by novel strategies to combine these frame-level features to yield video-level features (Section 8.2.2). Note that previous work in this area has simply used the mean value of each feature across the video [83], which fails to capture the video dynamics and the peculiarities associated with human perception [86]. Finally, we evaluate the proposed approach with the collected 160 videos, compare our results with the state-of-the-art (Section 8.3), discuss the implications of our findings (Section 8.4).

In this chapter, we focus on *building computational models of the aesthetic appeal of consumer videos*. Note that video aesthetic assessment differs from video quality assessment (VQA) [122] in that the former seeks to evaluate the holistic appeal of a video and hence encompasses the latter, as we described in Chapter 4. For example, a low quality video with severe blockiness will have low aesthetic appeal. However, a poorly lit un-distorted video with washed-out colors may have high quality but may also be aesthetically unappealing.

8.1 Ground truth data collection

Unlike in the previous chapter where we used images from online image-sharing websites, we decided to perform our own controlled user study in order to gather ground truth, in order to avoid the noise associated with the ratings in the datasets obtained from online image-sharing websites [10]. For instance, when rating a media object on-line, users are influenced in their aesthetic ratings by factors such as the artist who took the photograph or video, the relation of the subject to the photographer or videographer, the content of the scene and the context under which the rating is performed. Hence, a controlled study to collect aesthetic rating data is preferred over ratings obtained from a website. As noted in [36], web-based ratings are mainly used due to a lack of controlled experimental ground truth data on the aesthetic appeal of images or videos. In the area of image aesthetics, we shall highlight two controlled user studies – the one we presented in Chapter 5 and another one presented in [101]– even though neither of these datasets has been made public.

To the best of our knowledge, the only dataset in the area of video aesthetics is that used by Luo and Tang [83]. It consists of 4000 high quality (professional) and 4000 low quality (amateurish) YouTube videos. However, the authors do not explain how the dataset was obtained or how the videos were ranked. The number of subjects that participated in the ranking is unknown. It is unclear if the videos were all of the same length. Note that the length of the video has been shown to influence the ratings [93]. The content of the videos is unknown and since the rating method is undisclosed, it is unclear if the participants were influenced by the content when providing their ratings. Finally, the authors do not specify if the rated videos had audible audio or not. It is known that the presence of audio influences the overall rating of a video [16].

In order to address the above mentioned drawbacks and to create a publicly available dataset for further research[1], we conducted a controlled user study where 33 participants rated the aesthetic appeal of 160 videos[2]. The result of the study is a collection of 160 videos with their corresponding aesthetic ratings which was used as ground truth in our experiments. In this section, we detail how the videos were selected and acquired, and how the study was conducted.

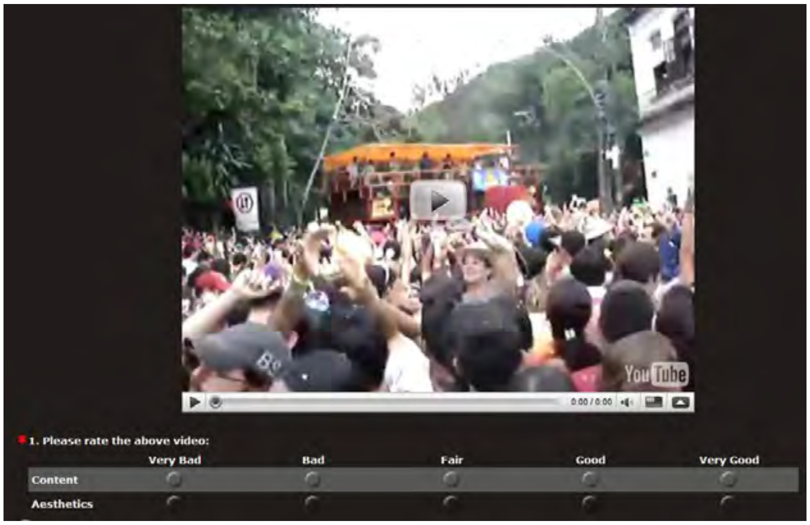
Video Selection: Since the focus of our work is consumer videos, we crawled the YouTube categories that were more likely to contain consumer generated content: Pets & Animals, Travel & Events, Howto & Style, and so on. To collect the videos, we used popular YouTube queries from the aforementioned categories (*i.e.*, text associated with the most viewed videos in those categories), for instance, "puppy playing with ball" and "baby laughing". In addition and in order to have a wide diversity of video types, we included semantically different queries that retrieved large numbers (>1000) of consumer videos, such as "Rio de Janeiro carnival" and "meet Mickey Mouse Disney". In total, we downloaded 1600 videos (100 videos × 16 queries). A 15 second segment was extracted from

[1] We have made available the video dataset along with the aesthetics ratings at http://mm2.tid.es/videoAestheticsUserStudy.

[2] Each video received 16 different ratings by a subset of 16 participants.

the middle part of each of the videos in order to reduce potential biases induced by varying video lengths [93]. Each of the 1600 videos was viewed by two video processing *experts* who rated the aesthetic appeal of the videos on a 5-point Likert scale. The videos that were not semantically relevant to the search query were discarded (*e.g,* “puppy playing with ball” produced videos which had children and puppies playing together or just children playing together); videos that were professionally generated were also discarded. A total of 992 videos were retained from the initial 1600. Based on the mean ratings of the videos – from the two sets of scores by the video processing *experts* after converting them to Z-scores [114], 10 videos were picked for each query such that they uniformly covered the 5-point range of aesthetic ratings. Thus, a total of 160 videos – 10 videos × 16 queries – were selected for the study. The selected videos were uploaded to YouTube to ensure that they would be available for the study and future research.

User Study: An important reason for conducting a controlled study is the role that content (*i.e.*, ”what” is recorded in the video) plays in video ratings. As noted in [36], the assessment of videos is influenced by both their *content* and their *aesthetic* value. We recognize that these two factors are not completely independent of each other. However in order to create a content-independent algorithm that relies on low-level features to measure the aesthetic value of a video, the ground truth study design must somehow segregate these two factors. Hence, our study required users to rate the videos on two scales: *content* and *aesthetics*, in order to reduce the influence of the former in the latter.



*1. Please rate the above video:

	Very Bad	Bad	Fair	Good	Very Good
Content	○	○	○	○	○
Aesthetics	○	○	○	○	○

Figure 8.1: User interface for the video aesthetics study.

A total of 33 participants (25 male) took part in the study. They had been recruited by email advertisement in a large corporation. Their ages ranged from

24 to 45 years ($\mu = 29.1$) and most participants were students, researchers or programmers. All participants were computer savvy and 96.8 % reported regularly using video sharing sites such as YouTube. The participants were not tested for acuity of vision, but a verbal confirmation of visual acuity was obtained. Participants were not paid for their time, but they were entered in a $USD 150 raffle. The study consisted of 30 minute rating sessions where participants were asked to rate both the *content* and the *aesthetic* appeal of 40 videos (10 videos × 4 queries). Subjects were allowed to participate in no more than two rating sessions (separated by at least 24 hours).

The first task in the study consisted of a short training session involving 10 videos from a "dance" query; the data collected during this training session was not used for the study. The actual study followed. The order of presentation of queries for each subject followed a Latin-square pattern in order to avoid presentation biases. In addition, the order in which the videos were viewed within each query was randomized. The videos were displayed in the center of a 17-inch LCD screen with a refresh rate of 60 Hz and a resolution of 1024 × 768 pixels, on a mid-gray background, and at a viewing distance of 5 times the height of the videos [20]. Furthermore, since our focus is *visual* appeal, the videos were shown without any audio [16].

Before the session began, each participant was instructed as follows. *You will be shown a set of videos on your screen. Each video is 15 seconds long. You have to rate the video on two scales: Content and Aesthetics from very bad (-2) to very good (+2). By content we mean whether you liked the activities in the video, whether you found them cute or ugly for example.*[3] *You are required to watch each video entirely before rating it.* We were careful not to bias participants toward any particular low-level measure of aesthetics. In fact, we left the definition fairly open in order to allow participants to form their own opinion on what parameters they believed video aesthetics should be rated on.

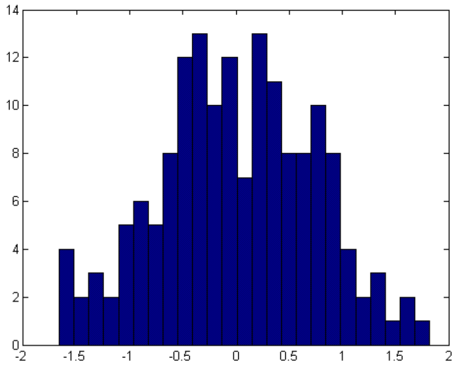
During the training session, participants were allowed to ask as many questions as needed. Most questions centered around our definition of *content*. In general, subjects did not seem to have a hard time rating the aesthetics of the videos. At the end of each query, participants were asked to describe in their own words the reasons for their aesthetic ratings of the videos. With this questionnaire, we aimed to capture information about the low-level features that they were using to rate video aesthetics in order to guide the design of our low-level features.

The study yielded a total of 16 different ratings (across subjects) of video aes-

[3] *Each video was embedded into the web interface with two rating scales underneath: one for* content *and the other for* aesthetics. *The scales were: Very Bad (-2), Bad (-1), Fair (0), Good (1), Very Good (2).*

thetics for each of the 160 videos. A single per-video visual aesthetic appeal score was created: First, the scores of each participant were normalized by subtracting the mean score per participant and per session from each of the participant's scores, in order to reduce the bias of the ratings in each session. Next, the average score per video and across all participants was computed to generate a mean opinion score (MOS). This approach is similar to that followed for Z-scores [114]. Thus, a total of 160 videos with ground truth about their aesthetic appeal in the form of MOS were obtained. Figure 8.2 depicts the histogram of the aesthetic MOS for the 160 videos, where 82 videos were rated below zero, and 78 videos were rated above zero. Even though 160 videos may seem small compared to previous work [83], datasets of the same size are common in state-of-the-art controlled user studies of video quality assessment [117].

Figure 8.2: Histogram of aesthetic MOS from the user study.

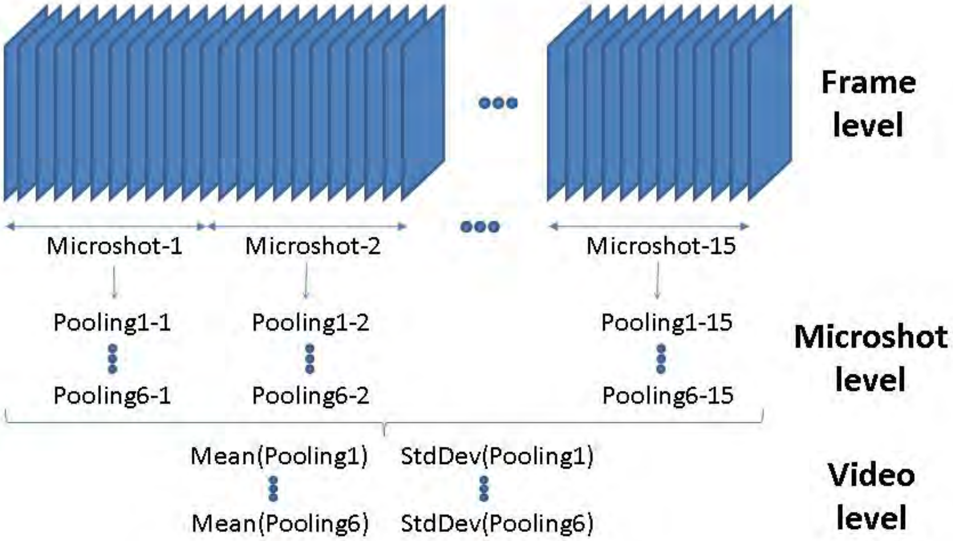


8.2 Feature computation

The features presented here were formulated based on previous work, the feedback from our user study and our own intuition.

The main difference between an image and a video is the presence of the temporal dimension. In fact, humans do not perceive a series of images in the same fashion as they perceive a video [122]. Hence, the features to be extracted from the videos should incorporate information about this temporal dimension. In this work, we propose a hierarchical *pooling* approach to collapse each of the features extracted on a frame-by-frame basis into a single value for the entire video, where *pooling* [86] is defined as the process of collapsing a set of features, either spatially or temporally. In particular, we perform a two-level *pooling* approach, as seen in Fig. 8.3. First, basic features are extracted on a frame-by-frame basis. Next, the

Figure 8.3: Proposed 2-level pooling approach, from frame to microshot (level 1) and video (level 2) features.



frame-level features are pooled within each microshot[4] using 6 different pooling techniques, generating 6 microshot-level features for each basic feature. Finally, the microshot-level features are pooled across the entire video using two methods (mean and standard deviation), thus generating a set of 12 video-level features for each of the basic frame-level features.

In the following sections we describe the basic frame-level features and their relationship (if any) to previous work, followed by the hierarchical pooling strategy used to collapse frame-level values into video-level descriptors.

8.2.1 Frame-level features

- Actual Frame Rate (f_1, actual-fps): 29% of the downloaded videos contained repeated frames. In an extreme case, a video which claimed to have

[4] In our implementation a microshot is a set of frames amounting to one second of video footage.

a frame-rate of 30 fps had an actual new frame every 10 repetitions of the previous frame. Since frame-rate is an integral part of perceived quality [122] – and hence aesthetics, our first feature, f_1, is the "true" frame-rate of the video. In order to detect frame repetition, we use the structural similarity index (SSIM) [120].

A measure of the perceptual similarity of consecutive frames is given by $Q = 1 - SSIM$ (small Q indicates high similarity), and is computed between neighboring frames creating a vector $\vec{m}$. To measure periodicity due to frame insertions, we compute $\vec{m^{th}} = \{ind(m_i) | m_i \leq 0.02\}$, where the set threshold allows for a small amount of dissimilarity between adjacent frames (due to encoding artifacts). This signal is differentiated (with a first order filter $h[i] = [1 - 1]$) to obtain $\vec{dm}$. If this is a periodic signal then we conclude that frames have been inserted, and the true frame rate is calculated as: $f_1 = fps \times \frac{MAX(\vec{dm})-1}{T_m}$, where T_m is the number of samples in $\vec{m}$ corresponding to the period in $\vec{dm}$. Note that this feature has not been used before to assess video aesthetics.

- Motion Features (f_2, motion-ratio, and f_3, size-ratio): The human visual system devotes a significant amount of resources for motion processing. Jerky camera motion, camera shake and fast object motion in video are distracting and they may significantly affect the aesthetic appeal of the video. While other researchers have proposed techniques to measure shakiness in video [130], our approach stems from the hypothesis that a good consumer video contains two regions: the foreground and the background. We further hypothesize that the ratio of motion magnitudes between these two regions and their relative sizes have a direct impact on video aesthetic appeal.

  A block-based motion estimation algorithm is applied to compute motion vectors between adjacent frames. Since the videos in our set are compressed videos from YouTube, blocking artifacts may hamper the motion estimates. Hence, motion estimation is performed after low-pass filtering and downsampling by 2 in each dimension, each video frame. For each pixel location in a frame, the magnitude of the motion vector is computed. Then, a k-means algorithm with 2 clusters is run in order to segregate the motion vectors into two classes. Within each class, the motion vector magnitudes are histogrammed and the magnitude of the motion vector corresponding to the peak of the histogram is chosen as a representative vector for that class. Let m_f and m_b denote the magnitude of the motion vectors for each of the classes, where $m_f > m_b$, and let s_f and s_b denote the size (in pixels) of each of the regions respectively. We compute $f_2 = \frac{m_b+1}{m_f+1}$ and $f_3 = \frac{s_b+1}{s_f+1}$.

The constant 1 is added in order to prevent numerical instabilities in cases where the magnitude of motion or size tends to zero. These features have not been used before to characterize video aesthetics.

- Sharpness/Focus of the Region of Interest (f_4, focus): Sharpness is of utmost importance when assessing visual aesthetics, as we described in Chapter 5. Note that our focus lies in consumer videos where the cameras are typically focused at optical infinity, such that measuring regions in focus is challenging. In order to extract the in-focus region, we use the algorithm proposed in [34] and set the median of the level of focus of the ROI as our feature f_4.

- Colorfulness (f_5, colorfulness): Videos which are colorful tend to be seen as more attractive than those in which the colors are "washed out" [55]. The colorfulness of a frame (f_5) is evaluated using the technique proposed in [55]. In Chapter 5 we showed how this feature was useful for image aesthetics.

- Luminance (f_6, luminance): Luminance has been shown to play a role in the aesthetic appeal of images [35]. Images (and videos) in either end of the luminance scale (*i.e.*, poorly lit or with extremely high luminance) are typically rated as having low aesthetic value[5]. Hence, we compute the luminance feature f_6 as the mean value of the luminance within a frame.

- Color Harmony (f_7, harmony): The colorfulness measure does not take into account the effect that the combination of different colors has on the aesthetic value of each frame. To this effect, we evaluate color harmony using a variation of the technique by Cohen-Or *et al.* [30] where they propose eight harmonic types or templates over the hue channel in the HSV space [108]. Note that one of these templates (N-type) corresponds to grayscale images and hence does not apply to the videos in our study. We compute the (normalized) hue-histogram of each frame and convolve this histogram with each of the 7 templates[6]. The peak of the convolution is selected as a measure of similarity of the frame's histogram to a particular template. The maximum value of these 7 harmony similarity measures (one for each template) is chosen as our color harmony feature. Other color harmony measures have been used to assess the aesthetic quality of paintings [73], and photos and video [83].

- Blockiness Quality (f_8, quality): The block-based approach used in current video compression algorithms leads to the presence of blocking artifacts in

[5] A video with alternating low and high luminance values may also have low aesthetic appeal.

[6] The template definitions are the same as the ones proposed in [30].

videos. Blockiness is an important aspect of quality and for compressed videos it has been shown to overshadow other artifacts [121]. The YouTube consumer videos from our dataset are subject to video compression and hence we evaluate their quality by looking for blocking artifacts as in [121]. Since this algorithm was proposed for JPEG compression, it is defined for 8×8 blocks only. However, some YouTube videos are compressed using H.264/AVC which allows for multiple block sizes [97]. Hence, we modified the algorithm in [121] to account for multiple block sizes. In our experiments, however, we found that different block sizes did not improve the performance of the quality feature. Therefore, in our evaluation we use the 8×8 block-based quality assessment as in [121] and denote this quality feature as f_8. We are not aware of any previously proposed aesthetic assessment algorithm that includes a blockiness quality measure.

- Rule of thirds (f_9, thirds): The *rule of thirds* is an important visual composition guideline, as presented in Chapter 2. This rule states that important compositional elements of the photograph should be situated in one of the four possible *power points* in an image (*i.e.*, in one of the four intersections of the lines that divide the image into nine equal rectangles, as seen in Fig. 2.2). In order to evaluate a feature corresponding to the rule of thirds, we utilize the region of interest (ROI) extracted as described above. Similarly to [83], our measure of the rule of thirds (f_9) is the minimum distance of the centroid of the ROI to these four points.

8.2.2 Microshot and video-level features

Once the 8 frame-level features (f_2 to f_9) have been computed on every frame, they are combined to generate features at the microshot level which are further combined to yield features at the video level. As mentioned above, in our implementation a microshot is a set of frames amounting to one second of video footage.

We compute 6 different feature pooling techniques for each basic frame level feature – *mean, median, min, max, first quartile (labeled as fourth) and third quartile (labeled as three-fourths)* – in order to generate the microshot-level features, and we let our classifier automatically select the most discriminative features. In this work we pool microshot-level features with two strategies in order to generate video-level features: *average*, computed as the mean (labeled as *mean*) of the features across all microshots; and standard deviation (labeled as *std*), again computed across all microshots in the video. Thus, a bag of 97 video-level features is generated for each video: 8 frame-level basic features × 6 pooling techniques at the microshot level × 2 pooling techniques at the video level + f_1.

In the remainder of this chapter, we shall use the following nomenclature: *videoLevel-microshotLevel-basicFeature*, to refer to each of the 97 features. For example, the basic feature *harmony* (f_7), pooled using the median at the microshot level and the mean at the video level would be referred as: *mean-median-harmony*. The use of these pooling techniques is one of the main contributions to the video aesthetics model. Previous work [83] has only considered a downsampling approach at the microshot level (at 1 fps), and an averaging pooling technique at the video level, generating one single video level feature for each basic feature, which cannot model their temporal variability.

8.3 Experimental results

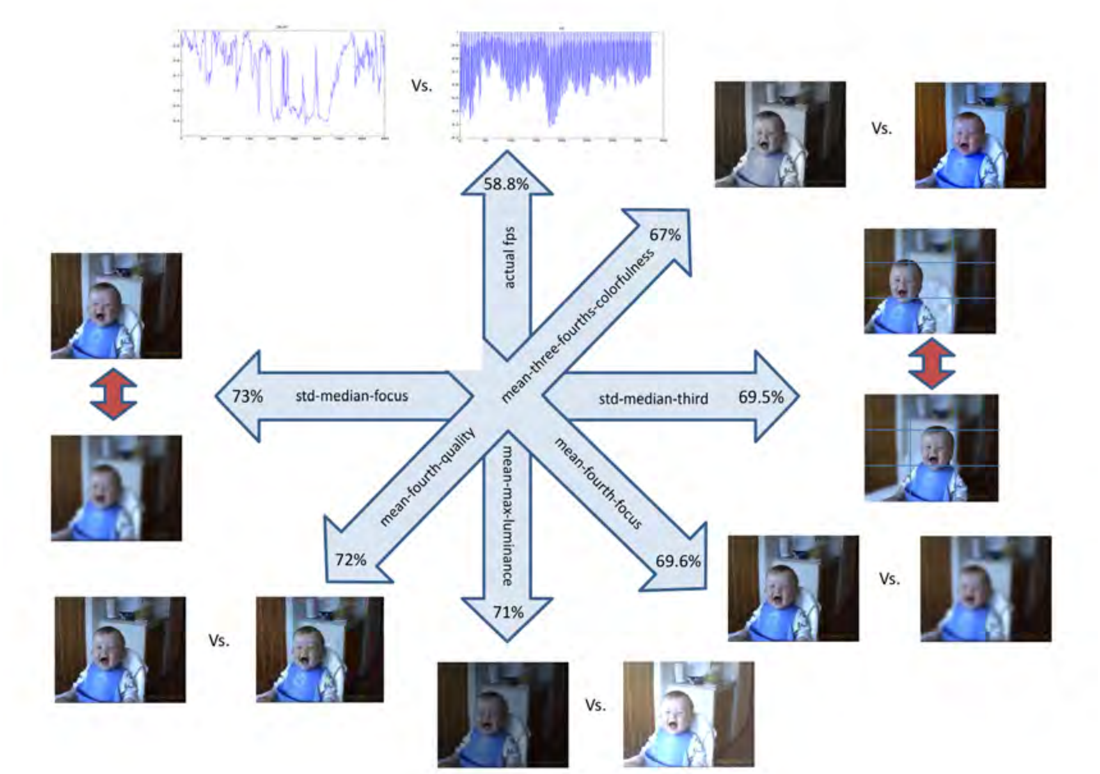
Even though one may seek to automatically estimate the aesthetic ratings of the videos, the subjectivity of the task makes it a very difficult problem to solve [36]. Therefore, akin to previous work in this area, we focus on automatically classifying the videos into two categories: aesthetically appealing *vs.* aesthetically unappealing. The ground truth obtained in our user study is hence split into these two categories, where the median of the aesthetic scores is considered as the threshold. All scores above the median value are labeled as *appealing (80 videos) and those below are labeled as* unappealing (80 videos). In order to classify the videos into these two classes, we use a support vector machine (SVM) [115] with a radial basis function (RBF) kernel $(C, \gamma) = (1, 3.7)$ and the LibSVM package [24] for implementation.

We perform a five-fold cross-validation where 200 train/test runs are carried out with the feature sets that are being tested. We first evaluate the classification performance of each of the 97 video-level features individually. The best performing 14 features in these cross-validation tests are shown in Table 8.1. The classification performance of these features is fairly stable: the average standard deviation of the classification accuracy across features and over the 200 runs is 2.1211 (min = 0.5397, max = 3.2779).

In order to combine individual features, we use a hybrid of a filter-based and wrapper-based approach, similar to [35]. We only consider the video-level features that individually perform above 50%. We first pick the video-level feature which classifies the data the best. All the other video-level features derived from the same basic feature and pooled with the same video-level pooling method (*i.e.*, either mean or standard deviation) are discarded from the bag before the next feature is selected. The next selected feature is the one that classifies the data the best *in conjunction with* the first selected feature, and so on. A 7-

dimensional feature vector[7] is thus formed. The selected features, in order of their classification performance after being combined with the previously selected features are: actual fps (acc=58.8%, $\sigma = 1.5$); mean-three-fourth-colorfulness (acc=67%, $\sigma = 1.8$); std-median-thirds (acc=69.5%, $\sigma = 1.9$); mean-fourth-focus (acc=69.6%, $\sigma = 2.2$); mean-max-luminance (acc=71%, $\sigma = 1.9$); mean-fourth-quality (acc=72.0%, $\sigma = 1.9$); and std-median-focus (acc=73.0%, $\sigma = 2.0$). See Fig. 8.4 for a visual interpretation of each feature and how they contribute to the overall classification accuracy.

Figure 8.4: Selected features in order of their classification performance after being combined with the previously selected features. The red double arrow indicates variation, *i.e.*, *standard deviation* feature, otherwise it is a *mean* feature.



An overall classification accuracy of 73.03% is thus obtained. In order to provide a comparison with previous work, we implemented the algorithm proposed in [83], achieving a classification accuracy of 53.5%. The poor performance of

[7] The feature vector is restricted to 7-dimensions due to the relatively small number of videos in the ground truth (160) and in order to prevent overfitting.

Feature	Accuracy
1. actual-fps	58.77%
2. mean-max-size-ratio	58.68%
3. std-fourth-motion-ratio	58.06%
4. mean-fourth-quality	57.67%
5. mean-three-fourth-colorfulness	56.86%
6. mean-max-colorfulness	56.80%
7. mean-max-quality	56.62%
8. mean-mean-colorfulness	56.34%
9. mean-med-colorfulness	56.21%
10. mean-mean-quality	55.73%
11. mean-three-fourth-quality	55.70%
12. mean-max-luminance	55.62%
13. std-three-fourth-motion-ratio	55.19%
14. mean-three-fourth-luminance	55.16%

Table 8.1: Individual classification accuracy of the top 14-features in descending order of performance.

this algorithm may be attributed to the fact that it was designed for professional *vs.* amateur video classification rather than for classifying consumer videos into high or low visual aesthetic appeal.

Personalization: Personalization involves using technology to accommodate the differences between individuals, using the individual's interests and/or past behavior. Personalization has not been explored before in the area of media aesthetics, even though it is known that certain aspects of aesthetic sensitivities depend on individual factors [36]. In this section, we carry out a preliminary analysis of the personalization of aesthetic ratings. Recall that two video processing *experts* rated the aesthetic value of 1600 videos. All videos which were semantically irrelevant or professionally generated were excluded from the analysis (608 videos or 38%). Video-level features were computed for the remaining 992 videos. Using the 7-dimensional feature vector previously described, we obtain classification accuracies of 61.66% (expert 1) and 58.17% (expert 2).

In order to evaluate the impact that personalization would have on this dataset, we select the optimum feature combination – using the approach described above – for each of the video processing *experts*. Tables 8.2 and 8.3 depict the selected features and their contributions to classification accuracy, yielding classification accuracies of 63.24% (*expert* 1) and 66.46% (*expert* 2), significantly larger in the case of *expert* 2 than the accuracies obtained with the *non-personalized* feature vector.

Expert 1		
Feature	Accuracy	StdDev
actual-fps	58.4%	0.1
+ mean-mean-quality	60.2%	0.3
+ mean-mean-size-ratio	61.2%	0.4
+ mean-fourth-harmony	62.3%	0.7
+ std-max-quality	63.2%	0.7
+ std-max-size-ratio	63.1%	0.7
+ mean-max-luminance	63.1%	0.8
+ std-fourth-thirds	63.2 %	0.9

Table 8.2: Classification accuracy with personalized feature vectors for video processing *expert* 1. Features selected for expert 1 and their contribution to accuracy - ’+’ indicates that the result was obtained by combining this feature with the one right above it.

Expert 2		
Feature	Accuracy	StdDev
mean-fourth-luminance	58.0%	0.2
+ mean-max-harmony	62.1%	0.5
+ std-max-quality	64.1%	0.6
+ mean-median-size-ratio	65.0%	0.5
+ mean-fourth-focus	66.0%	0.7
+ std-fourth-size-ratio	66.1%	0.6
+ mean-max-thirds	66.4%	0.6
+ std-mean-focus	66.5%	0.7

Table 8.3: Classification accuracy with personalized feature vectors for video processing *expert* 2. Features selected for expert 2 and their contribution to accuracy - ’+’ indicates that the result was obtained by combining this feature with the one right above it.

Aesthetics *vs.* Quality: As we mentioned in the introduction, *quality* does not capture all aspects of the aesthetic appeal of a video, but a holistic definition of aesthetics must include the quality of a video. In order to illustrate the role that quality plays on aesthetics, we evaluate the performance of the quality features – blockiness quality (f_8) and actual frames-per-second (f_1) – on the aesthetics classification. Hence, a *quality feature vector* is created by combining the actual fps measure (f_1) and the blocking quality pooling strategy that gives the best performance (mean-fourth-quality). This vector when used for classification yields an accuracy of 58.0%, which suggests that even though quality is an integral part of aesthetics, the aesthetic value of a video encompasses elements beyond traditional measures of quality. When adding the focus feature (f_4), arguably a quality feature also (particularly the *std-median-focus* feature) the overall performance increases to 60.0%, still well below the performance obtained when using the best performing 3 aesthetics features: 69.5%, as previously explained.

8.4 Discussion

Apart from the actual-fps feature (f_1), the rest of the features that were automatically selected to classify the aesthetic value of videos correlate well with previous research and intuition. For example, the third quartile of the colorfulness feature (f_5) would indicate that the maximum colorfulness value is probably noise, and the statistical measure of third quartile is a stable indicator of colorfulness. Again, the first quartile of the quality feature (f_8) correlates with research in image quality assessment [86]. Furthermore, quality features alone do not seem to capture all the elements that characterize the aesthetic value of consumer videos.

The standard deviation of the focus feature (f_4) is again intuitive in the sense that humans tend be more sensitive to changes in focus rather than its absolute value. This is also true for the rule-of-thirds feature (f_9), which is a measure of how well the main subject is framed in the video. Even though the motion features that we computed were not selected in the final feature vector, on their own these features performed well (see Table 8.1) and seemed to be useful for personalization (Table 8.2 and 8.3). Given that the number of videos in the personalization dataset is large and that motion features on their own seem to correlate well with perception, we hypothesize that increasing the number of videos in the current dataset (which we plan to undertake in the future) will result in a selection of the motion features as well.

CHAPTER 9

Application: aesthetics based image re-ranking user study

Current day image search systems are able to retrieve thousands or even millions of images relevant to a specific text query; they do a very good job by analyzing text around the images or within their captions –*e.g.* google images, or by analyzing associated tags and descriptions in social image sites –*e.g.* Flickr. But even in this situation there are still many images with very similar relevance at the top of the retrieved results. We hypothesize that re-ranking these images –*i.e.*, with the same relevance– based on their aesthetic value may improve the overall user experience –*i.e.*, the most aesthetic images retrieved at the top of that group with the same relevance. In the area of consumer images, aesthetics also plays an important role, since users are currently taking more and more images, many of which never see the light of day again –*i.e.*, information overload; in this situation, a way to select images based on their aesthetic value will also improve the end user experience.

In this chapter we analyze the role that query relevance and image aesthetics play in the decisions that users make when selecting images in a consumer image search task. In order to measure the impact of each of these factors, we have implemented an image re-ranking algorithm that takes into account the aesthetic appeal of the images retrieved by a standard image-search engine –Google's Picasa Web Album. The results of a controlled user study with 37 participants reveal that query relevance and type, image aesthetics and the presence of people in the photographs have an influence when searching for images. Therefore, and particularly in the context of consumer image search, traditional query-based search approaches tend to return a very large number of photos of varying levels of quality, with the same or very similar tags which may or may not be related to the image content. In these cases, we believe that there is a need for additional –human-centric– criteria to rank the images beyond query similarity and increase the user's satisfaction with the search results.

To the best of our knowledge, this was the first work that combined query-based image search with content-based image aesthetic appeal, at the time of publication in [C9], in an effort to understand the role that each factor plays in the selection of images in an image search task.

In the re-ranking application that we propose, we evaluate the fusion of tag relevance, obtained through a generic text based image search engine – picasaWeb.com– with the aesthetic appeal of each image, as described in Section 5.5, and more specifically, by using the model presented in Eq. 5.24. This specific model was used since a high percentage of the images retrieved from picasaWeb.com had been taken by low quality camera-phones, some of them in very bad lighting conditions; this favored this specific aesthetics model since it incorporates the noise factor NOF.

In order to combine the tag relevance with the aesthetic appeal of each image, we describe a *score aggregation function* that fuses both values together in an effective way, see Section 9.1. The user study is presented in Section 9.3. Its results and the implications for the design of consumer image search tools are described in Sections 9.4 and 9.5, followed by our conclusions and future work in Section 9.6.

9.1 Score aggregation function

Given an input query k, we propose the combination of two rankings or scores: relevance, provided by a standard text query-based search engine (R_k) and the result of ranking the images by their image aesthetic appeal (A_k), using Eq. 5.24 from Chapter 5. They are combined into a re-ranking of the retrieved images (FA_k).

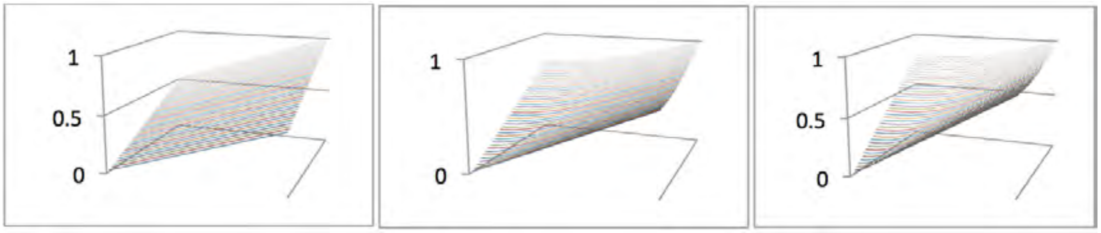
The search score fusion method provides a way for normalizing, combining and re-scoring lists of ranked search results by means of an aggregation function. The aggregation function determines how the scores from multiple ranked lists are combined, in order to obtain a final ranked list that includes information from each individual list [109].

Let $D_k(n)$ be the generalized score for item n in the results set corresponding to query k. The normalized score $D_{norm_k}(n)$ is given by Eq. 9.1:

$$D_{norm_k}(n) = \frac{D_k(n) - min\,(D_k(n))}{max\,(D_k(n) - min(D_k(n))} \tag{9.1}$$

Both the text-based relevance list, R_k, obtained through Picasa Web Album's search, and the aesthetic appeal score, A_k, are normalized with the previous formula. For each query, 1000 images are retrieved using the standard Picasa Web Album's search engine, such that the top ranked image will have a normalized score of 1 and the 1000^{th} ranked image will have a normalized score of 0. This generates the R_{norm_k} score. In addition, each of the retrieved images is analyzed, its aesthetic appeal calculated and they are re-ranked according to their aesthetic

Figure 9.1: Fusion aggregation formula (FA_k) with m=1 (left), m=2 (middle), and m=3 (right), as a function of the normalized aesthetic score A_{norm_k}, and the normalized text-based relevance R_{norm_k}.



appeal. Once normalized, a score of 0 is assigned to the least aesthetically appealing image whereas a normalized score of 1 is assigned to the most aesthetically appealing image, generating the $A_{norm_k}(n)$ score.

After normalization, the $A_{norm_k}(n)$ score gives the normalized aesthetic score of image n and query k, and $R_{norm_k}(n)$ is the normalized text-based score of image n and query k. We define an aggregation function, $FA_k(n) = f(A_{norm_k}(n), R_{norm_k}(n))$, that re-scores each item n, taking into account both the aesthetics and text-based scores.

Desired properties of the aggregation function include:

1. Images with highly relevant tags to the input query should receive a high score after the aggregation;

2. Highly aesthetically appealing images should receive a high score after the aggregation;

3. Images that are both highly relevant and aesthetically appealing should receive a higher score after the aggregation than in the 2 previous cases;

4. The weights given to relevance and appeal in the final ranking would be user and task dependent.

We propose a simple aggregation function, as a first approximation:

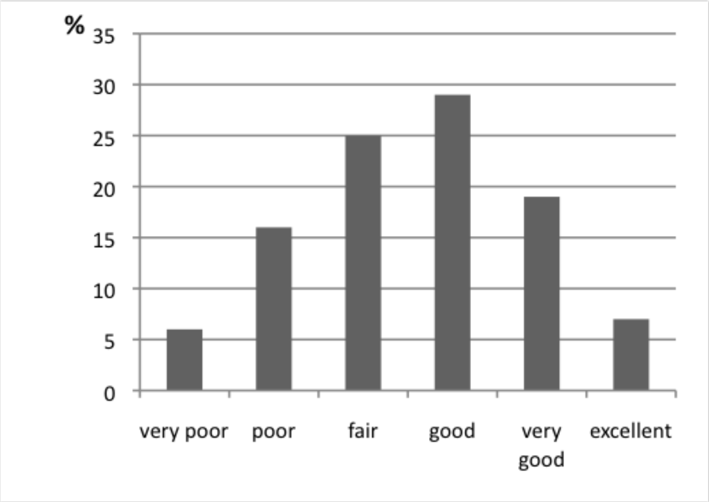
$$FA_k(n) = f(A_{norm_k}, R_{norm_k}) = [\alpha \cdot (A_{norm_k})^m + (1-\alpha) \cdot (R_{norm_k})^m]^{1/m} \quad (9.2)$$

The optimal settings of α and m depend on the user, the particular image collection and the task at hand. In the experimental results presented in this chapter, we use $\alpha = 1/2$, *i.e.*, aesthetics are as important as relevance. Dynamic optimization of this parameter was left for future research. In preliminary experiments, we implemented and validated three different aggregation functions, corresponding to $m = 1$, $m = 2$ and $m = 3$ –see Fig. 9.1 for a graphical representation of these 3 aggregation functions. After experimentation, it was found that the best combination of relevance and aesthetics was accomplished with $m = 2$, which would strike a balance between images that are either relevant or aesthetically appealing, and images that are both relevant and aesthetically appealing.

9.2 Image database

A large portion of the photographs existing in public photo sharing sites (*e.g.* Flickr), and even more so in the more professional sites (*e.g.* photo.net, DPChallenge.com), have different characteristics than the photographs found in a typical consumer image database. As described in [131], photo forum sites tend to include photographs that are more aesthetically appealing than those existing in personal collections. In addition, they rarely represent a full photographic session since photographers may only upload their favorite photos that can help differentiate themselves in the online community. Finally, photo forum sites users tend to rate mostly the photos they like [131]: in the case of photos with low aesthetic appeal, there may be very few or no ratings at all.

Figure 9.2: Histogram of 2000 ground truth images from 4 unedited photo collections, different photographers with different photographic skill levels, tagged by several observers.



Top 15 Tags in Flickr
Wedding, party, travel, family, beach, nature, vacation, friends, music, trip, birthday, Christmas, Flowers, summer, water.
Top 11 Geo-Tags in Flickr
Japan, London, California, Italy, USA, France, Paris, China, Europe, NYC, New York.
Top 15 Tags in Picasa Web Album
Wedding, trip, Christmas, party, family, birthday, vacation, park, beach, summer, people, city, lake, house, cruise.
Top 9 Geo-Tags in Picasa Web Album
Europe, Italy, USA, Paris, China, France, India, New York, Hawaii.

Table 9.1: Top tags in order of popularity.

We decided to create our experimental database with photographs from an online consumer image repository. The photos in such personal repositories typically have a wide range of aesthetic levels, ranging from very poor to excellent, peaking in the fair and good categories –see Fig. 9.2. After inspecting the image aesthetics distribution from a sampling of Picasa Web Album, we concluded that it closely resembled a typical consumer image database. Thus, we used Picasa Web Album photos to generate our image database. Note that in an ideal situation, we should have used the personal photos of each of our study participants. However, due to logistical limitations and privacy concerns, we were unable to do so.

In the following, we describe in detail the process that we followed to select the photos that would be included in the image database used in the study. In order to make the study manageable (and not too tiring) while covering a wide range of queries and topics, participants were asked to perform 10 image search queries: nine common or general queries to all participants and one personal query of their choice.

9.2.1 General queries

Nine of the queries performed by the participants in the study were general queries. They were designed to: (1) represent very popular queries executed in popular online personal photo sharing sites; and (2) retrieve as diverse a set of images as possible.

The most popular tags and geo-tags from Flickr and Picasa Web Album were extracted and analyzed, and the top tags were used to create the general queries (see Table 9.1). Note that we removed non-consumer related tags (*e.g. Nikon,*

Query	Query Text	#images (K)
1	Birthday party	13203
2	Inauguration of Barack Obama	107
3	Trip to Japan	3383
4	New York buildings	445
5	Wedding in the park	1351
6	Hawaii beach	510
7	Mountains of China	130
8	Summer in Paris	305
9	Vacation in Italy	630

Table 9.2: General queries and number of images retrieved in Picasa Web Album (in thousands), as of February 2009.

Canon, *Art*). Interestingly, the most popular tag in Flickr and Picasa Web Album is *wedding* with 15 and 140 millions of images retrieved, respectively.

We created eight of the final general queries by pairing one of the most popular general tags with another popular tag or geo-tag. The tags were very broad. Therefore, by combining two popular tags (or a tag and a geo-tag) we narrowed the semantic scope of the query and thus ensured that the retrieved images belonged to a coherent theme. In all cases, we made sure that the retrieved number of images on Picasa Web Album would be above 100k (*i.e.*, still a very popular query).

We also added a celebrity query, which could be a personal hero, public figure or artist that sometimes appear in consumer image collections (*e.g.* music concert or political rally events). An icon in the world, at the time of writing this dissertation, was President Barack Obama. In an effort to avoid emotional biases in the query (*i.e.*, republican vs. democrat), we selected an emotionally neutral query: "inauguration of Barack Obama".

Table 9.2 summarizes the nine general queries that we used in our experiments and the number of images (in thousands) retrieved by the Picasa Web Album search engine.

9.2.2 Personal query

The tenth query was formulated by each participant. We asked participants to create a query corresponding to one of these categories: daily life, locations, or nature, which have been reported to represent a large percentage of the image queries in the Microsoft Live search engine [131]. In addition, we made sure that all the personal queries retrieved at least 1000 images, so that the algorithms could be tested in the same conditions for all queries.

The distribution of the personal queries of the 37 participants in our study was: Locations: 22 queries (*e.g.* "Costa Rica", "Disney World"), Daily life: 8 queries (*e.g.* "walking", "sleeping"), Nature: 7 queries (*e.g.* "tulips", "lakes"), with a minimum number of pictures retrieved of 1.3k ("web surfing"), an average of 6400k, a standard deviation of 11600k, and a maximum of 37000k ("London") images. The average word length of the personal queries was 1.6 words.

The selected queries encompass a large image set that contained over 18 million photos about memorable events (birthday, wedding and inauguration) and locations (Japan, New York, Hawaii, China), both in urban (New York buildings) and natural (park, beach, mountains) settings, with and without people in them. Interestingly, in the set of images that were shown to the participants of our study –see Section 9.3– there were 20% more images with than without people in them.

9.2.3 Image database implementation details

After selecting the queries, we had to retrieve the associated images from the Picasa Web Album repository in order to populate the image database with actual photos. In addition to the images, we retrieved the normalized *order* of relevance (R_{norm_k} ranking) of each of the images as established by Picasa's retrieval algorithms. We used the Picasa Web Album Data API [2] to obtain, for each query, an XML-based result list which contains each of the retrieved images together with all their associated metadata. The ranking of the retrieved photos was used to generate the first ranking of our experiment: *Picasa*, as explained in Section 9.3.

Images were analyzed by the image aesthetics algorithm described in Section 5.5, with Eq. 5.24, generating a normalized aesthetics score, A_{norm_k}. This score was used to re-rank the retrieved images, hence generating the second ranking of our experiment: *Aesthetics*. Finally, the two previous rankings (*Picasa* and *Aesthetics*) were fused together using the aggregation function described in Section 9.1, generating the third ranked list, FA_k: *Fusion*. Once we had created the image database and associated rankings for each of the queries, we were ready to deploy the user study with real users.

9.3 User study

The research questions that the user study was designed to answer were:

- R1: Are users influenced by image aesthetics when searching for images in a *Consumer Image Search* setting?
- R2: What are the factors that play a role in determining the importance of image aesthetics in consumer image search tasks?

9.3.1 Participants

We carried out a controlled study with 37 volunteers (27 male) whose ages ranged from 23 to 49 years old (mean 30.6 years), 7 of them (19%) having one or more children. They were all computer literate and held a variety of occupations, including researchers, administrative assistants, engineers, accountants, infrastructure specialists, students, financers, people managers, front desk clerks and human resources specialists, from a diverse set of nationalities. All participants filled out an online demographic pre-study questionnaire that included questions about their digital picture taking habits and expertise.

We summarize next the findings of such questionnaire. The average monitor size of the users was 17.8". Fourteen users (38%) owned a subcompact camera (including camera-phones), 12 (32%) owned a compact camera and 11 (30%) owned an SLR (Single Lens Reflex, *i.e.*, with exchangeable lenses) camera. Thirty-one participants (83%) took pictures in order to capture memories and/or share those memories, 21 (57%) as a hobby, 16 (43%) for artistic reasons and 5 (14%) for work. Most of the users reported taking pictures every week (N=18; 49%), followed by those who took pictures every month (N=15; 40%). Interestingly, only 1 user reported taking pictures every day and 3 (8%) only took pictures occasionally.

In terms of their picture deleting practices, the highest portion of participants reported rarely deleting pictures (N=14; 38%), or sometimes (N=8; 22%), while 6 participants (16%) deleted pictures often and 8 (22%) very often. Only one participant reported never deleting any pictures. When asked about how often they edit/retouch their pictures in order to improve them, 15 participants (40%) never do it, 11 participants (30%) retouch 1 or 2 photos per photo session, 7 participants (19%) retouch 3 or 5 photos per session, and a small percentage of participants retouch 5-10 or more pictures per session, 3% and 8%, respectively.

With respect to the size of their digital image libraries, most of our participants had between 1,001 and 5,000 photos in their collections (N=15; 40%), followed by those who had between 5,001 and 10,000 photos (N=8; 22%). Almost the same number of users had more than 10,001 (N=7; 19%), or fewer than 1,000 photos (N=6; 16%). Our participants considered themselves to be average photographers (N=20; 54%).

When asked about their image search needs and habits, the largest portion of our participants searched for pictures in their personal collection on a monthly (N=14; 38%) or weekly basis (N=10; 27%), followed by those who searched less frequently (N=6; 16% every two months, and N=6; 16% every six months). In terms of their satisfaction with existing search technologies, the largest portion of our participants (N=16; 43%) is satisfied with the search results after inspecting the second page of results – *i.e.*, 40 images, followed by those who only look at the first page (N=7; 19%) – *i.e.*, 20 images. A small fraction of participants look beyond the second page (N=5; 13%) or have to search again with a different query once (N=6; 16%) or more times (N=3; 8%).

We also asked participants to rate their level of familiarity with the depicted locations in the test set in a 5-point Likert scale (1: not familiar at all, to 5: very familiar). The average familiarity score was 2.7 (std=3.9), ranging from the most familiar locations (Paris and Italy, mean=3.4 and 3.1; std=1.9, respectively) to the least familiar location (Hawaii, mean=1.8, std=1.1).

Finally, only 11 participants (29%) reported adding tags to their photos. In such cases, the majority of participants (56%) only annotated 21% of the photos. These user statistics are consistent with those previously reported in the literature of personal photo management and search [101, 9].

9.3.2 Apparatus and task

We conducted the user study using the Safari web browser on a MacBook computer, on a 17 inch DELL LCD display, set at 1280x1024 pixels in 32 bit color.

Upon arrival, participants were described the task that they had to do: they would be presented with the top 15 results of executing ten image search queries, one at a time; they would have to inspect each image at full screen resolution, and then select the 3 best photos of the result list in response to each of the ten queries.

Before starting the experiment, all participants did a trial query in order to learn about the query interface and task. Fig. 9.3 depicts the query interface corresponding to query Q5:"wedding in the park". Once they felt comfortable with the study (typically after 5 to 10 min), they started with the first query. The duration of the study was on average 46 minutes (std=9 minutes). Participants took all the time they needed to select the 3 pictures and all of them were able to complete the task. In addition to selecting the images, participants were asked to enter the reasons for their selection in a text box at the bottom of the page (see Fig. 9.3).

All the computer interactions were logged and a video-camera recorded the

Figure 9.3: User study user interface for query #5: *"wedding in the park"*. Only 13 images are presented, since there were 2 collisions: images that appeared in 2 treatments within the top 5.

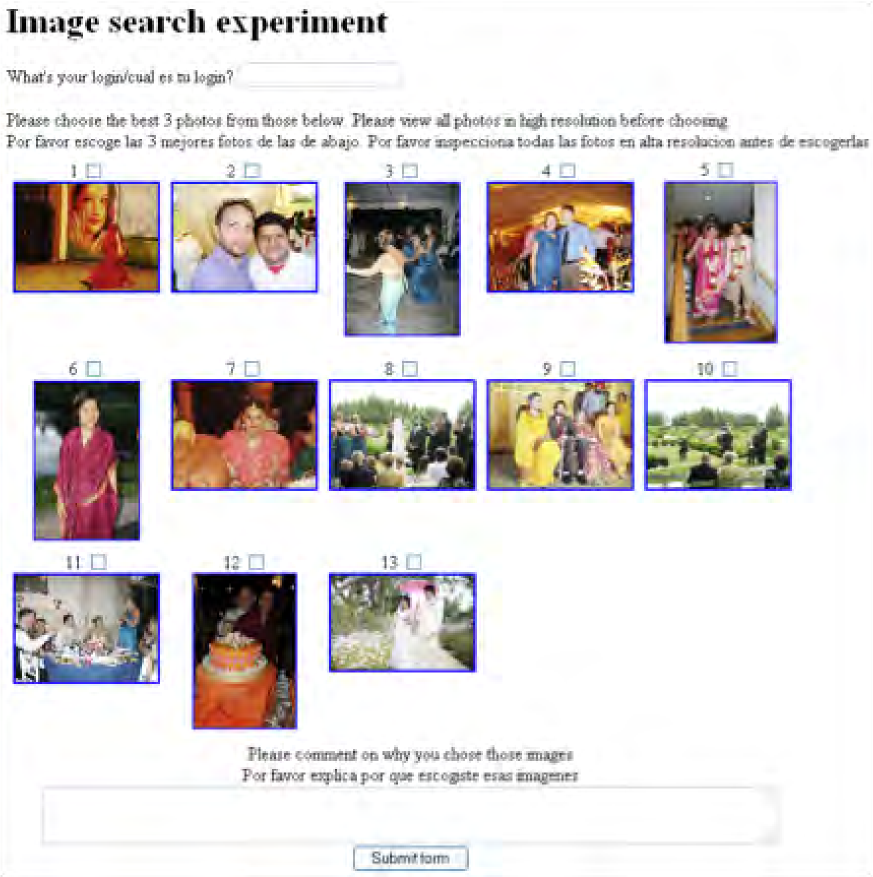


Image search experiment

What's your login/cual es tu login?

Please choose the best 3 photos from those below. Please view all photos in high resolution before choosing

Por favor escoge las 3 mejores fotos de las de abajo. Por favor inspecciona todas las fotos en alta resolucion antes de escogerlas

Please comment on why you chose those images

Por favor explica por que escogiste esas imagenes

computer screen and audio comments of participants for the duration of the study.

Participants were not told how the images had been selected or combined for presentation.

9.3.3 Treatments

Participants were exposed to three different rankings of images (treatments): (a) the original relevance ranking provided by Picasa Web Album (*Picasa*); (b) the image aesthetic appeal-based ranking (*Aesthetics*); and (c) the ranking resulting from applying the aggregation function described in Section 9.1, (*Fusion*).

The top five images from each treatment were presented to the user at the same time, in a randomized manner. Therefore, a total of at most 15 images were shown to each participant with each query: the images that were ranked in the top five of more than one treatment (*i.e.*, a collision), were only presented once (see Fig. 9.3 for an example). There were a total of 19 (12%) collisions for the nine general queries.

As previously explained, participants were asked to select the 3 best images from each presented set. Participants did not know the treatment (ranking) that each image was coming from. The order in which the queries were presented to each user was also randomized from user to user.

We use 3 performance measures to evaluate the results:

Treatment winner (TM): It quantifies the number of times that the selected photos came from each of the 3 treatments, as given by Eq. 9.3, with ($\sum TM_i = 1$), where i is the treatment under consideration. In order to take into account collisions between treatments, a second term is added to the right of the equation, such that when there is a collision, the reward is equally split between the treatments that generated that collision. For example, if one of the selected images appeared in the top 5 of both the *Picasa* and *Aesthetic* rankings, *i.e.*, $collisions(j) = 1$, their corresponding TM measure would be halved.

$$TM_i = \sum_{j \in PhotosInTreatment(i)} \frac{PhotoWasSelected(j)}{3} \cdot \frac{1}{collisions(j) + 1} \tag{9.3}$$

In Eq. 9.3, $PhotosInTreatment(i)$ is the set of images that belong to treatment i, and $PhotoWasSelected(j)$ returns 1 if that specific photo was selected by the participant, and 0 otherwise; and $collisions(j)$ is 0 if that specific image appeared only in this treatment, 1 if it appeared in this treatment and another one, and 2 if it appeared in all treatments.

Re-ranking performance (RM): The re-ranking performance (RM) quantifies how well each treatment ranked the images that were selected by the user. For instance, one of the selected images might have been ranked in position #6 by a treatment –and hence did not get any points from the TM measure. However, its RM would be significantly higher than the RM of another treatment that would have ranked the same image in position #100. To this effect, we propose the RM formula for treatment i, see Eq. 9.4.

$$RM_i = \frac{1}{3} \cdot \left[\frac{S - Pos_i(Ph1)}{S-1} + \frac{S - Pos_i(Ph2)}{S-2} + \frac{S - Pos_i(Ph3)}{S-3} \right] \tag{9.4}$$

where $Pos_i(Ph)$ is the position that each photo occupies in the treatment's ranking, where $Pos_i(Ph1) > Pos_i(Ph2) > Pos_i(Ph3)$; and S is the scope, or maximum rank considered –most participants of our study reported looking in at most the first two pages of results (scope=40). This scope has also been reported in the image search literature [60] as one of the most common ones. Therefore, we set the scope in our experiments to 40, in order to account for the most common user scenario.

Note that $RM_i = 1$ if all the selected photos are ranked at the top for treatment i, and $RM_i = 0$ when the selected photos are ranked at the bottom for treatment i (at or below the scope).

Overall performance (OM): The measures introduced above are averaged into a single measure (Eq. 9.5) that provides the overall performance for a specific treatment and query.

$$OM_i = \frac{TM_i + RM_i}{2} \tag{9.5}$$

9.4 Analysis of qualitative feedback

In this section, we summarize first the qualitative feedback provided by participants (audio transcripts and explicit textual feedback entered in the text box), as it is helpful in understanding the quantitative performance results –see Section 9.5.

All users in our study provided feedback on the reasons why they picked or did not pick specific images in the experiment. Most users (N=32; 86%) provided detailed feedback. After careful analysis of their audio and textual feedback, we identified five variables that users take into account when searching for images. We summarize the findings and highlight a few representative comments (positive and negative) about each of the variables.

9.4.1 Presence of people in the photo

A few participants enjoyed seeing happy people in the pictures ("she's happy / surprised") and candid pictures of people that tell a story in a daily activity setting (not posing). However, the vast majority of participants (N=32; 86%) were not interested in photos of people they did not know. Five users (14%) mentioned enjoying people when they tell a story within the image or when they are small (full body shot or smaller), so that the scenery can be seen. Finally, a subset of participants (N=3; 8%) reported liking images with people when the people were relevant to the query (*i.e.*, of expected ethnicity).

9.4.2 Emotional content in the photo

The emotional content in the photos did play a role in the participants' preferences. Some of the participants (N=15; 41%) enjoyed the pictures that made them feel better (*e.g.* "peaceful", "quiet", "smile"). A few participants (N=4; 11%) mentioned liking images that evoke a place that they long to be, or mesmerize them (*e.g.* "hypnotizes me", "I can even smell the rice field"). Two users had metaphorical observations interpreting the meaning of the picture as life passages and events.

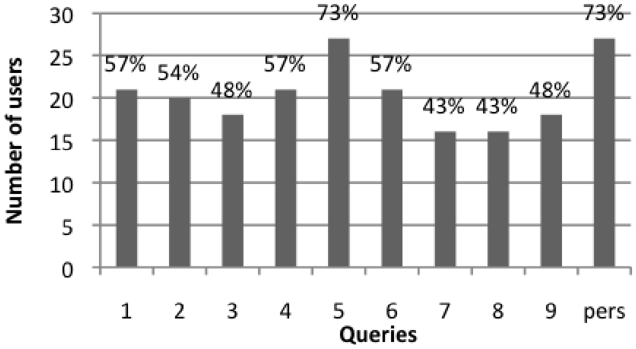
9.4.3 Preferences and personal experiences

Twelve participants (32%) liked the pictures that depicted something they enjoy in real life or surprised them (*e.g.* "I didn't know this existed"). A different subset of participants (N=4; 11%) liked images that sparked fond memories or that reminded them of the type of pictures they usually take. Conversely, they did not like the pictures that were very different from the type of pictures that they tend to take.

9.4.4 Expectations about search results

Most of the participants (N=35; 95%) felt reassured when they could understand the relationship between the retrieved images and the input query. Conversely, they did not like the images if they could not recognize an object or landmark that confirmed the query (*e.g.* "this could be anywhere", "I cannot see the park").

Figure 9.4: Percentage of users that questioned the relevance of at least one image as a function of the query.



Interestingly, users were very vocal about the relevance of the queries, even when they were not familiar with the subject of the query. We logged every time a participant would question the relevance of one or more images in a query. Fig. 9.4, shows the % of participants that questioned each query. The reason for questioning the relevance include: (1) the query subject (*e.g.* a beach for the query Q6:"Hawaii beach") did not appear clearly in some of the retrieved images; or (2) they could not identify the query location (*i.e.*, Hawaii for the query Q6:"Hawaii beach"), either because it did not show a landmark, or because it did not match their mental depiction of the place – which they may have acquired through documentaries or movies: "this is not what I'm expecting to find when I'm looking for Hawaii".

In our experience, low query relevance is a common problem in consumer image search, where not every single image is tagged or commented. When a user tags a set of images in bulk mode as *Hawaii beach*, a few of them will definitely show beaches, while others may depict content related to the trip to the beach (restaurant, same city, etc.), but not related at all with an actual beach. Kennedy et al. [65] analyze the user behavior in consumer image tagging in Flickr, and they conclude that the tags not necessarily describe the image content.

As seen in Fig. 9.4, participants were very picky with their personal query and with a few of the general queries. In particular, Q5: "wedding in the park" raised the same level of concern as the personal query:"this may be a wedding, but I cannot see the park". Similarly, the relevance of the search results was highly questioned in Q1: "Birthday party" ("this could be any celebration"), Q4: "New York buildings" ("these are images of people, not buildings"), and Q6: "Hawaii beach" ("I cannot see any beach").

9.4.5 Image aesthetics

Finally, most of the participants (N=33; 89%) mentioned aesthetic properties of the images at one point or another during the experiment. Comments about the image's composition (*e.g.* "the main subject is nicely isolated from the background"), good lighting, bright colors and sharpness were amongst the most frequent (N=18; 49% in average for every query).

Conversely, almost all participants (N=35; 95%) mentioned not liking low quality images with high levels of noise (*e.g.* "this has been taken with a cameraphone"), that were over/under exposed or out of focus –particularly the object of interest. In addition, some users made negative comments when the faces of the people in the photographs were too small in order to be identified.

Finally, two of the participants would extract the semantic meaning of the images and use that information to make their selections regardless of the low aesthetic quality of some of the images (they even selected extremely noisy night

Query 1.	Picasa 2.	Aesthetics	3. Fusion	P
1	0.41 [0.18]a	0.36 [0.15]a	0.22 [0.15]b	*
2	0.36 [0.18]a	0.43 [0.22]a	0.21 [0.16]b	*
3	0.18 [0.20]b	0.57 [0.21]a	0.25 [0.20]b	*
4	0.74 [0.26]a	0.15 [0.17]b	0.11 [0.18]b	*
5	0.63 [0.27]a	0.25 [0.25]b	0.12 [0.14]c	*
6	0.33 [0.23]a	0.34 [0.23]a	0.33 [0.20]a	
7	0.63 [0.26]a	0.12 [0.17]c	0.25 [0.20]b	*
8	0.39 [0.25]a	0.19 [0.15]b	0.42 [0.19]a	*
9	0.48 [0.29]a	0.33 [0.22]a	0.19 [0.16]b	*
Pers	0.53 [0.31]a	0.17 [0.16]c	0.30 [0.25]b	*

Table 9.3: Comparison between the treatments' Treatment Winner (TM) for each condition: Mean [std] score, ($*p < 0.05$). Descriptive statistics with different subscripts (*e.g.*, a, b, c) in the same row differ significantly for $p < 0.05$.

images). Both users were in the younger age bracket and with low photography skills, they owned either a compact or sub-compact camera, took pictures infrequently (from once a month to 3/4 times a year), and had between 500 and 5000 images in their personal collections. They are representative of a type of user that seems very little concerned with image aesthetics, perhaps due to their familiarity with low quality images taken with camera-phones.

9.5 Quantitative results

In this section, we present quantitative results of our user study. Please refer to the original paper for the details [C9]. We present the results for the three performance measures described in Section 9.3.3.

Tables 9.3, 9.4, and 9.5 summarize the results of the statistical analysis for each treatment, and for each of the queries. Descriptive statistics with different subscripts (*e.g.* a, b, c) in the same row differ significantly for $p < 0.05$, *i.e.*, if the subscripts are the same, then there is no statistically significant difference between the measures. For example, for Q1 in Table 9.4, there is significant difference between Aesthetics and Fusion, but there is neither a significant difference between Picasa and Aesthetics, nor between Picasa and Fusion since they share the same subscript. Table 9.3 summarizes the results of the Treatment Winner (TM). We can see that Picasa performs the best with statistically significant results, followed by Aesthetics. Similarly, Table 9.4 summarizes the results for the Re-ranking Performance (RM), where both Picasa and Fusion perform the

Query 1.	Picasa 2.	Aesthetics	3. Fusion	P
1	0.41 [0.18]ab	0.51 [0.18]a	0.38 [0.21]b	*
2	0.36 [0.18]b	0.55 [0.18]a	0.57 [0.16]a	*
3	0.23 [0.21]c	0.75 [0.18]a	0.34 [0.18]b	*
4	0.72 [0.26]a	0.24 [0.24]b	0.15 [0.21]c	*
5	0.60 [0.27]a	0.32 [0.25]b	0.18 [0.16]c	*
6	0.43 [0.25]b	0.39 [0.21]b	0.59 [0.20]a	*
7	0.61 [0.25]a	0.29 [0.22]b	0.58 [0.20]a	*
8	0.37 [0.25]b	0.50 [0.22]b	0.60 [0.23]a	*
9	0.49 [0.29]a	0.48 [0.28]a	0.49 [0.27]a	
Pers	0.58 [0.28]a	0.31 [0.26]b	0.53 [0.26]a	*

Table 9.4: Comparison between the treatments' Re-ranking performance (RM) for each condition: Mean [std] score, ($*p < 0.05$). Descriptive statistics with different subscripts (*e.g.*, a, b, c) in the same row differ significantly for $p < 0.05$.

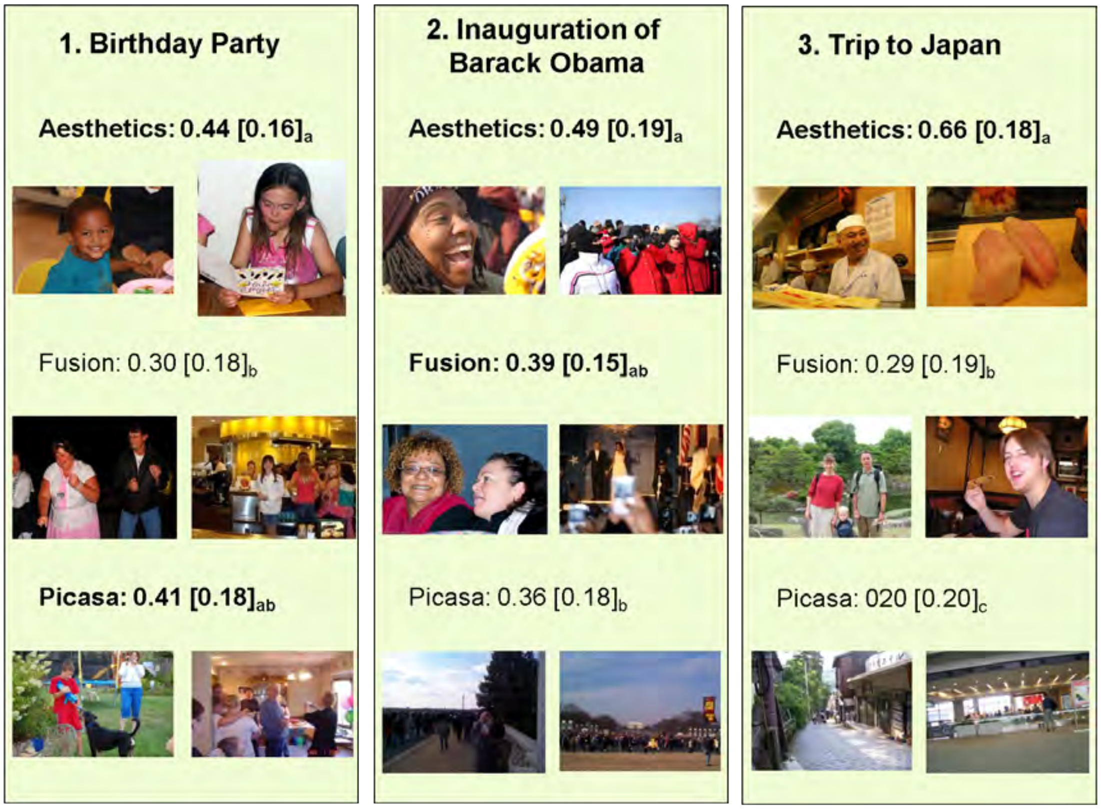
Query 1.	Picasa 2.	Aesthetics	3. Fusion	P
1	0.41 [0.18]ab	0.44 [0.16]a	0.30 [0.18]b	*
2	0.36 [0.18]b	0.49 [0.19]a	0.39 [0.15]ab	*
3	0.20 [0.20]c	0.66 [0.18]a	0.29 [0.19]b	*
4	0.73 [0.26]a	0.19 [0.19]b	0.13 [0.19]c	*
5	0.62 [0.27]a	0.29 [0.24]b	0.15 [0.14]c	*
6	0.38 [0.23]a	0.36 [0.21]a	0.46 [0.18]a	
7	0.62 [0.26]a	0.21 [0.18]c	0.42 [0.19]b	*
8	0.38 [0.25]ab	0.35 [0.17]b	0.51 [0.20]a	*
9	0.49 [0.29]a	0.40 [0.24]a	0.34 [0.20]a	
Pers	0.55 [0.29]ab	0.24 [0.20]c	0.41 [0.23]b	*

Table 9.5: Comparison between the treatments' Overall Score (OM) for each condition: Mean [std] score, ($*p < 0.05$). Descriptive statistics with different subscripts (*e.g.*, a, b, c) in the same row differ significantly for $p < 0.05$.

best with statistical significance in 2 queries, followed by Aesthetics. Table 9.5 presents the overall measure (OM), where the Picasa treatment performs the best (best 3 times with statistical significance), followed by Aesthetics (best once with statistical significance). In the following sections we analyze the OM results and discuss the implications of our findings in the design of consumer image search tools.

Figures 9.5, 9.6 and 9.7, present the results from Table 9.5, along with the top two ranked images for each query and treatment.

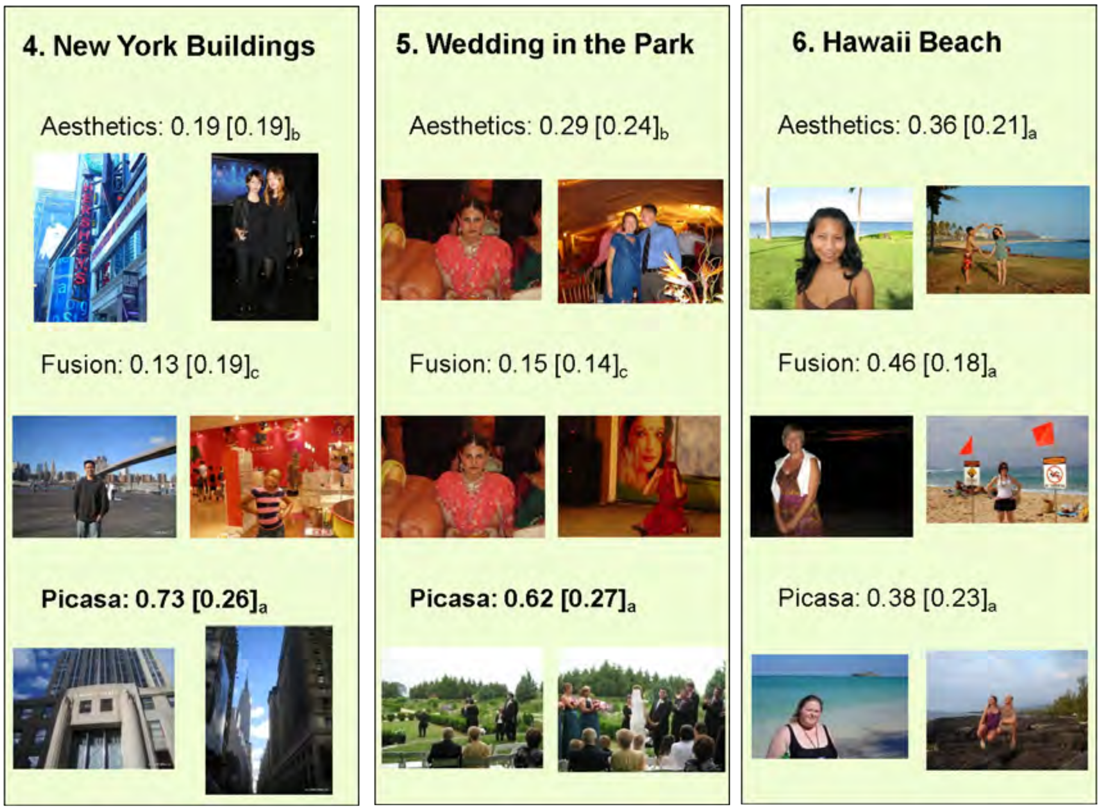
Figure 9.5: Top two ranked images for each of the treatments, for queries *Birthday Party*, *Inauguration of Barack Obama* and *Trip to Japan*.


1. Birthday Party
Aesthetics: 0.44 [0.16]$_a$
Fusion: 0.30 [0.18]$_b$
Picasa: 0.41 [0.18]$_{ab}$
2. Inauguration of Barack Obama
Aesthetics: 0.49 [0.19]$_a$
Fusion: 0.39 [0.15]$_{ab}$
Picasa: 0.36 [0.18]$_b$
3. Trip to Japan
Aesthetics: 0.66 [0.18]$_a$
Fusion: 0.29 [0.19]$_b$
Picasa: 020 [0.20]$_c$

9.6 Discussion and implications for design

In this section, we carry out a detailed analysis of the quantitative results (Overall Measure), modulated by the qualitative feedback previously described. In our analysis, we highlight key human-centric factors that play a role in the users'

Figure 9.6: Top two ranked images for each of the treatments, for queries *New York Buildings*, *Wedding in the Park* and *Hawaii Beach*.



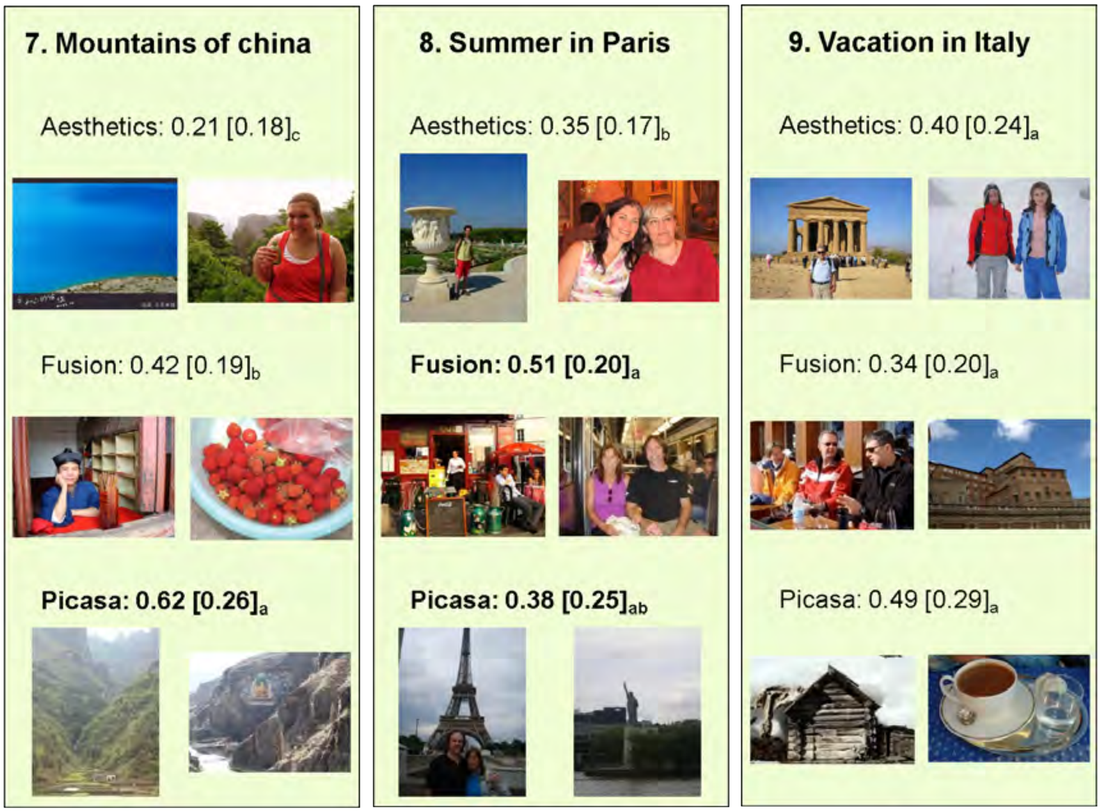
satisfaction with the images included in the results of a consumer image search task.

First, we turn our attention to the queries where the *Picasa* treatment is superior to the other 2 treatments with statistical significance:

Q4: "New York buildings". The relevance of the results returned by this query was highly questioned by participants (Fig. 9.4). In addition, none of the images returned by *Picasa* had any people in them –they portrayed buildings, whereas most images in *Fusion* and *Aesthetics* portrayed people in them. The fact that some traditional aesthetic measures (*e.g.* colorfulness) do not usually apply to buildings, caused the low relevance in *Fusion* and *Aesthetics*; in addition, the fact that those treatments' images portrayed mostly people, made *Picasa* the winner.

Q5: "Wedding in the Park". In a similar way to Q4, participants were very skeptical of the relevance of the images shown in response to this query

Figure 9.7: Top two ranked images for each of the treatments, for queries *Mountains of China*, *Summer in Paris* and *Vacation in Italy*.



(Fig. 9.4): most of the retrieved images did not show a park, and some of the weddings were Hindu, Sikh and Thai (all of them much more colorful than western weddings), which most participants could not recognize as weddings. The images of western-style weddings and most images of parks belonged to *Picasa*, hence the winner.

Q7: "Mountains of China". The photographs from the *Picasa* treatment had no people in them and mostly mountains, whereas both the *Aesthetics* and *Fusion* treatments included mostly pictures of tourists. Both the high relevance and lack of people make *Picasa* the winner.

Conversely, the *Aesthetics* treatment clearly outperformed all other treatments in query **Q3: "Trip to Japan"** (see Table 9.5). In this case, the images returned by all treatments included people in them. However, most of the photos in the *Aesthetics* treatment included Asian persons of medium to large size. Interestingly, a photo that was often selected by participants depicted three Asian chefs in a sushi restaurant.

For **Q2: "Inauguration of Barack Obama"**, both *Aesthetics* and *Fusion* are better than *Picasa* with statistical significance. The photos retrieved by the *Picasa* treatment were of Washington D.C. on Inauguration Day, but of poor quality (camera-phone probably). The *Fusion* images all included people – in one case a bad quality picture of Barack Obama and wife. Finally, all the images retrieved by the *Aesthetics* treatment had people in them, one clearly related to the inauguration event (*e.g.* a very happy lady wearing an Obama hat while watching the event).

Finally, in the case of **QP: personal query**, *Picasa* and *Fusion* are statistically better than *Aesthetics*. In this case, participants were also concerned with the relevance of the images (Fig. 9.4).

From the discussion above, we shall highlight the following implications for the design of consumer image search tools. Most of the findings are in alignment with our own intuitions, but supported by empirical data provided by our user study:

1. More relevant tags are needed: As previously explained, users do not typically label their personal images and when they do, they tend to do it in "bulk" [71]. Moreover, sometimes the tags do not describe the image contents [65]. Therefore, the quality of the results generated by tag-based consumer image search engines tends to be lower than that of general image search. Interestingly, [9] showed how user satisfaction in consumer web image search recall-based tasks, did not have a statistically significant correlation with the effectiveness of the search system, which somehow contradicts our findings. In our study, participants were happy (did not express concerns) with the relevance of the retrieved results only 44% of the time (ranging from 27 to 55%, depending on the query as shown in Figure 3). Recent work in the literature has tried to improve the frustrating low performance of tag-based search engines for consumer images. For instance, [66] successfully combines tags with location metadata and visual cues in order to boost performance. Content-based analysis techniques are particularly relevant in this domain, in order to automatically or semi-automatically label the high percentages of unlabeled images that are stored in the personal repositories of users and therefore improve the relevance of the retrieved results.

2. Relevance is more important than pure Aesthetics: The aesthetics reranking algorithm is designed to re-rank the retrieved images by their aesthetic appeal, assuming that all have similar (high) relevance. Therefore, when the results of the query-based search engine are poor (as in Q4 and Q5), images with high aesthetic appeal –that are promoted by the image

aesthetics algorithm– are likely to be irrelevant and thus unsatisfactory to the user. The work in [36] presents 3 different user intents in image retrieval: searcher (knows what he is looking for), surfer (not sure of the purpose) and browser (no purpose at all). In the case of surfer and browser-type of users, we expect aesthetics to be more important than for searchers. We plan on carrying out additional studies to verify this hypothesis.

3. Personalization and context are needed: Our study has confirmed that different users weigh aesthetic appeal differently when selecting images in a search task. User modeling techniques would greatly help in understanding whether aesthetics are important for each user, and in what measure. Note that the task or reason why the user is searching for images plays a role in the decision making process. This factor is beyond the scope of this work, but part of our research agenda.

4. Image aesthetic measures are not universal: Different types of topics have completely different aesthetic connotations, *i.e.*, buildings vs. weddings. Therefore, we believe that a general image aesthetics measure –such as the one proposed in this chapter– is unable to represent the range of parameters that might play a role in defining the aesthetic appeal of an image. Thus, we are working on a category-dependent aesthetic measure. This is related to the work presented in [33], where the authors use intent categorization in order to perform better relevance re-ranking. As an example of category-based analysis, we already presented a model for aesthetic appeal of faces in Chapter 6.

5. Fusion is not straight-forward: The proposed fusion algorithm performed poorly when compared to the Picasa and Aesthetics treatments. However, our experiments suggest that Fusion might be appropriate in the cases of queries that produce highly relevant results (Fusion performs best in 2 queries for the RM measure, see Table 4). We are investigating alternative, category and task-dependent fusion functions.

6. Sensitivity to the Presence of People: Finally, participants in our study vastly preferred images without people than with people. This behavior is probably due to the fact that the Picasa Web Album photos are personal photos of mostly unknown individuals. Therefore, all the persons depicted in the photos that we used in our study –except for President Obama– were unknown to the participants. An interesting exception is the case of travel-related queries to exotic countries, where participants preferred images that showed native people of the countries related to the queries. We understand that this is a drawback of our experimental database. In the

case of running the experiment on the personal collections of the participants, we would expect users to be more interested in photographs with familiar (*e.g.* friends) people in them [101]. Future research should attempt to use the participants' personal image collections.

CHAPTER 10

Media aesthetics conclusions

In this Part I of the dissertation, we have described a series of media aesthetics computational models. Two regression based image aesthetics models [C9,C11,C13] have been described in detail. These models segment out the relevant region out of the image, and calculate specific low level features on that region, as well as on the rest of the image, including *sharpness, contrast, colorfulness, light exposure, noise* and *background homogeneity.* A computational model for face aesthetics has also been described [C11] which is helpful when aesthetically ranking images of the same person. These models are used in Part II of this dissertation as building blocks of the photo storytelling systems described therein.

We have also covered two different classification models. In the first one [C4] images are classified into high aesthetic appeal images, or low aesthetic appeal images, using only visual composition features. In the second one [C5] video clips are classified into high aesthetic appeal videos, or low aesthetic appeal videos.

Finally, we have looked at the real application of image search re-ranking by using one of the regression based image aesthetic models, *i.e.*, the rank order generated by the search engine is modified by the aesthetic appeal of each of the images in the list. One of the main conclusions is that human observers seem to key into different features when judging the aesthetic appeal of an image, depending on the general category that the observed image belongs to –*e.g.* landscape, buildings, people. Research on category dependent media aesthetics will be part of our future work.

Part II

Media Storytelling

CHAPTER 11

Media storytelling introduction

In recent years, and mainly due to the pervasiveness of digital cameras and camera-phones, there has been an exponential increase in the overall number of photos taken by users. This dramatic growth in the amount of digital personal media has led to increasingly large media libraries in local hard drives and/or online repositories, such as Flickr!, Picasa Web Album or Facebook. Picasa Web currently allows up to 1000 images per album[1].

Unfortunately, large photo collections turn the manual task of selecting images into a tedious and time consuming process [47, 125], for instance, Facebook has taken some initial steps at making the upload process a bit more user friendly[2]. In addition, the familiarity that users have with the photos belonging to a specific event will decay over time [125], turning the photo selection task more difficult with time.

Our user studies show that people do not create as many photo books as they would like to, one of the reasons being the image selection process is too painful in the current digital photography landscape, in which hundreds or even thousands of photos are taken in one single event. For instance, a few of the participants reported spending one full weekend in order to accomplish the selection of a set of images, out of a collection of 800+, and laying them out on a fixed layout photo book template. These results were corroborated in our user study presented in Chapter 14.

On the other hand, the social narrative use of photos – *i.e.*, social photo storytelling – plays an important role in people's lives as it serves to structure and share personal and interpersonal experiences and to express personal and group identities [57]. At the same time, Rodden et al. [98] revealed that "the most important use of digital photographs is to record holidays or other significant events, and then show those pictures to friends and family". Hence, it does not come as a surprise that automatic approaches to personal photo collection summarization and event detection have recently been of interest in the research community [31, 50, 80, 87]. Unfortunately, none of these approaches addresses social aspects of these photo stories, such as its target audience, *i.e.*, who is going

[1] http://picasa.google.com/support/bin/answer.py ?hl=en&answer=98898

[2] http://blog.facebook.com/blog.php?post=206178097130

to enjoy the photo story. This motivated us to propose the photo storytelling system presented in Chapter 14 that takes analyzes the user's On-line Social Network (OSN) in order to adapt to the potential audience on that particular OSN.

In certain storytelling settings –photo book or slideshow– targeting a specific image count may be of high importance. For instance, targeting the exact number of images in a template photo book, or targeting a specific slideshow duration. In order to accomplish this task, and still ensure the best coverage of the story, with the most relevant images, a versatile image collection representation is introduced in Chapter 13, which allows for automatic scalable selection in order to target a specific final image count. A hierarchical time clustering is presented, which is traversed at a specific hierarchy level in order to select images by alternating among all time clusters, and selecting the most relevant images in each of the clusters. The relevance ordering we use is based on a combination of features, namely, important people, smile detection, image aesthetic appeal measures, and whether a near-duplicate of the image has already been selected. Once this Hierarchical Scalable Representation has been created, it can be reused to generate any target size summary. In Chapter 13 we present two automatic image selection algorithms, one that selects images from clusters with high average image relevance more frequently, and another one that selects images from larger clusters more frequently. This overall system –the implementation of which is introduced in Part III of this dissertation– was used over the period of one year on several large image collections –each above 800 images; the resulting selection was presented to their owners in the form of photo-books in order to get feedback, validating the presented approach, see Section 13.4.

With the advent of photo and video capabilities in OSN, an increasing portion of the users' social photo storytelling activities are migrating to these sites, where friends and family members update each other on their daily lives, recent events, trips or vacations. For instance, FaceBook is the largest online repository of personal photos in the world with more than 3 billion photos being uploaded monthly[3]. Hence, there are opportunities to mine existing photo albums in OSN in order to automatically create *relevant* and *meaningful* photo stories for users to share online.

Fully automatic personal photo collection summarization for storytelling purposes is a very hard problem, since each end-user may have very different interests, tastes, photo skills, etc. In addition, meaningful and relevant photo stories require some knowledge of the social context surrounding the photos [80], such as who the user and the target audience are. Hence, we believe that automatic summarization algorithms should incorporate this information.

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

In Chapter 14 we present a photo collection summarization which builds on the main concepts presented in Chapter 13, and also learns some of the users' social context by analyzing their online photo albums. In an in-depth user study conducted with 12 subjects, the proposed system was validated as a first step in the photo album creation process, helping users reduce workload to accomplish such a task. Our findings suggest that a human audio/video professional with cinematographic skills does not perform better than our proposed system.

We want to note that the author also worked on the topic of *video storytelling*, filing two patents [C56,C75] to protect those ideas.

In this Part II of the dissertation we describe two main photo storytelling approaches. The first one was part of the author's research while he was at Hewlett-Packard, and the storytelling aspects were geared towards a Photo-Book story, which is described in Chapter 13. The second one was part of the author's research while he was at Telefonica I+D, and the storytelling application was focused on slideshow story sharing on a social networking site, which is described in Chapter 14.

CHAPTER 12

Media storytelling prior art

In this chapter we will describe some of the related work to multimedia storytelling that has been done in two different communities, the Human Computer Interaction (HCI) community, and the image/video processing community.

12.1 How do people tell stories with photos

In this section we present the prior work done in the HCI community, in which the overall process of how people tell stories with photos has been analyzed, and certain conclusions have been reached as to how the best way to help users in this task would be. In Chapter 14 we present a new photo storytelling algorithm and analyze it in a user study with HCI techniques.

Kirk et al. [68] define a number of stages in the image acquisition pipeline prior to photo storytelling:

1. *Pre-download stage*, where users do a certain amount of on-camera collection editing, –*i.e.*, selecting one image from a set of near-duplicate images to keep, and deleting the others, or deleting non aesthetically appealing images.

2. *Download stage*, where around one half of the users just file the images away, and the other half actually does some collection editing.

3. *Pre-share stage*, where users undertake a significant amount of work sorting and preparing photos for sharing. One of the most common and time consuming activities was selecting and/or sorting of images, *i.e.*, considering any one photo against a collection of others, and making decisions about what to keep/delete/share/not-share.

In a series of HCI studies researchers seem to agree that large photo collections turn the manual task of selecting images –*i.e.*, the *pre-share stage*– into a tedious and time consuming process [47, 125]. In [125], Whittaker et al. examine the effects of new technologies for digital photography on people's longer term storage and access to collections of personal photos. They report an empirical study of parents' ability to retrieve photos related to salient family events from more than a year ago. Performance was relatively poor with people failing to find almost 40% of pictures. Possible reasons for retrieval failure include: storing too many pictures, rudimentary organization, use of multiple storage systems, failure to maintain collections and participants' false beliefs about their ability to access photos. In [47] Frohlich et al. state that the discipline required to filter and arrange favorite photos into albums was too much for many mothers who reported frustration in wanting to make photo albums while lacking the time and motivation to do so. As life gets more hectic and as additional children are born, it appeared to get harder for families to keep up with the backlog of images. In another HCI work, now looking at collaborative storytelling, Crabtree et al. [32] explore the interactional ways in which people naturally collaborate around and share collections of photographs.

In the HCI community there have been various attempts at creating useful user interfaces in order to help the users create their media stories. In one of the earliest works on storytelling with digital photos Balabanovic et al. [15] extended the common practice of storytelling around print photos to digital photos using a tablet PC, where they found that people would rather "select, and then narrate" the story, rather than "select while narrating". In a different work [72] Laundry et al. present a digital narrative composition tool using digital images, in which users had to go through four exhaustive steps in order to accomplish the final story, namely: *brainstorm*, *organize*, *write*, and finally, *add personal media*. In [12] Ames et al. show a user interface for generating photo stories, in which the users need to manually place their photos on a dramatic arc. In [105] Shen et al. present a video editing system that helps authors compose a sequence of scenes that tell a story, by selecting from a corpus of annotated video clips, including *characters*, *emotions*, *themes*, and *story structure*.

Unfortunately, all the systems noted above need the user to perform all the tasks in a manual way. In the next section we look at automatic algorithms that can actually help the user in reducing the burden of the media organization and selection.

The main implications for design in [68] were to design tools to help users sort images, cluster poor quality images, cluster similar images, and keeping representative images; also not to over-automate, since it confuses users. At the same time, in [72] it was noticed that participants excluded portions of their experience when they did not have media to visually represent them, and they listed a sequence of events that took place in their story, as opposed to documenting the dramatic arc. These implications have been great motivators in our research on media storytelling. See Chapters 13 and 14 for two algorithm examples that try to solve the problems stated above.

12.2 Recent work on automatic photo storytelling algorithms

This area has seen a lot of activity in the last ten years. We will describe here some of the most relevant approaches in the literature.

Most of the prior art related to our proposed approach relies on the information extracted from the photos to process only, *i.e.*, they do not automatically analyze the social context around those collections, like, for instance, the people they are going to be shared with. They either analyze the images in a personal collection, or a set of images retrieved from the web, by clustering them into events, either for collection navigation or summarization. In the summarization case representative images are selected from each of those events.

In [94], Platt presents a simple time clustering algorithm which starts a new cluster if a new photo is taken more than a certain amount of time since the previous photo was taken. Clusters are merged based on content analysis until the desired number of clusters is reached. The photo in the center of the time cluster is selected as its representative image. This algorithm was improved in [95], by means of an adaptive temporal threshold and a new approach to select the representative image of each cluster (the most distinctive image in the Kullback-Leibler divergence sense). Loui *et al.* present in [80] an automatic albuming system in which collection summarization is performed by event detection using time clustering and sub-event clustering based on color similarity; in addition, very low quality images – with underexposure, low contrast and camera de-focus, all related to image aesthetics – are discarded. In [50] a browsing interface is presented that summarizes photo collections by exploiting the capture time information in an adaptive way, similar to [95]; the allocated space for each event is roughly proportional to the number of photos taken in that cluster, and the representative images for each event are selected by identifying very close or very distant images in time. Naaman *et al.* [87] present a system that utilizes the time and location information – *i.e.*, GPS coordinates– to automatically organize a personal photo collection in a set of event and location hierarchies.

An unsupervised automatic similarity-based method to cluster digital photos by time and image content is presented in [31]. The approach is general, and makes minimal assumptions regarding the structure or statistics of the photo collection. Inter-photo similarity is quantified at multiple temporal scales, and the *best* scale is selected by the algorithm. In [49] the algorithm is used as the back end for an image browsing user interface, where the representative pictures in the clusters are selected based on user ratings. Chu et al. [28] present a multimedia summarization approach in which correlation between photos and videos of the same event is exploited in order to identify relevant photos and video snippets. In a different work [29] Chu et al. only analyze images, and generate a similarity graph based on their SIFT similarity; the degree of centrality of each photo in the graph is calculated, and is used as the main feature to select images for summarization.

Finally, there has also been some work in web (*i.e.*, Flickr) multiuser collection summaries. For instance, Simon *et al.* [106] have recently proposed a solution to the problem of landmark summarization, *i.e.*, the Pantheon in Rome. They use multi-user image collections from the Internet, and select a set of canonical views – by taking image likelihood, coverage and orthogonality into account – to form the scene summary. A similar approach is extended in [67] by adding location metadata, new visual features and a more sophisticated representative image selection by clustering the images into visually similar groups, and generating links between images that contain the same visual objects. In [63] Joshi et al. extract semantic keywords from a written story and an annotated image database is searched, generating the final photo story in an unsupervised manner. Unfortunately, users are typically reluctant to annotate images with text [99], and therefore such a system may not be suited to generate personal photo stories.

As mentioned in Section 11, algorithms that take advantage of information regarding the social context of the user in order to perform a better job at collection summarization are of interest. For instance, in [8] an algorithm is presented that improves a multimedia browser based on social metadata – *i.e.*, places the users spend time at, and people they meet with – obtained via GPS traces of daily life routines. More recently, Loui *et al.* [79], have presented an image value assessment algorithm that takes into account social relationships between detected people in the photographs, where a higher weight is given to photos of close relatives and lower weight to the photos of, for instance, neighbors. Unfortunately the social relationships need to be entered manually by the user.

Recently, Li et al. [74] have presented an image selection system based on the assessment of the aesthetic visual quality of consumer photos, focusing on one single photo genre: photos with faces.

None of the previous work approaches address the social aspects of the photo stories, such as their target audience in an automatic way. In the same spirit as the work in [8, 79], in Chapter 14 we propose a photo storytelling system that leverages information from the user's OSN photo albums in order to create a *personalized* photo story, *i.e.*, a photo story adapted to the user's style and target audience.

CHAPTER 13

Scalable automatic photo storytelling

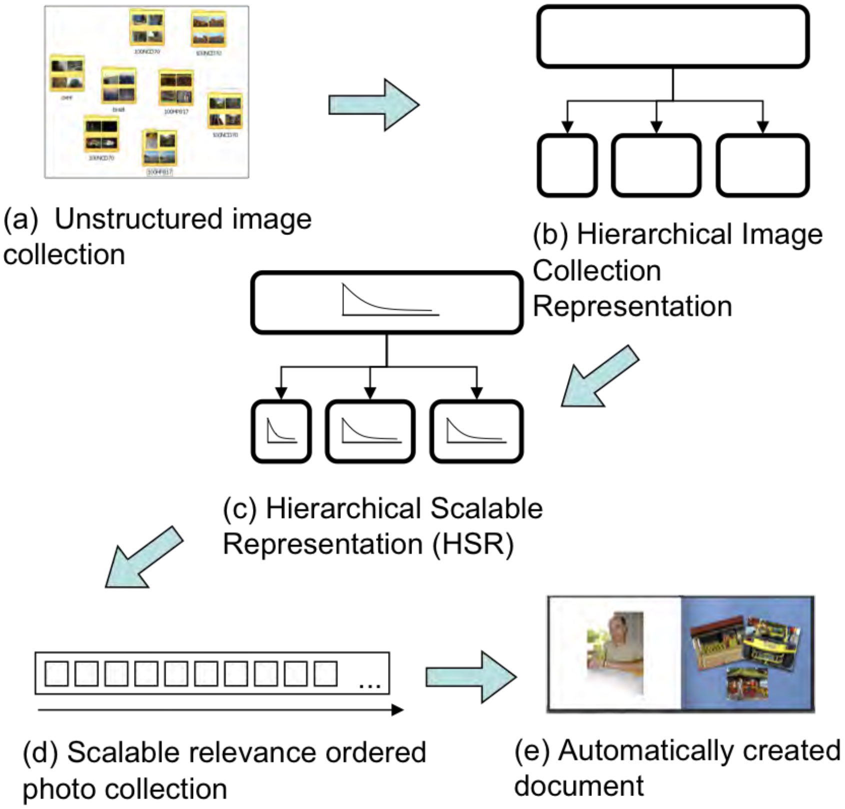
In this chapter we consider how a Hierarchical Scalable Representation (HSR), can be constructed from an unstructured image collection (see Fig. 13.1). This HSR is a taxonomy based on hierarchical image clustering (*e.g.* time and similarity based; see Section 13.2.1), image relevance features (*e.g.* measured image aesthetic appeal, face clustering and smile detection; see Section 13.2.2); and the user input that allows the system to generate a targeted relevance metric from the individual relevance features. All of the pre-processing steps are designed to have predictable and reasonable default behaviors. However in many cases users may elect to fine-tune these results, such as opting for more people's images, more finely sampling of a specific event (*e.g.* time cluster), or include *favorite* images. Taking into account the user relevance configuration, a hierarchy with relevance sorting at each node (Fig. 13.1c) is created. The outcome of the presented algorithm, is a single scalable relevance ordered image list (Fig. 13.1d) that covers the whole image collection. The inherent scalability of such representation allows for straightforward photo-book population by selecting the top NN images from the relevance ordered list (Fig. 13.1e). The work presented in this Chapter was originally published in [C10] with an emphasis on photo-book storytelling. These algorithms have been protected in these patent applications [C71,C79].

13.1 Photo-book creation user studies

In many cases photo-books are created with considerable manual intervention, in which users carefully select a greatly reduced sub-set of images, and then, using drag-and-drop interfaces, manually position images on pages. A recent survey of web-based printing [37] found that 4 out of 7 vendors surveyed used drag and drop modality during the publication creation process. The image selection process tends to be subjective but there are certain trends that can be observed.

Our user studies show that people tend to remove redundant or near-duplicate images. In some cases there are simple rules that tend to be used, for instance if a user has multiple photos of a child it is often the case that the photo with the child smiling is preferable. Next people tend to group images by important events

Figure 13.1: From unstructured image collections, to automatically fulfilled photo-books. A Hierarchical Scalable image collection Representation (HSR) is created, which helps drive the automatic image selection in the form of a Scalable Relevance Ordering. This, in turn, can easily auto-populate any size photo-book (NN images in this case), due to its scalability properties.


(a) Unstructured image collection

(b) Hierarchical Image Collection Representation

(c) Hierarchical Scalable Representation (HSR)

(d) Scalable relevance ordered photo collection

(e) Automatically created document

such as people, time, location and topic. Another broad category is the basic quality of the images, such as sharpness and contrast. The image composition and colorfulness are also considered. All of these factors have been considered in our image aesthetic models described in Part I of this dissertation.

Deriving a sub-set of images from a large collection of images is a time consuming task. Our tests have found that starting with 500 images and reducing a collection down to 100 images can take one or more hours depending on the user. For significant collections, such as over 1000 images, this can be sufficiently time consuming as to prevent or limit this process to very a few number of key important occasions.

Below, we describe a series of constraints that should be met by an automatic image selection algorithm.

13.1.1 Selection constraints

From our user studies we also identified a series of functionalities the users would welcome in a semi-automatic photo-book creation process:

1. the ability to start with an automatic first pass photo-book where selection of images and layout were done for them,
2. the flexibility to change layout, size of certain section (*i.e.*, prominence of it),
3. the ability to easily swap images by semantically related ones (*e.g.* near-duplicates, same person in picture),
4. suggest a small number of images as cover pictures for the photo book.

13.1.2 Storytelling constraints

It is very important to have a good coverage of all important events from an image collection: time events, as well as relevant people or characters in the story. Also good storytelling avoids redundancies and boring repetitions. Finally, good storytellers try to explain the more interesting passages within a certain act (*i.e.*, time cluster), which we interpret as to covering the most aesthetically appealing images of that act [68] in our algorithm.

13.1.3 Layout constraints

A key layout constraint is driven by the combination of the number of pages in the book and roughly the number of images per page. This layout constraint in turn drives the approximate number of images to be selected from a larger collection. Specifically consider three different cases where the overall collection is approximated with a highly reduced number of images, say 10 percent, a moderately reduced set of images, say 30 percent and a lightly reduced set of images, say 50 percent. These three examples, for a collection of about 800 images, may correspond to a lower volume publication, such as a saddle stitched document with a low page count, a medium sized perfect bound book and a high capacity perfect bound book. In each case the configurable image taxonomy presented in this chapter can be used to drive the image selection given the book and layout constraints.

13.1.4 Photo-book prominent images

A specific definition of prominent images is challenging and likely contextual. For example, for a photo-book cover image, it may be limited to landscape oriented images that are highly appealing, have a face on the right side of the image and have a more uniform region for a book title.

13.2 Image collection taxonomy

The findings from the previous section motivated us to design a versatile image collection representation which could be easily reconfigured, and at the same time, be highly scalable in order to meet a specific image count to fit any document size, while providing a good coverage of the whole collection and allow for selection of highly relevant images.

In Section 13.2.1 we describe a hierarchical representation that takes into account time and similarity clustering, which allows for coarse scalability –*i.e.*, targeting a rough photo-book size– and also allows for easy reconfiguration by the user (*i.e.*, select alternative near-duplicates, give more prominence to a certain time cluster, etc.). In Section 13.2.2 we describe a scalable representation that allows for fine grain scalability; this is accomplished by ordering the images from each time cluster based on image relevance; this is a user configurable relevance ordering based on features like an aesthetic appeal metric, face clustering, smile detection and, optionally, user favorite tagging.

13.2.1 Hierarchical representation

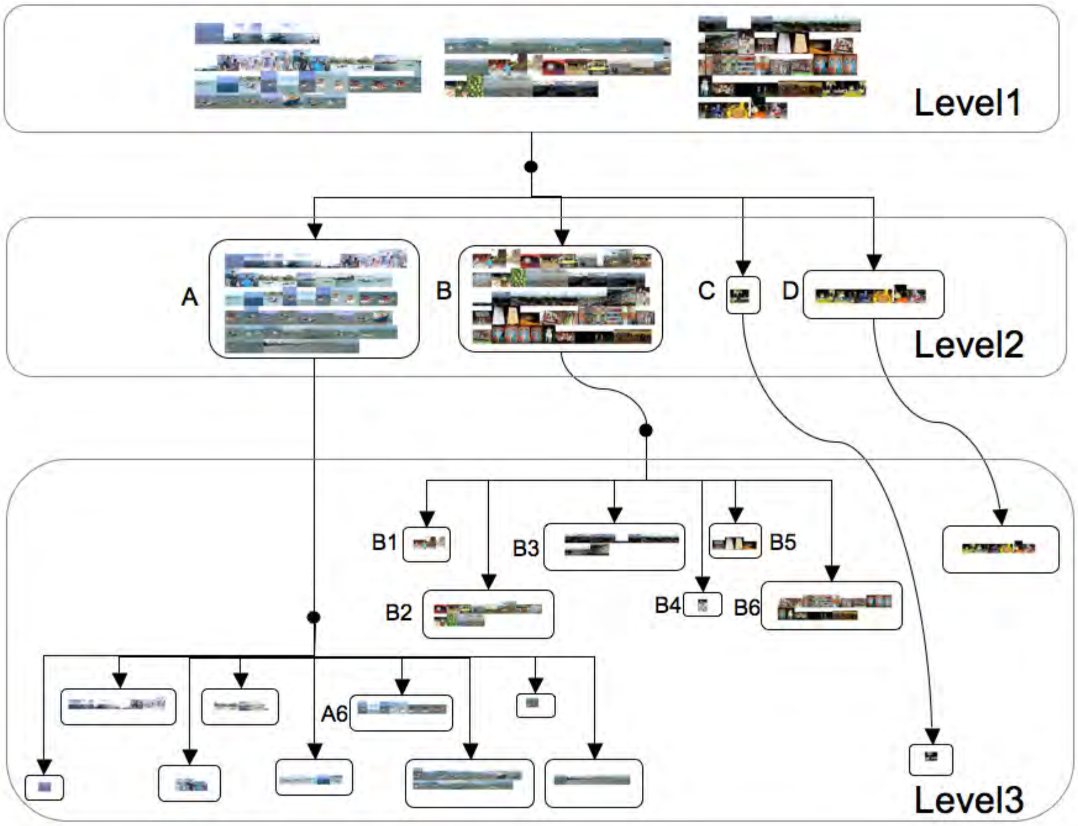
At the time of creating a photo-book, the user usually has a storytelling purpose. This story may be centered around a time event (*e.g.* holiday, wedding); it may be centered around a place (*e.g.* London, Paris); or it may be centered around a certain person, or character (*e.g.* John, our newborn).

Face clustering [129], can be used to help select images of a certain person –*i.e.*, character; unfortunately, such systems have low reliability to generate the main image collection representation. Geographical Position System based coordinates (geo-tagging) can be used to create a hierarchical geography based clustering; unfortunately, very few images incorporate such geo-tagging information. All digital cameras do capture the time information and save it in the EXIF (EXchangeable Image File [6]) header, and time clustering has been used for some time [31, 48, 50, 94] for image collection browsing; unfortunately, we have learnt from our experiments that it is highly improbable that multiple cameras will be time synchronized in a multi-editorial photo-book.

From a storytelling perspective, as mentioned above, the user does not want redundant images to be included. Near-duplicate detection and similarity clustering are therefore important aspects of a good image collection representation.

Section 13.2.1.1. presents a hierarchical time clustering representation and an algorithm to accomplish such representation in an automatic manner. Section 13.2.1.2. presents a hierarchical similarity clustering, that is built embedded into the hierarchical time clustering presented in Section 13.2.1.1.

Figure 13.2: Real life example of a time cluster hierarchy. Top cluster (Level1), Level2 (4 sub-clusters) and Level3 (17 sub-clusters). The hierarchical time clustering algorithm works on the statistical distribution of capture times, and therefore certain clusters are much larger than others, depicting different "events".



13.2.1.1 Hierarchical time clustering

In our user studies, as well as in our experiments, we found that users will normally not go below 12/16 images for a photo-book (*e.g.* photo-calendar, 8-sheets photo-booklet); on the other extreme, some users may actually print the whole collection, avoiding only non-appealing images, and near-duplicates.

One of the main storytelling practices is to have a good coverage of the story. To this effect the images are time clustered in a hierarchical way, so that the selection algorithm can select images from each cluster, at a specific hierarchy level (see Section 13.3). The top hierarchy of the time clustering should be devised to accommodate the smallest size photo-books in order to accomplish the need for good story coverage, by at least selecting one image from each time cluster.

In order to have coarse scalability in our image selection algorithm, we have implemented a hierarchical time clustering based on the algorithm presented in Section 3.1.1., which subdivides the collection into increasingly smaller event clusters.

Fig. 13.2 shows the actual output of the presented time clustering algorithm: (level1) presents the image collection, and how it is hierarchically sub-clustered into 4 sub-clusters (level2), and further sub-clustered into 17 sub-sub-clusters (level3). Each of these time clusters represents a different event within the larger cluster, therefore allowing for a selection algorithm (Section 13.3) to choose images from each of these clusters alternating among all of them, and generate a good story coverage.

Fig. 13.3 and 13.4 show a two level hierarchy time clustering example. This is going to be used in the following section to explain how the similarity hierarchy is embedded into the time clustering hierarchy.

Figure 13.3: Level 1 time clustering into coarse time events.

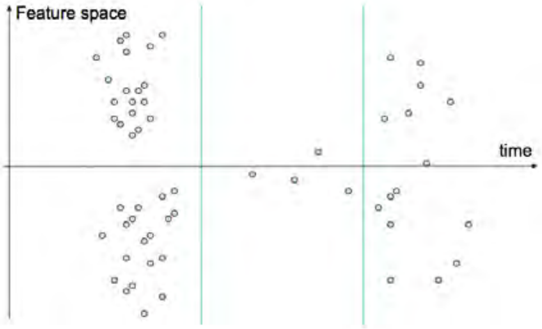
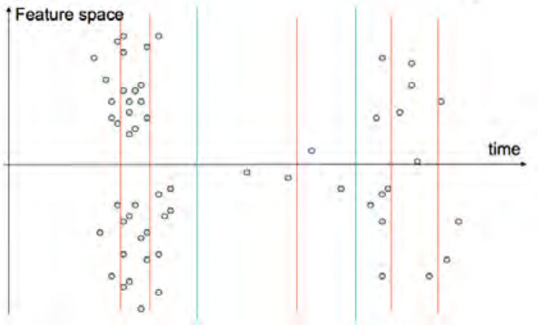


Figure 13.4: Level 2 time clustering: finer grain time events.



13.2.1.2 Hierarchical similarity image clustering

Another important storytelling practice is to avoid repetition. To this effect, embedded within each of the time clusters, images are clustered based on similarity, which allows for near-duplicate detection. A representative image from each similarity cluster, the most aesthetically appealing one, is selected.

Figure 13.5: Near-duplicate detection within each Level2 time cluster.

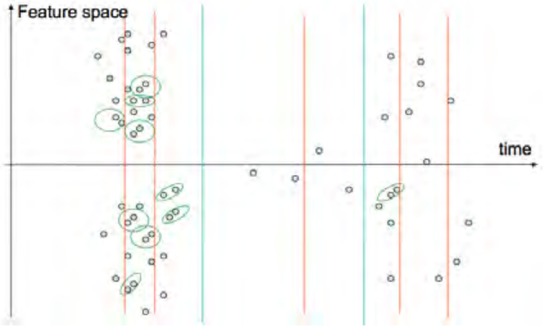


Image similarity is calculated by a region based approach, as described in Section 3.2.2.

In order to successfully create an embedded similarity hierarchy within the time event hierarchy, similarity clusters are first created within the time clusters at the bottom of the hierarchy (smallest time clusters); then, the similarity threshold is lowered and similarity clusters are calculated within each time cluster at the next hierarchy level up; this is repeated until reaching the top of the time clustering hierarchy. This is important at reconfiguration time, *i.e.*, if the end-user wants to give more prominence to a specific time cluster, then the selection for that time cluster will move to a lower level, with smaller time sub-clusters; the way our embedded similarity hierarchy is created avoids the possibility of the similarity clusters overflowing into nearby time clusters (this might well happen

Figure 13.6: Similarity detection within each Level1 time cluster. The centroid of the near-duplicate cluster at Level2 is used for the calculations.

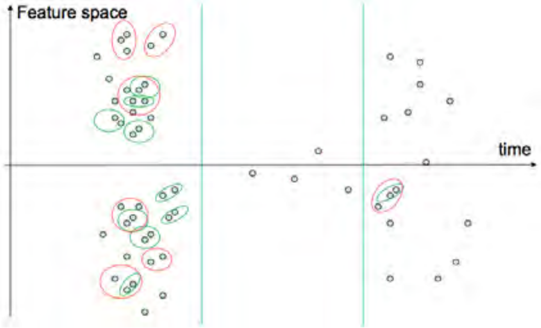
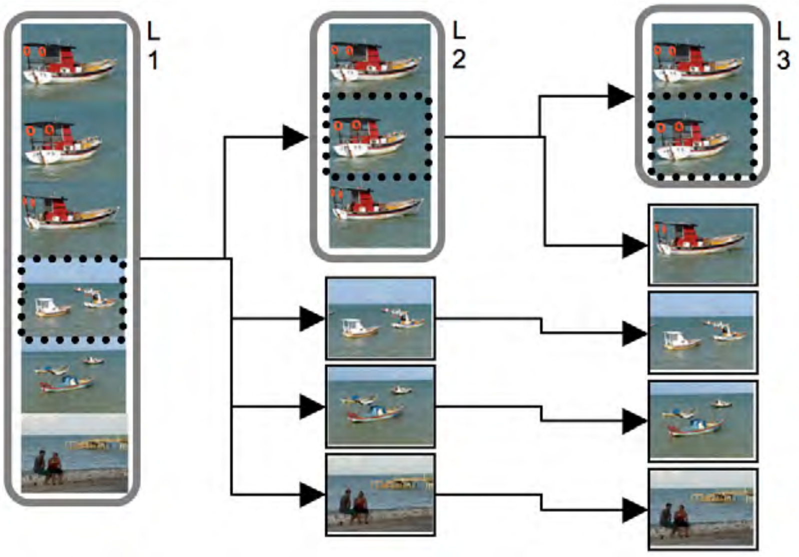


Figure 13.7: Three level similarity hierarchy. This figure shows how time cluster A6 (Level3) in Fig. 13.2 has a similarity cluster with 2 images in it. When moving to Level2, there is a larger similarity cluster (3 images) which embeds the smaller cluster at Level3. Finally, at Level1, we find one even larger similarity cluster, which embeds the smaller cluster at Level2. Within each similarity cluster, the most aesthetically appealing image is marked, and will represent that similarity cluster (*i.e.*, representative image).



if the similarity hierarchy were to be started at the top of the time cluster hierarchy). See Fig. 13.5 and 13.6 for an example on how the similarity clusters are embedded within the time clusters at each level of the hierarchy. On the other hand, Fig. 13.7 shows a real example with images from one event, and how they are clustered using similarity.

In figures Fig. 13.8 through Fig. 13.11, we present another example. In

Figure 13.8: Two level time clustering hierarchy, with embedded two level similarity hierarchy. At Level1, similar images (3,4,5,6,7,8), are decomposed at Level2 into singular images (3,4,8), and a cluster of near-duplicates (5,6,7) within time sub-cluster **TSC1**.

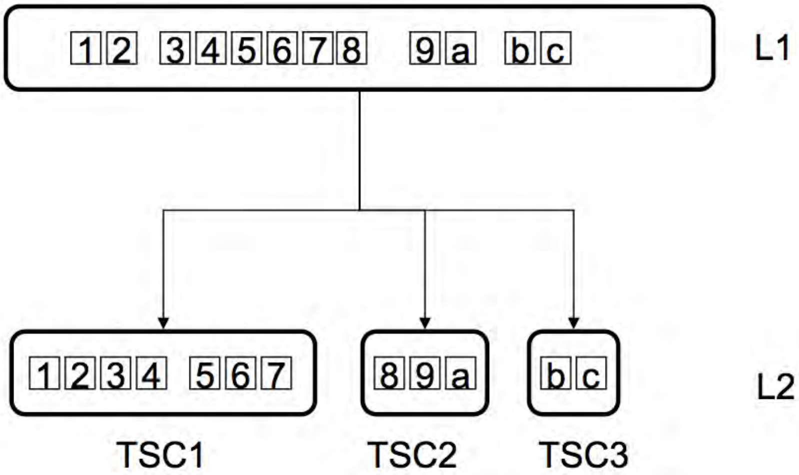


Figure 13.9: Two level time clustering hierarchy, with embedded two level similarity hierarchy. At Level1, similar images (3,4,5,6,7,8), are decomposed at Level2 into singular images (3,4,8), and a cluster of near-duplicates (5,6,7) within time sub-cluster **TSC1**.

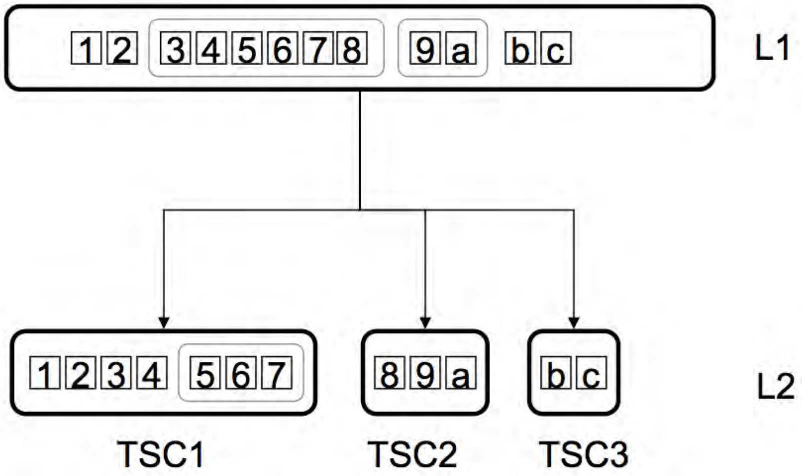
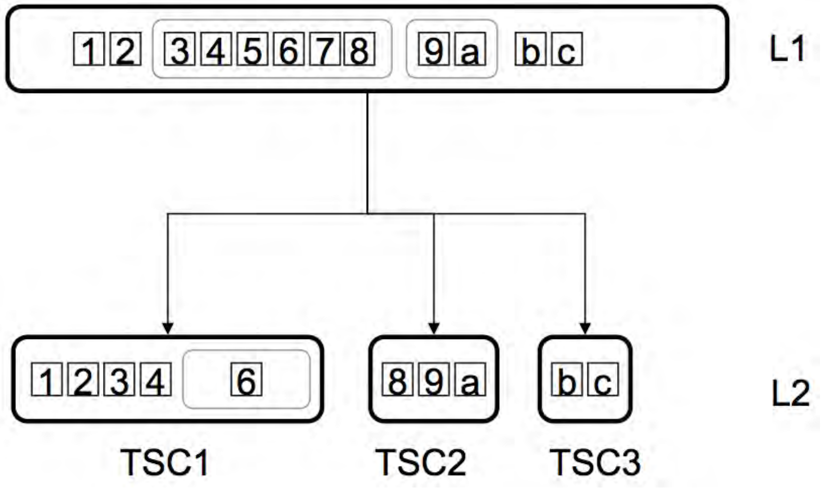


Fig. 13.8 we present the two level hierarchy (L1 and L2), with one time sub-cluster at level L1, and 3 time sub-clusters at level L2 ($TSC1$, $TSC2$ and $TSC3$). In Fig. 13.9 two similarity clusters have been created at level L1, composed of images $3, 4, 5, 6, 7, 8$, and another similarity cluster composed of images $9, a$; at

Figure 13.10: Two level time clustering hierarchy, with embedded two level similarity hierarchy. At Level1, similar images (3,4,5,6,7,8), are decomposed at Level2 into singular images (3,4,8), and a cluster of near-duplicates (5,6,7) within time sub-cluster **TSC1**.



the same time, it presents one near-duplicate cluster at level L2, composed of images $5, 6, 7$. In Fig. 13.10 a representative image –*i.e.*, the most aesthetic one– is selected from each of the near-duplicate clusters at level L2. And finally, in Fig. 13.11, a representative image –*i.e.*, the most aesthetic one– is selected from the similarity clusters at level L1.

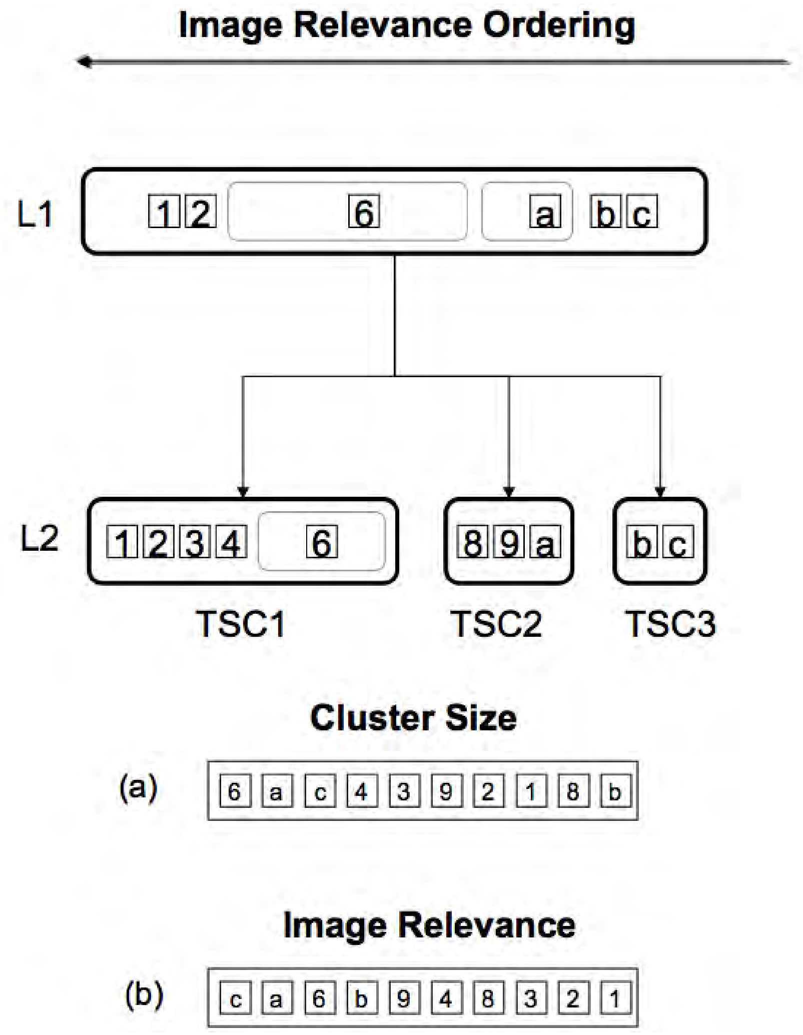
At the top of the hierarchy, time clusters, as well as similarity clusters, are very large. This level will be used when selecting a very small number of images from the collection, providing a good coverage of large scale events. Conversely, at the bottom of the hierarchy, the time and similarity clusters are much smaller, which is convenient when selecting a large number of images (a large percentage of the original collection), since it will give a better coverage of all events, both large and small.

13.2.2 Fine grain scalability

Hitting a specific image count requires fine grain scalability. This has been implemented in the form of a modular scalable relevance representation at each hierarchy node (Fig. 13.1c). The relevance modules that we have used are: face clusters, smile detection, aesthetic appeal measure, and near-duplicate image sets (*i.e.*, a by-product of the similarity clusters generated in Section 13.2.1.2).

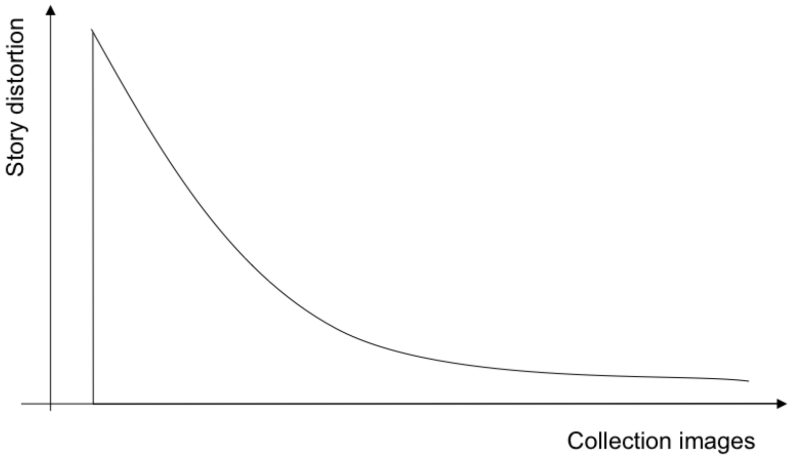
Ideally, at each hierarchy node, the images would be ordered in such a way that any given set of images that were selected from it, the outcome would always

Figure 13.11: Same time clustering hierarchy as in Fig. 13.8-13.10. For the sake of simplicity, near-duplicate images are not shown, but in a real world example, they would be relevance ordered with lower relevance than the representative images (as in Fig. 13.1). The images and clusters have been re-ordered based on image relevance (highest relevance on the right). (a) shows the result of selecting the images from the Level2 clusters using the size proportional method. (b) shows the result of selecting the images from the Level2 clusters using the average relevance method.



approximate the story in the best possible way, *i.e.*, with the least story distortion, see Fig. 13.12. These ideas have been borrowed from image scalable compression, like the standard JPEG2000 [110], in which the image coefficients within each image block are ordered in such a way that any subset of the coefficients always renders the minimum image distortion in the rate-distortion sense.

Figure 13.12: Ideal ordering of the collection images. The first image is the one that would represent the best the story on its own, *i.e.*, the story distortion would be the least; the second image is the one that combined with the first one would reduce the story distortion the most; and so on and so forth.

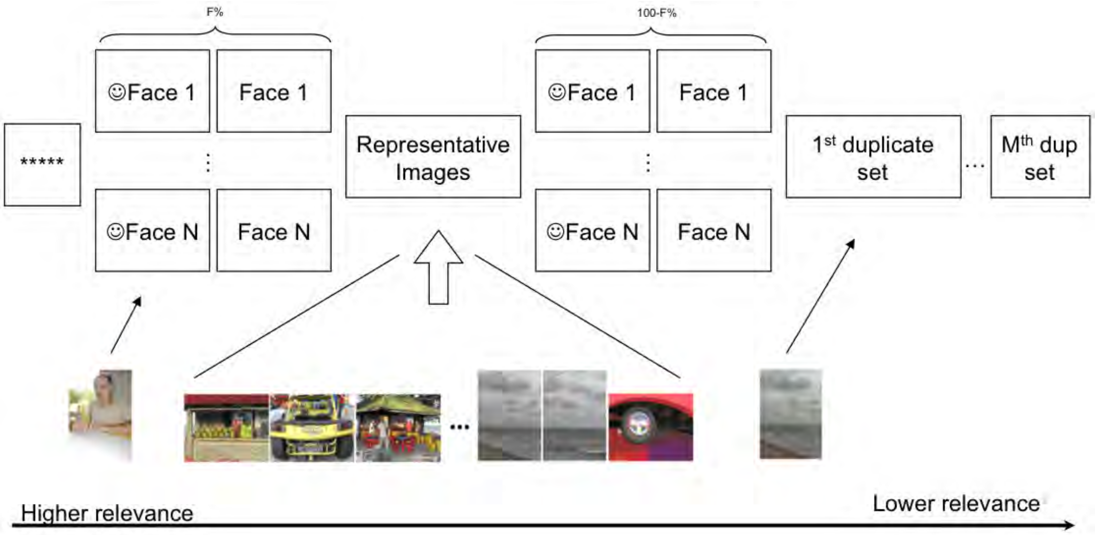


From the storytelling standpoint it is very important to cover the most relevant people, or characters of the story. Due to the subjectivity of this matter, we integrated a face clustering algorithm based on the one presented in [129]. In a personal collection it becomes obvious who the main characters are, based on the size of the face clusters. Any face cluster with more than 2x the size of the average face cluster size is marked as a character's cluster. In our experiments, though, we found that end-users really want to have control over this feature, and therefore we created a user interface in which they can select the characters of their story –see Part III of this dissertation. The aesthetic appeal of people's faces is also very relevant, and to this effect we use the algorithm presented in [116], which allows us to give higher relevance to the smiling characters' photographs, by including the smile probability into the overall face based aesthetic appeal formulation; this face based aesthetic appeal is used to rank order these characters' photos in those face clusters. Note that this face aesthetics model is a simplification of the model presented in Chapter 6 since only the smile detection probability is used.

From the general image aesthetics standpoint, it is important to select the most appealing images (*i.e.*, sharp, contrasted, colorful, good composition, as listed in [101]). In Part I of this dissertation a series of image and video aesthetics models were introduced. For the photo-storytelling approach presented in this Chapter, we used the model presented in Section 5.5, and given by Eq. 5.23.

The algorithm relevance orders the images within each time cluster (see the

Figure 13.13: Image relevance ranking within a specific time cluster; each relevance sub-block contains its images ranked by image aesthetic appeal. In this example, the user has configured the relevance sub-blocks to be ordered like (higher to lower relevance): user favorites; F% of characters' photos (first, in which their faces have high aesthetic appeal, *i.e.*, high detected smile probability; second, rest of characters' photos); representative images; rest (100-F%) of the characters; first set of duplicates; second set of duplicates, and so on and so forth. An example corresponding to cluster B2 in Fig. 13.2 is presented; note that the smiling character image has moved to the top of the relevance ordering, even though its aesthetic appeal is just average; also note this cluster has only one duplicate image.



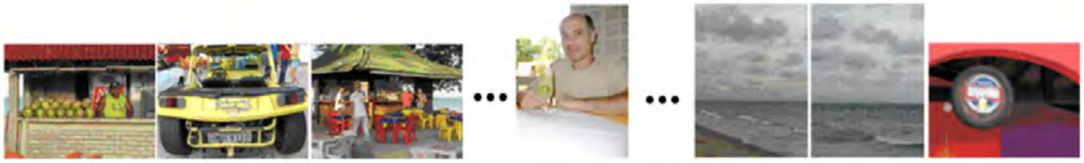
example in Fig. 13.13) by:

1. the creation of a "representative image" set, defined as the set of images not belonging to any similarity cluster, plus the most aesthetic image from each of the similarity clusters.

2. the creation M "duplicate" sets, where each cluster is composed of one image from each of the remaining similarity clusters, selected based on their aesthetic value. This is repeated until no images remain in any of the similarity clusters.

3. Finally, a parameter (F) is defined, which is the percentage of characters' images that should be more relevant than the representative images (*i.e.*, a user that likes landscape photography user may set $F = 10\%$, which means that the aesthetically appealing top 10% characters' images will be selected

Figure 13.14: Aesthetic appeal ranking (left is highest appeal) of a level 3 time cluster (B2 in Fig. 13.2), after removing near-duplicates at that level. Original time/similarity cluster size = 14 photos. Notice how for the images towards the right, some are blurrier, and/or have less contrast and/or are less colorful. In the center, with average aesthetic appeal, lies an image of the main character in this collection, smiling.



before the non-character representative images, and then the remaining 90%, while a user that likes mostly people's photographs might set $F = 80\%$).

The user has control over the final image count that needs to be selected from the image collection; the user may drive with this number an automatic layout algorithm. On the other hand, when creating a certain fixed page document with design templates, it is the document that will dictate the total number of images to be selected. The latter option is the usual situation when ordering a photo-book from one of the web-based photo-book publishers, who have standard book sizes with a few selectable templates.

The user can decide to change the default relevance of the characters' clusters that have been detected, *i.e.*, certain users might like a more artistic photo-book deciding not to give relevance to images of characters' faces, or, on the other hand, the user may decide to give high relevance to specific face clusters –main characters. The users can manually give more prominence to one or multiple time clusters, because they might be more relevant to the story they want to tell. In such case a better coverage of those time clusters will be performed, albeit with worse coverage of other clusters, since the overall number of images to be selected is fixed.

In the system implementation presented in Part III of this dissertation, the end user was allowed to select the percentage of size reduction from the original image collection to the summarized one, as well as the parameter F, *i.e.*, the percentage of characters' images to select, and also which N main characters' clusters to use, by selecting them through a graphical user interface.

Fig. 13.13 shows a specific relevance ordering which we use as the default relevance ordering. Future research in this area is bound to produce new relevance sub-blocks that would be added to this representation. In this example top relevance is given to the user's favorites (*star system* tagging), followed by N main characters (F% of the characters' faces), favoring the most aesthetic faces; the rest of representative images (with no duplicates) follow; the rest of the characters' images are next; and finally all the near-duplicates sets follow. Within each of these sub-blocks, the images are ordered by their overall aesthetic appeal given by the aesthetic model in Eq. 5.23 in Chapter 5. This figure can be contrasted with the results shown in Fig. 13.14 in which only aesthetic appeal is considered.

13.3 Image selection algorithm

Given a document size (*i.e.*, exact number of images that fit in a certain size photo-book), and given the HSR introduced above, we describe how to perform the image selection that will target the final image count. Two automatic selection algorithms based on relevance are described in Section 13.3.1. Each time cluster in the hierarchy will have its images ordered by relevance.

13.3.1 Scalable selection

Different selection algorithms may be designed to take advantage of the HSR introduced in Section 13.2, and the different relevance features introduced in Section 13.2.2 that can enhance user configurability. These algorithms generate a scalable image selection ordered list for the selected hierarchy level; they do so by selecting images from each time cluster at that level in a specific order.

This is a fully scalable representation in the sense that it can be used to fit any size photo-book in a straightforward manner: just select the first NN images of this relevance ordered list. This final selection of NN images is then reordered chronologically for storytelling purposes, and laid out to create the final photo-book.

We present the two main selection algorithms that have been used in our field experiments: one favors clusters with high average image relevance, and the other favors larger clusters.

13.3.1.1 Time hierarchy level selection

Given a certain document size, the first decision is to select the right level in the time clustering hierarchy on which to operate. Once a certain hierarchy level has been selected, the whole image selection process will happen at that level unless the end user has specified otherwise. The overall number of images (NN) to be selected for the photo-book will drive this decision. Being NC_i the number of time clusters at each hierarchy level i, with i greater or equal to one. The best time hierarchy level for best coverage of the image collection is given by Eq. 13.1.

$$NC_{select} = \begin{cases} NC_1 & \text{if } NN < NC_1 \\ NC_i & \text{if } NC_{i-1} > NN > NC_i \end{cases} \tag{13.1}$$

If the end user has given more prominence to a certain time cluster at the selected level, more images from it need to be selected (expressed as a percentage of the final photo-book size). In this case, the selection for this time cluster is performed at the immediate lower hierarchy level (*i.e.*, on its sub-clusters) in order to have a better coverage of that specific event. This would modify the way the selection algorithms weigh each of the time clusters, which are described below.

13.3.1.2 Time cluster size based selection

Once the time hierarchy level has been identified, images are going to be selected from each of the time clusters at that level. The selection algorithm will alternate among different time clusters based on specific rules. Time cluster size based selection favors the larger clusters, not counting near-duplicates/similar images. It does so in a proportional way.

Fig. 13.11 shows the same time clustering hierarchy as in Fig. 13.8-13.10. One representative image has been selected from each similarity cluster (the one with highest aesthetic appeal), and the clusters, and the images within the clusters have been reordered right (highest relevance; image **c**) to left (lowest relevance, image **1**), in order to explain the selection algorithms.

In this example, the overall number of representative images at level $L2$ is 10, and **TSC1**'s size is 5 images, or 50% of the total not counting near-duplicates/similar images; **TSC2** is 30% and **TSC3** is 20%. The algorithm starts off selecting at least one image from each time sub-cluster (for storytelling reasons, *i.e.*, good coverage of all time based events), starting with the larger one (**TSC1**): **6**, **a**, **c**; and adding images from each time cluster in a proportional way. See Fig. 13.11a. This turns out to be the preferred solution when storytelling is the main reason for the image selection.

13.3.1.3 Average Image Relevance Based Selection

Average image relevance based selection favors the clusters whose average relevance of its representative images is higher (*i.e.*, not counting near-duplicates/similar images).

In the example of Fig. 13.11b, the algorithm starts off selecting at least one image from each time sub-cluster (for storytelling reasons), starting with the highest average relevance one, and progressively selecting one image from each of the other sub-clusters with decreasing average relevance (*i.e.*, first **TSC3**, then **TSC2**, and finally **TSC1**): **c**, **a**, **6**; and adding images from each time cluster following that same order.

13.4 Experimental results

Over the period of one year, a set of experiments were run in order to validate the automatic image selection methods. A call for image collections was made within Hewlett-Packard. Several dozen collections were received (each between 800 and 3200 images). An automated system for creation of photo-books, with minimal user interaction, was implemented [C14] for this purpose (see Part III of this dissertation). The automatically selected images are automatically laid out and a certain amount of user interactivity is allowed, where the system presents near-duplicates or similar images to the user for a possible swap. The final layout is converted to PDF, printed and perfect bound into a nicely finished photo-book –see Fig. 18.4 for examples of such finished photo-books.

Certain parameters were fixed for all collections in our experiments:

1. The collection owners would select the main characters themselves, as they would in a real world scenario. Main character's image selection set to 50%. (*i.e.*, 50% of images of all selected main characters will appear in the final document).

2. Top hierarchy level to be 12 time clusters, to allow for yearly calendar experiments.

3. The selection algorithm used is the *time cluster size based selection*, presented in Section 13.3.1.

4. Three different size photo-books were created for each experiment: *LARGE* (composed by 50% of the images in the original collection), *MEDIUM* (composed by 30% of the images in the original collection) and *SMALL* (composed by 15% of the images in the original collection). Of each of these, 2 versions were create: *method A* in which the images within each of the relevance sub-blocks (Fig. 13.13), were ordered by their aesthetic appeal; and *method B* in which the images within each of the relevance sub-blocks were ordered randomly.

5. A simple layout was used (Fig. 16.1), so that it would not influence the end-user at the time of the experiment. A cover image was selected (*i.e.*, the most aesthetic landscape image of the collection), the same for all the six photo-book versions, and the books were perfect bound.

6. Only the books were shared with the end-users. The interactive user interface will be the subject of future experiments.

The end-users were asked 2 questions:

a) Select which selection, *method A* or *method B* did a better job.

b) Select which photo-book size represented their recollection of the specific event/trip better, and was more likely to be shared for storytelling purposes.

We were able to complete 18 such experiments. The main feedback from the experiments is that people like the fact of an automatically generated photo-book, but they all express the need of being able to tweak the results (*i.e.*, selecting other near-duplicates, or giving higher/lower prominence to specific time clusters).

Question (a): Select which selection, *method A* or *method B* did a better job.

18/18 users preferred the aesthetic appeal ordering at book sizes *LARGE* and *MEDIUM.*

16/18 users preferred the aesthetic appeal ordering at book size *SMALL.* After interviewing the two end users that chose the random versions, we discovered that high reduction factors (15% and below) will force the system to perform the image selection from very large time clusters, in which, sub-events may have a lot of variability in terms of aesthetic appeal (*i.e.*, most of the highly aesthetic images might actually belong to one single sub-event); therefore, a random selection will give a better representation of the event which might be more representative for the end user. Aesthetic appeal ordering makes more sense at the very small cluster level, since conditions in which those images were captured are going to be similar (*e.g.* illumination, similar content).

Question (b): Select which photo-book size represented their recollection of the specific event/trip better, and was more likely to be shared for storytelling purposes.

13/18 prefer the *MEDIUM* size book.

4/14 prefer the *LARGE* size book, but when reminded of higher price concerns, they all decided it was fine to have just the 30% version.

1/14 prefers the *SMALL* size book. This person was more interested in portability.

CHAPTER 14

Social network driven automatic photo storytelling

The main contribution of this approach to photo storytelling, over the approach presented in Chapter 13, is the fact that a certain context is analyzed in order to learn how to perform a better selection. The author did some research in the past in the area of context based image selection, or retrieval, see Appendix C of this dissertation [C15, C16, C17]; in that work, the document page in which the image had to be inserted was analyzed for document visual balance, and also for color harmony –both measures of document aesthetic appeal– and the images from a database were ranked based on a metric that combined both features. The context was, therefore, the document into which the image had to be inserted. On a different line of work, the author proposed a framework for video indexing based on the viewers' context, including their behavior and face expression [C32]. All this prior work served as the motivation for the work presented in this Chapter, in which the context are the images that the user has already uploaded to his/her online social network as photo stories. In this way, the selection algorithm can be optimized to mirror some of the photo storytelling traits that the user has portrayed in the past. The work presented in this Chapter was originally published by the author in [C2,C3], and protected by one patent application [C80].

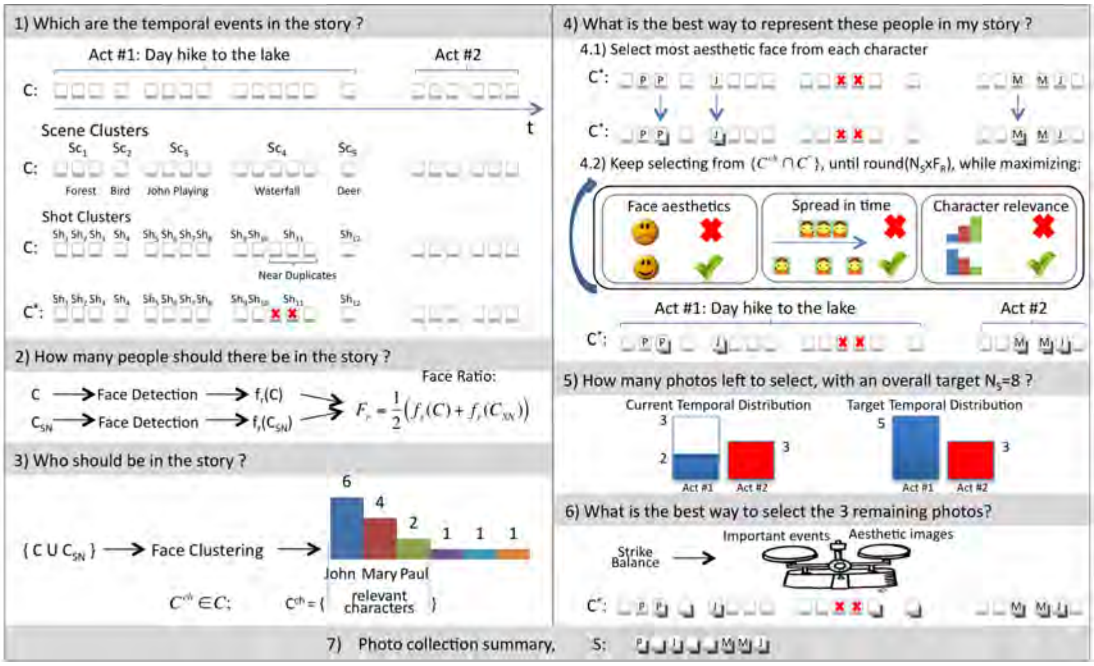
As shown in Section 14.3 and in previous research [68], users typically enjoy the creative process involved in photo story creation and they rely heavily on emotional and contextual information in order to select images [80]. Therefore, we hypothesize that the proposed system should be seen as the first component of an iterative, incremental loop based on a construct, examine and improve cycle [56], which leads to the final story to be shared. In other words, by starting from a *half baked* story or *draft*, the user would be *satisficing*[1] rather than *optimizing* the full story creation process from scratch [17]. In our user study described in Section 14.2, we corroborate this hypothesis. We also hypothesize that a human audio/video (A/V) professional with storytelling skills performs better than the

[1] Satisficing is a stopping rule for a sequential search, where an aspiration level is fixed in advance, and the search is terminated as soon as an alternative exceeds that level.

proposed system, which we could not validate in our study.

The rest of this chapter is structured as follows: in Section 14.1 describes the proposed photo storytelling system. In Section 14.2, we describe our user study and its results, whilst a few implications for the design of multimedia storytelling applications are summarized in Section 14.3.

Figure 14.1: The photo selection process (See Section 14.1.1.5 for notation explanation): 1) Event clustering; 2) number of images with faces; 3) face clustering; 4) *people photos* selection, striking a balance among face aesthetics, spread in time and *character* relevance; 5) slot allocation for remaining photos to be selected; 6) selection of remaining photos striking a balance between photos from important events and highly aesthetic photos. 7) final summary.


1) Which are the temporal events in the story ?
Act #1: Day hike to the lake
Act #2
C:
t
Scene Clusters
Sc_1 Sc_2 Sc_3 Sc_4 Sc_5
C:
Forest Bird John Playing Waterfall Deer
Shot Clusters
Sh_1 Sh_2 Sh_3 Sh_4 Sh_5 Sh_6 Sh_7 Sh_8 Sh_9 Sh_{10} Sh_{11} Sh_{12}
C:
Near Duplicates
Sh_1 Sh_2 Sh_3 Sh_4 Sh_5 Sh_6 Sh_7 Sh_8 Sh_9 Sh_{10} Sh_{11} Sh_{12}
C^*:
2) How many people should there be in the story ?
Face Ratio:
C → Face Detection → $f_r(C)$
C_{SN} → Face Detection → $f_r(C_{SN})$
$F_r = \frac{1}{2}(f_r(C) + f_r(C_{SN}))$
3) Who should be in the story ?
$\{C \cup C_{SN}\}$ → Face Clustering →
6
4
2
1
1
1
John Mary Paul
$C^{ch} \in C$;
C^{ch} = { relevant characters }
4) What is the best way to represent these people in my story ?
4.1) Select most aesthetic face from each character
C^*:
P P J M M J
C^*:
4.2) Keep selecting from $(C^{ch} \cap C^*)$, until round(N_SxF_R), while maximizing:
Face aesthetics
Spread in time
Character relevance
Act #1: Day hike to the lake
Act #2
C^*:
5) How many photos left to select, with an overall target N_S=8 ?
Current Temporal Distribution
3
2
3
Act #1
Act #2
Target Temporal Distribution
5
3
Act #1
Act #2
6) What is the best way to select the 3 remaining photos?
Important events
Aesthetic images
Strike Balance →
C^*:
7) Photo collection summary,
S:

14.1 Storytelling for social albums

The proposed photo summarization system is inspired by principles of dramaturgy and cinematography. Each generated summary, album or photo story[2] contains a set of elements that are first described in this section, followed by a detailed description of the algorithms that compose the proposed system.

[2] In the following, we shall indistinctively refer to photo stories, albums or summaries.

14.1.1 Photo story elements

A good story includes essential elements such as a certain narrative structure, with identifiable beginnings, middles and ends, and a substantial focus on characters and characterization which is arguably the most important single component of the story [78]. In the case of personal photo storytelling, many times users want to show off their experiences [47] – emphasizing good/happy times with friends and family, and aesthetic imagery [68].

14.1.1.1 Narrative structure

In our approach to storytelling, the photos are grouped into meaningful events, which will generate a certain narrative structure. We divide the story into a three level hierarchy of *acts*, *scenes* and *shots* – see Figure 14.1 part 1. Since this three level hierarchy provides a good level of granularity, we use relatively non-sophisticated – and hence faster – clustering methods to group the images, and still obtain a good overall performance.

1. *acts*: An *act* is major section of a play (dramaturgy), in which all story elements are related to one another; in our application this can be seen as a relatively large group of photos representing a well defined period in time. Users typically give some structure to their image collection by the temporal patterns (*i.e.*, bursts) with which they take their photos [50]. Hence, *acts* are detected by an algorithm similar to that in [94], as described in Section 3.1.2, where a photo is included into a new *act* if it was captured more than a certain amount of time T_t since the previous photo was captured. This allows us to target a specific number of *acts* just by varying T_t, which is an important feature as explained below.

The number of *acts*, $N_{ActClusters}$, into which the photo collection should be partitioned will depend on the average number of images per *act* $\overline{N_{Act}}$, and the overall number of images in the collection N_C: $N_{ActClusters} = \frac{N_C}{\overline{N_{Act}}}$. Given N_C and $\overline{N_{Act}}$, the proposed *act* clustering algorithm will vary the time threshold T_t until $N_{ActClusters}$ is reached.

Each summarized act should present enough photo variety so as to allow the user to indulge in as many different aspects of the story as possible, *i.e.*, important moments, or just plain aesthetically beautiful images. Therefore in section 14.1.2 a selection algorithm that alternates between important events and plain aesthetic imagery is presented. After early experiments we found that, in general, selecting 2.5 images in average from each act generates enough photo variety for the user, *i.e.*, in average the algorithm might select 1 character photos, 0.75 important event photos, and 0.75 aesthetic photos from each cluster, assuming $F_r(C, SN) =$

0.4.

The average number of images selected from each act drives the overall number of acts into which the collection should be partitioned: $\overline{N_{Act}} = \frac{N}{2.5}$, where $N = |S|$ is the target number of images that should make up the summarized story. Our time clustering algorithm varies the threshold T_t until the desired number of clusters N is reached.

2. *scenes*: Each *act* within a photo story is divided into *scenes*, in which the setting is fixed. In our algorithm a *scene* is composed of images from one specific *act* that are *similar* to each other, using global color similarity.

3. *shots*: Finally, each *scene* is divided into *shots* – borrowing now from cinematography – which are single video sequences captured by one single camera without interruption. Each consecutive frame in a video *shot* is almost identical to the previous one, and therefore we use this term in our algorithm to refer to a set of near-duplicate photos – *i.e.*, images that were taken from almost the same camera angle, presenting almost identical foreground and background.

Note that we follow a bottom-up approach to accomplish a hierarchical *scene-shot* representation. First, similar images within a specific *act* are clustered into *shots* using the normalized SIFT [81] feature similarity function described in [106]. Next, only *one representative* image from each *shot* is selected using an aesthetic measure (see Section 14.1.1.3 below). All the representative pictures selected at the *shot* level are then clustered together using a global color similarity function (the normalized histogram intersection in HSV – hue, saturation, value – color space [108], as described in Section 3.2.1), generating the *scenes* for this particular *act*. Also note that a specific *shot* or *scene* may be composed of one single image.

In this approach we also use the time-varying similarity threshold described in Section 3.2, both for *scenes*, as well as for *shots*.

14.1.1.2 Characters

The *characters* in the story are probably one of its most important elements [78]. Hence, it is not surprising that users tend to be very sensitive to the selection of *characters* in their social photo stories. For photo albums to be shared on OSNs, users tend to prioritize photos with members of the network.

Our system takes into account three *character* related features: (1) *Face ratio*: the proportion of images with people, *people photos*[3], that should appear in the story; (2) *characters*: who the people in the selected photos should be; and (3)

[3] In the rest of the chapter, we shall refer to images with people in them as *people photos* or *images*.

aesthetics: the aesthetic value of the *characters*' faces in the photos where they appear, including whether they are smiling or not.

Since the goal of our system is to help users create photo stories that will be shared on their OSN, we use two sources of information to determine the target *face ratio* and the *characters* in the story: The specific photo collection to be summarized (C) *and* the set of photos in the user's OSN albums (C_{SN}). This allows our system to approximate the user's style – *i.e.*, average face ratio in an album, which we found to be a very personal trait (see Table 14.1 in section 14.2.1) – and adapt to the target audience – *i.e.*, friends that appear prominently in the user's albums are probably socially closer, and should therefore be favored in future summaries.

The *face ratio* is given by the ratio of number of *people photos* in a collection when compared to the total number of photos in that collection. Since different photo collections do not necessarily have the same face ratios, *i.e.*, the user may have lots of people images in one collection and barely any in another collection, the target *face ratio* in the photo story, F_r, is given by a linear combination of the face ratios in C ($f_r(C)$) and in C_{SN} ($f_r(C_{SN})$): $F_r = \frac{1}{2}(\ f_r(C) + f_r(C_{SN}))$ –see Figure 14.1 part 2. In this way we reach a compromise between the user's social storytelling style and the actual collection to summarize.

In addition, a specific photo collection to be summarized does not necessarily include photos from all the people that are relevant to the user (*e.g.* family, friends). In order to identify the main story *characters*, we combine C and C_{SN} into a single photo collection $\{C \cup C_{SN}\}$, which we use to identify the user's *character* set by clustering the faces using a face detection and recognition engine based on [62] –see Figure 14.1 part 3. Each face cluster that has at least two images is considered relevant enough to correspond to a *character* important to the user. This gives a good estimation of the people the user cares about. For instance, one of these relevant people may appear only once in C but many times in C_{SN} and hence our system would include that person as a *character* in the summary. In addition, we infer the *importance* of the *characters* from the number of images in each face cluster.

Finally, the aesthetic value of the faces in *people photos* is also computed as described in Section 14.1.1.3 –see Figure 14.1 part 4.

14.1.1.3 Aesthetics

As previously mentioned, users typically share images of important events, relevant *characters*, or images that may be important to them mainly for aesthetic reasons [68]. In addition, if a low quality photograph is selected to summarize an event, it will not be a mnemonic for the user to remember that event [95]. In the case of image selection for storytelling purposes it makes more sense ranking the images, based on their aesthetic appeal, within a cluster, rather than classifying them –*i.e.*, so that any number of images can be selected from the top of the list, depending on the size of the selection, as presented in Chapter 13. Hence, we use the regression-based computational image aesthetics model described by Eq. 5.24 in Chapter 5. Our system also includes the regression-based computational face aesthetics model introduced in Chapter 6, since it has been shown that different image categories –*i.e.*, general image aesthetics vs. face aesthetics– would benefit from different aesthetic metrics –as we found in Chapter 9–, and the best high level categorization regarding aesthetics is usually obtained by partitioning the set into *people* and *non-people* photos[4] [23].

a. Face Aesthetics As we mentioned in Chapter 6, there has been some research in trying to understand facial attractiveness [111] using face features including symmetry. Unfortunately, these type of approaches would favor a *character* over another based on their looks, which would go against the storytelling principles described above. In order to avoid this kind of bias, we have used a normalized face aesthetic measure (A_f) that takes into account normalized face sharpness, combined with the relative size of the face with respect to the overall image size [C11], and smile detection [124]; see Chapter 6 for details on the face aesthetics regression model.

This face aesthetic measure turns out to be very effective when comparing aesthetics of the same *character*'s face, *i.e.*, within the same *character*'s face cluster. For the rest of the images with faces, but no *characters* in them, the algorithm rates the aesthetics of the largest face in the photo, since smaller faces might not be relevant or could have been photographed accidentally.

b. Image Aesthetics As previously explained, different methods of selecting representative images from within an image cluster have been proposed in the literature [50, 106, 95]. In this work, we take a similar approach to the one presented in Chapter 13 where the representative images within a specific event cluster will be selected based on their aesthetic value [68], and images within a cluster will be ranked based on their aesthetic value as given by Eq. 5.24 in the model presented in Section 5.5.

[4] Note that we will consider photos to be *people* photos if they have at least 1 face detected by the face detection algorithm.

14.1.1.4 Visual variety or diversity

Each summarized *act* should present enough photo variety so as to allow the user to indulge in as many different aspects of the story as possible: relevant people and moments combined with aesthetically beautiful images [68]. Therefore, the photo selection algorithm presented in the next section takes into account these three elements: relevant people and events together with aesthetically beautiful images.

Before delving into the details of our approach, we shall summarize the rest of the notation used in this chapter.

14.1.1.5 Notation

A photo collection C is formed of $N_C = |C|$ images (c_i) in capture time order[5]: $C = \{c_i,\, 0 \leq i < N_C\}$. The photo summary, S, and the collection of photos available in the user's OSN, C_{SN}, are similarly defined.

We shall define next two subsets of C, C^{ch} and C^*:

(1) C^{ch}, which is the subset of C with all the photos that have *characters* in them. It is represented as a collection of M *characters*, or face clusters, that are obtained from the combined set $\{C \cup C_{SN}\}$. Note that some of the clusters might be empty if there are no photos in C where a particular *character* appears – *i.e.*, he/she only appears in C_{SN};

(2) C^*, which is the subset of C that contains no near-duplicate photos, *i.e.*, in C^* all *shots* contain only one image.

As previously explained, C is subdivided into a series of *acts*, each *act* into a series of *scenes*, and each *scene* into a series of *shots*:

$Act = \{Act_i,\, 0 \leq i < N_{ActClusters}\}$,

where $N_{ActClusters}$ is the number of *acts* in C. *Scenes* and *shots* are similarly defined.

One of the constraints that we impose on the photo summary S to be created is to preserve the temporal distribution of photos – characterized by normalized histograms – in *acts*, *scenes* and *shots* of the original collection C, where:

$$H_{Act}(C) = \left\{ \frac{N_{Act_i}}{N_C},\, 0 \leq i < N_{ActClusters} \right\}$$

is the histogram of *acts* in collection C. $H_{Scene}(C)$ and $H_{Shot}(C)$ are similarly defined.

[5] From now on, all our representations of image or event lists will be in capture time order since it is the most common way of ordering personal photos [99].

Finally, the generated summary should also approximate the user's *character* normalized histogram, $H_{Character}(C \cup C_{SN})$, while maximizing aesthetics and variety, as explained in the following section.

14.1.2 The Photo Selection Algorithm

Given a particular user, his/her online photo collection C_{SN} and a photo collection to be summarized C, the goal of the photo selection algorithm is to generate a photo summary S, from C, that contains a pre-defined number of photos, $N_S << N_C$, and conveys the essence of the story to be shared by the user on his/her OSN. This is achieved by ensuring that the photo summary satisfies the following requirements:

1. *People vs. non-people*: The summary's *face ratio* $f_r(S)$ should approximate the target *face ratio* F_r;
2. Characters: $H_{Character}(S) \approx H_{Character}(C \cup C_{SN})$. Note that we use C instead of C^* because near-duplicate photos of *characters* informs us of their importance.
3. Narrative: $H_{Act}(S)$, $H_{Scene}(S)$ and $H_{Shot}(S)$ approximate $H_{Act}(C^*)$, $H_{Scene}(C^*)$ and $H_{Shot}(C)$ respectively. In this case, C^* is used for *acts* and *scenes* because it better represents their event distribution.
4. Aesthetics: High normalized aesthetic value of the summary (A_S), considering both aesthetics of faces in *people photos*, and *non-people photos* image aesthetics.
5. Variety: The selected images are visually diverse.

Therefore, the problem at hand may be described as to select N elements from C in an optimal way to create the summary photo story S, while satisfying all the requirements.

The objective function to maximize is formulated as follows

$$\begin{aligned} O(C,S,SN) = {} & \alpha\, A_S - \\ & \beta\,(F_r(C,SN) - f_r(S)) - \gamma\, d(H_{Character}(S), H_{Character}(C \cup SN) - \\ & \delta\, d(H_{Act}(S), H_{Act}(C^*)) - \epsilon\, d(H_{Scene}(S), H_{Scene}(C^*)) - \\ & \zeta\, d(H_{Shot}(S), H_{Shot}(C)) - \kappa \sum_{S_j \epsilon S} \sum_{S_{k>j} \epsilon S} sim(S_j, S_k) \\ & \qquad\qquad subject \quad to \quad |S| = N \end{aligned}$$

Where C^* is the subset of C after selecting only the most aesthetic photo within each of the shots; therefore, $H_{Act}(C^*)$ is a much more faithful representation of the event temporal distribution, since it avoids near-duplicates. $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ and κ are weighting factors, A_S is the normalized aesthetic value of S, $d(.)$ is the normalized L1 distance metric between histograms, and $sim(.)$ is an *ideal* image similarity measure that takes global color information and local geometric descriptors into account. The last term, borrowed from [106], ensures diversity in the selected images, and at the same time it minimizes the number of selected images coming from the same *scene* or *shot*.

In order to solve this closed form equation we resort to two greedy algorithms which take advantage of some heuristics, as well as of the different categories introduced above, *i.e.*, faces vs. no-faces. This is implemented in a two-step process, in which we first select the *people photos* that will appear in S (step 1 below), and then select the rest of images up to N_S images (step 2 below).

14.1.2.1 Step 1: People photo selection

The goal of this first step is to add to S all the needed *people photos* by selecting $N_S^f = round(N_S \times F_r)$ faces from C^*, according to the following algorithm:

1.a. Rank the face clusters in $\{C \cup C_{SN}\}$ by number of images. Select the image with the most aesthetic *character* face that belongs to $\{C^{ch} \cap C^*\}$ from each of the face clusters – starting from the largest cluster in the rank, which ensures coverage of relevant *characters* in the story while avoiding near-duplicates –see Figure 14.1 part 4.1. If the image has already been selected – *i.e.*, there are two or more *characters* in the same picture, pick the following image in the aesthetically ordered list from the most popular *character* of the two – *i.e.*, largest histogram bin in $H_{Character}(C \cup C_{SN})$, and so on and so forth.

1.b. Keep selecting face images from $\{C^{ch} \cap C^*\}$ while $|S| < N_S^f$ and maximizing the objective function O_f, see Figure 14.1 part 4.2:

$$\begin{aligned} O_f(C, C^*, S, C_{SN}) = \alpha_f \; A_f(S) - \\ \gamma_f \; d(H_{Character}(S), H_{Character}(C \cup C_{SN}) - \\ \delta_f \; d(H_{Act}(S), H_{Act}(C^*)) \end{aligned}$$

where $A_f(S)$ is the normalized aesthetic value of the considered face in the people images in the summary, and $d(.)$ is the normalized L1 distance metric between histograms. More importance is given to the *character* histogram distance ($\gamma_f = 1$), followed by the face aesthetic value ($\alpha_f = 0.8$), and the *act* histogram distance ($\delta_f = 0.5$). This last term is important to ensure a certain amount of temporal coverage by the *characters*, since images with highly aesthetic people faces may be confined to specific *acts* – *i.e.*, better *vs.* worse light conditions at different times of the day. These parameters were fine-tuned by analyzing the images in four photo collections that were summarized by their collection owners –the same ground truth that was used in Section 3.2– along with the images in their OSN photo collections.

Note that if not enough *character* images are present, then the photos with most aesthetic faces of other people will be selected. If there are not enough *people photos* in the collection, *i.e.*, $N_S^f < round(N_S \times F_r)$, the algorithm moves on to step 2.

14.1.2.2 Step 2: Non-people photo selection

The previous step has selected the first N_S^f images of S. Now the algorithm will select the rest of the images $(N_S - N_S^f)$ from C^*.

From here on, we define a large *scene* (*L-scene*) or large *shot* (*L-shot*), as *scenes* or *shots* with at least 3 images, which ensures the importance of those sub-events, and avoids potentially noisier smaller clusters.

2.a. Similar to [50] and in order to ensure good temporal coverage of all *acts*, we start by ensuring that each *act* has had one representative image selected in step 1 above. If not, we allocate one image slot, out of the $(N_S - N_S^f)$ available empty slots, for each of the empty *acts*. If not enough empty slots are available, then the larger *acts* are favored.

2.b. Next, we optimally allocate the rest of the empty image slots to each *act* –see Figure 14.1 part 5– by minimizing:

$$O_a(C^*, S) = d(H_{Act}(S), H_{Act}(C^*)) \qquad subject \quad to \quad |S| = N_S$$

For each Act_i in C, we keep selecting images until Act_i has all its empty image slots filled. Similar to [80], and also similar to the algorithm presented in Chapter 13, the images are selected based on their aesthetic value. The algorithm alternates between *L-shots* or *L-scenes* and highly aesthetic images in order to provide good selection variety, as well as never selecting more than one representative image from a particular *scene* – see Figure 14.1 part 6:

- 2.b.1. Select the most aesthetic image from the largest unrepresented *L-shot* from an unrepresented *scene* in Act_i. If not available, then select the most aesthetic image from the largest unrepresented *L-scene* in Act_i. If not available, move to the following step.

- 2.b.2. Select the most aesthetic image in Act_i from any of the unrepresented *scenes*.
- 2.b.3. Go to 2.b.1.

We found that giving higher relevance to the largest *L-shot* is important since they usually represent the same object or landscape portrayed from the same viewpoint, implying a certain level of relevance for the user [106]. Conversely, highly aesthetic images tend to appear in smaller clusters or alone, and hence the alternate search for relevant and aesthetic images. Finally, all selected images are reordered chronologically before being presented to the end user. In the following sections, we describe an in-depth user study targeted at understanding the advantages and disadvantages of the proposed photo storytelling system.

14.2 User study

We designed and carried out a user study to assess the strengths and weaknesses of the proposed system. In order to motivate the need for an automatic storytelling approach, we also wanted to investigate if users consider the task of creating a photo story to be laborious and time consuming. Previous work supports this assumption [47] and confirms that photo retrieval is neither a fast nor an easy task [125]. With our study, we wanted to verify whether users perceive the effort and time demand associated with the photo storytelling task as the main source of workload instead of their concern to create a good story. In other words, if photo storytelling is too hard and users are not that demanding with the final results, an automatic approach could be appropriate. Hence, we formulate our first hypothesis as:

$\mathbb{H}1$ Users consider the task of creating personal photo stories to be laborious and time consuming, and this effort is more important than their concern to create a good photo story.

With respect to the proposed system, we wanted to evaluate the assumption that its storytelling features improve the quality of the automatically generated photo stories when compared to a simple automatic approach. Hence, the second hypothesis is formulated as:

ℍ2 Users prefer personal photo stories generated automatically by the proposed system *more often than* those generated by a random selection of photos in chronological order.

Furthermore, we carried out a comparison between the stories generated by the proposed system and a human expert in A/V photo story creation. Clearly, a human A/V professional can better filter photos based on aesthetics than our proposed system, such as removing high quality photos where the main *character* looks fat or with the eyes closed. Moreover, (s)he can use his/her storytelling skills to combine photos that were taken with different timestamps but at the same place, or even select photos that compose stylish stories with an artistic taste. Therefore, the third hypothesis is thus formulated as:

ℍ3 Users prefer personal photo stories created by a human A/V professional –who creates video/photo stories for a living– *more often than* those generated by the proposed system.

Our intention with hypothesis ℍ3 is to test whether an automatic approach that is aware of the user's photo sharing patterns can achieve a performance level similar to a human that does not take this information into consideration.

Finally, we believe that the photo stories generated by our system can be appreciated by users as an initial draft instead of creating the entire photo story by themselves from scratch. Therefore, we state our final hypothesis as:

ℍ4 Users prefer to reuse a personal photo story generated by the proposed system and upload it to their OSN after making the appropriate changes instead of creating the photo story from scratch.

Note that ℍ4 introduces the goal of sharing photo stories in social networks. Therefore, we asked participants in the initial questionnaire if they agree with this assumption.

In order to validate these hypotheses, we conducted two lab studies to: (1) measure the workload associated to the photo story creation process and (2) obtain the users' level of satisfaction with photo stories generated automatically and by the A/V professional. The next sections describe in detail the user study design and discuss the main results.

14.2.1 Participants

Twelve subjects (male: 6) were recruited via e-mail advertisement inside a large company. Subjects were considered eligible if they had an account on at least one OSN, were currently sharing photos with peers, and had at least 200 photos in their personal repository from a specific event that they would be willing to use during the user study (*e.g.* a vacation trip, a wedding party, a night out). Each participant was offered 20 Euro (about 27 USD) to be part of the experiment and a prize of 100 Euro (about 135 USD) was raffled among all of them. Mean age was 30 years old ($s = 3.89$) and occupations were somewhat varied, including students, researchers, software developers, a technology expert, a professional from human resources, a secretary, and a teacher. Participants self-assessed their photo shooting skills as slightly below average (1:novice, 5:expert, $\tilde{x} = 2.5, q1 = 2, q3 = 3$), and their ability to differentiate photos by image quality – *e.g.* contrast, sharpness, composition – as average ($\tilde{x} = 3, q1 = 3, q3 = 4$). They typically took photos from one to three times per month.

All participants had a Facebook account and 92% considered it as their main OSN, which they used to access at least two days per week ($\tilde{x} =$ every day). On average, they had 13 online photo albums each ($s = 9.746$), 36 photos per photo album ($s = 18.17$), and about 100 photos per folder in their personal collection ($s = 63.66$). These setting are similar to those from the experiment conducted by Kirk *et al.* [68]. Table 14.1 characterizes the profile of the participants and the event associated to the 200 photos that they lent to the user study. Note that most of the collections were about trips and holidays, which are a common source for storytelling between friends and family [68].

In addition to the 12 participants, an A/V professional with storytelling skills was recruited with the goal of creating one photo story for each of the participants' collections.

14.2.2 Apparatus

Photo stories were created and evaluated by the participants using the same apparatus, including a 21.5 inch flat panel monitor with a resolution of 1680x1050 pixels, a standard mouse with a scrolling wheel button and a keyboard. The Windows Explorer application (Windows Vista version) was used by the participants to create photo stories (see Section 14.2.3.1). Interviews were audio recorded.

14.2.3 Procedure

The lab study was divided in two trials. In the first trial, participants created a photo story using photos from the personal collection they lent to the user study. Workload was measured both objectively and subjectively to shed some light on $\mathbb{H}1$. In the following week, each subject attended the second trial in which they were presented with photo stories generated automatically and by the A/V

Table 14.1: Profile of the participants and their stories. Where *Face Ratio* is $f_r(C_{SN})$, and *#characters* is the number of characters with representation in C^{ch}.

Subj.	Sex	Age	Story Title	Story Relev.∗	Face Ratio	#char- acters
1	F	28	Peru	4	45.1%	8
2	F	25	Mexico	5	84.1%	8
3	M	29	Rome	5	46.5%	9
4	M	32	California	4	18.5%	3
5	F	26	Mjsa Lake	3	47.7%	3
6	M	23	Rome II	2	57.3%	9
7	M	33	Sardinia	4	31.8%	7
8	M	37	Russia	3	14.6%	1
9	F	29	Bolivia	5	57.8%	6
10	F	33	Calabria	3	68.9%	7
11	M	30	Nepal	2	40.0%	5
12	F	31	Seattle	5	59.4%	5

∗ Relevance rated by the subjects (*Not important:* 1, to *very important:* 5).

professional, and were asked to evaluate how good they were. This procedure was adopted to provide answers to $\mathbb{H}2$, $\mathbb{H}3$, and $\mathbb{H}4$. Next, each trial is explained in more detail.

14.2.3.1 First trial: Workload measurement

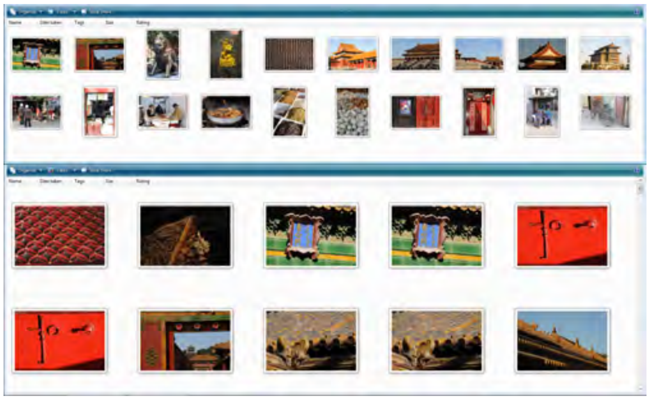
Workload is an individual experience and therefore very hard to be quantified effectively in different activities by different subjects. In the first trial, we used the NASA Task Load Index [5] to subjectively measure workload in a storytelling task, thus gathering information in six different dimensions: mental demand, temporal demand, physical demand, performance, effort, and frustration level. Furthermore, we also measured task completion time and logged the participants' interactions with the interface, including keystrokes, mouse clicks, mouse moves and mouse scrolling.

All participants were assigned the task of creating one story of 20 photos – from their initial pool of 200 – that they would be willing to share on their main OSN. The Windows Explorer application was used by the participants to create the photo stories for three main reasons: First, participants were familiar with it, thus reducing the interaction learning curve; second, it implements all interactions available on the Facebook Photo Album webpage (*i.e.,* photo selection, photo maximization, and *drag'n'drop*); and third, its popularity eases replication of the study by the scientific community. Figure 14.2 shows an example of the interface used by the participants.

Note that the pool size of 200 photos matches the total number of photos that one could share on a Facebook[6] photo album [4] and is also an approximation of the total number of photos per event considered in the work by Kirk *et al.* [68]. Moreover, the 20-photo story might have imposed a challenge to some participants as they shared an average of 36 photos per photo album in their main OSN ($s = 18.17$). As a consequence, if these participants were willing to share the automatic 20-photo generated stories, instead of manually creating them (with on average 36 photos), we would be reducing information overload by 44%.

After accomplishing the storytelling task, participants filled the NASA TLX questionnaire [5] and commented on the experience. Each session lasted an average of 34 minutes.

Figure 14.2: Storytelling interface used by the participants. Bottom window contains the initial pool of 200 photos (only 10 thumbnails can be seen at a time, as in Facebook), while the upper window retains the 20 photos that belong to the story created by the participants (ordered from left to right, top to bottom).



14.2.3.2 Second trial: Evaluation of stories

In the second trial, three photo stories with 20 photos each were generated for each collection of 200 photos. The following approaches were used to generate the stories:

[6] Facebook was taken as point of reference because 11 out of 12 participants used it as their main OSN.

1. **Random:** Photos chosen randomly and presented in chronological order;

2. **System:** Photos chosen and ordered by the proposed system;

3. **Professional:** Photos chosen and ordered by the human A/V professional with storytelling skills, who received instructions to create *appealing* photo stories that best *describe* the titles provided by the participants (see Table 14.1).

The second trial was conducted as follows: One week after the first trial, participants attended the second lab session that took an average of 36 minutes per session. First, they browsed their 200 photos to remind themselves of the event and the photos available to compose a story. Next, they were presented with one photo story containing 20 photos from the initial pool of 200 and were asked if they would share it in their social network. This procedure was repeated for the two remaining photo stories. After evaluating each story, subjects were asked to select the stories they would be most comfortable and least comfortable to share on their OSN. Finally, they answered if they would prefer to use the System story as an initial draft to compose the story or if they would rather create the photo story from scratch.

In the beginning of each session, the expert told participants that all photo stories were generated automatically. Deception was used in this case to avoid biasing the results towards either of the approaches. In addition, the presentation order of the photo stories was rotated in a Latin square basis to avoid biasing the results. Finally, participants were gender balanced in each of the presentation ordering groups to avoid gender biases.

Please see the original publication [C2] for an explanation on the statistical analysis we used, along with the numerical results they generated.

14.2.4 Results and discussion

The results obtained are presented and discussed in this section with the aim of evaluating each of the hypotheses stated in Section 14.2. The following subsection serves as an introduction to the validation of each of these hypotheses.

14.2.4.1 Online photo sharing behaviors

When asked about their OSN photo sharing habits, participants reported rarely reordering photos in their shared photo albums (1: never, 5: always $\tilde{x} = 2, q1 = 1.25, q3 = 3$) and mainly using chronological ordering (67% of the participants). This behavior can be explained by their major complains on the difficulties of sharing photos in OSNs:

1. *Effort to select photos:* This was the most cited issue by the participants.

Half of them indicated that there is a lot of effort in this process, including time demand (participants 3, 6 and 9), identification of the most special/appealing photos (participants 4 and 11), and selection of the best photo from the available near-duplicates (participant 5).

2. Delay to upload photos: Four participants mentioned the time demand to upload photos (participants 1, 2, 10, 12), and one of them also pointed the fact that sometimes errors happen in this process.

3. Effort to organize photos: Three participants were concerned about the effort to reorder photos after uploading them (participants 1, 2 and 6). One of the arguments reveals that shared and private photos do not necessarily follow the same organizational schema: *"...you have to reorder them if they are not stored in one single folder before uploading"* (participant 2).

Note that the first and third most cited problems by the participants motivate the study presented herein.

14.2.4.2 Validation of $\mathbb{H}1$

The task of creating photo stories is laborious and time consuming. As mentioned before, the effort to select/organize photos and the time required to do it are the self-reported main problems that our participants had with online photo sharing. However, participants were neither satisfied nor unsatisfied with current social networks as a means to share their photo stories (1: very unsatisfied, 5: very satisfied, $\tilde{x} = 3, q1 = 3, q3 = 4$). This is somewhat in accordance with our results from the NASA TLX questionnaire, in which the overall workload to create the 20-photo personal story from an initial pool of 200 photos was considered slightly low (0: low, 100: high, $\bar{x} = 35.2$, $s = 3.92, min = 16, max = 58$). Hence, although participants consider the time demand and mental demand in the photo selection process to be relevant problems, they do not seem to be important enough to make them dislike current online related services. These problems were also validated by the analysis of the workload source, in which performance (concern to create a good story), mental demand, time demand, and effort received the highest weights (no significant difference between them). Furthermore, objective workload measures are consistent with these results, given that a strong positive correlation was found between task duration and workload. In other words, the longer the task, the higher the workload.

From these results, we reject $\mathbb{H}1$ and rewrite it as:

$\mathbb{H}1_{new}$ Users consider the task of creating personal photo stories to be *mentally* laborious and time consuming, and this effort is *as high as* their concern to manually create a good photo story.

We revisit $\mathbb{H}1_{new}$ after validating $\mathbb{H}4$ to clarify whether the participants' concern to create a good photo story is such that they would persist in doing it by themselves.

14.2.4.3 Validation of $\mathbb{H}2$

The proposed system performed better than the random approach. A majority of nine participants (or 75%) preferred to share on their main OSN the stories generated by the proposed system instead of the stories by the random approach. Moreover, the difference between the subjects' rankings to the stories generated by these approaches was *significant*, thus confirming that the better performance of our system compared to the random approach is not due to chance. After carefully analyzing the participants' comments on the reasons why the Random method generated worse stories, we realized that the key advantages of the proposed system include:

1. Image aesthetics analysis: Several participants wanted to remove photos from the Random story that they considered to have low aesthetic value, *i.e.,* photos that were blurred (participants 2, 3, 4, 6, 8, and 11), had poor composition (participants 3 and 6), were too dark/bright (participant 6), or less colorful (participant 4). Conversely, it was rare the case when participants complained about the image aesthetics of the photos belonging to the System stories (only participants 1 and 6 considered one of the pictures to be, respectively, too bright and out of focus).

2. Balance of photos per act: In the Random stories, participants complained about the absence of photos showing all the places they visited (participants 7, 10 and 12), or the over representation of certain parts of the event: *"This photo I would keep as well... But there are too many of the aquarium... so... I would keep only these two"* (participant 12). There was also the case when participants were presented pictures they did not consider relevant for the event (participants 3, 5, 6, 9 and 10) or even photos they did not remember: *"This one I would remove because... I didn't even know of this, you know? It's a photograph that I almost don't even remember of being part of the 200."* (participant 9). Conversely, the System stories did not include photos from less relevant memories because it balanced the number of photos per *act* according to their relative importance (*i.e.,* ratio of photos taken per *act*, *scene* and *shot*).

3. Near-duplicate detection: Six subjects experienced near-duplicates in their Random story and opted for the one with better image quality (participants 7 and 10), well centered (participants 3 and 6), with their friends (participant 1), or without himself to highlight the landscape (participant 11).

4. Face aesthetics modeling, including smile detection: Poor face aesthetics was mentioned by four participants when analyzing the Random stories. More specifically, participant 2 did not like the way she looked in one of the random photos, participant 7 was concerned with the weird face his son was making, and participants 5 and 9 avoided photos that both themselves and their friends/family were not looking good: *"Oh no! My friend is not going to be happy with that photograph. She looks drunk! And she... and evil! Like, she'd kill me!"* (participant 9). None of the Random stories' rejected photos were selected by the System, with the exception of one from participant 5. Although image aesthetics was good in that photo, the main *character* – her boyfriend – was smiling in a somewhat aggressive way, and thus she would not be comfortable in sharing it with friends.

5. Character selection: By using face detection and generating photo stories with more/less people according to the *face ratio* of the participant's photo collections, our system was able to better adapt to individual preferences –see Table 14.1. Moreover, the use of face recognition and clustering helped in identifying the most relevant *characters*. This feature was well appreciated: *"This one –System story– is better because there are more photos where I am with my girlfriend."* (participant 3); Conversely, the Random approach had no leverage to opt between photos with different people: *"I don't know these people –Random story. I know I took these photos, but I wouldn't share them in my social network because my friends don't know them."* (participant 5).

Considering the prior observations and the fact that a *significant* difference was revealed between the participants' preference for the System and the Random stories, we corroborate hypothesis $\mathbb{H}2$.

14.2.4.4 Validation of $\mathbb{H}3$

No evidence was found that the human A/V professional performs better than the proposed system. Seven participants (58%) liked the System stories more than the Professional stories while the remaining five participants (42%) preferred the Professional stories. The difference between these preferences is not significant and therefore suggests that participants liked the Professional stories *as often as* they liked the System stories. At least two facts suggest that the Professional approach did not perform better than the System approach: (1) although not significant, the majority of participants preferred the System stories; and (2) the number of participants that would not like to upload the Professional story to their social network –participants 6 and 9– was the same when compared to those that also would not upload the System stories –participants 9 and 11[7].

[7] In this case, participants were allowed to consider both removing and reordering any of the 20 photos, but not including others from the initial photo collection, which

Therefore, we neither corroborate nor deny $\mathbb{H}3$, but instead highlight the tendency to obtain similar results with both Professional and System approaches.

Furthermore, after carefully listening to the audio recorded sessions, we confirmed that the Professional stories were lacking information regarding the participants photo sharing preferences, which was better captured by System stories:

GOOD: *"This one –System story– is better because there are more photos where I am with my girlfriend."* (participant 3)

GOOD: *"I prefer this story –System– because it gives me a more... warm feeling about it. The combination of colors, the brightness of pictures, and the... there is more people here, but usually I don't like much to have people in my pictures, but maybe for sharing"* (participant 4)

GOOD: *"I like it –System story– because there are more people that I know."* (participant 5)

BAD: *"This story –Professional– is not exactly what I would do because it lacks pictures with people. The pictures are really nice, but as far as the social network, I am much more interested in looking for some people out there. What I am looking for is some sort of experience."* (participant 6)

BAD: *"This one –Professional story– is too focused on me, isn't it? Very selfish. I prefer to do it like half and half. For instance, there is no photo with my husband. Oh, no. There's one here. In the end. But he's alone."* (participant 12)

The statements above confirm the importance of the *character* selection technique used by the System stories, which was based on both C and C_{SN}.

14.2.4.5 Validation of $\mathbb{H}4$

Participants preferred to use the System story as an initial draft instead of manually creating the entire photo story from scratch. Although results from $\mathbb{H}1$ reveal a slightly low effort to manually create the photo stories from scratch, 75% of the participants reported preferring to reuse the System story and make changes to it. Some of the reasons included the good aesthetics of the photos chosen by System (participants 1, 3, 4, 5, 8 and 12), the summarization/arrangement of sub-events (participants 1, 2 and 12), the possibility of reducing the effort associated to the photo selection process (participants 2, 6 and 10), and the lack of time (participants 2, 5, 6, 8, 10 and 12). These observations validate $\mathbb{H}4$ and confirm that both mental and time demands are indeed the sources of the storytelling workload, as indicated by the first trial.

would require them to browse the remaining 180 photos and thus be exposed to the information overload problem.

Finally, from the results obtained for $\mathbb{H}4$, a final version of $\mathbb{H}1_{new}$ can be written as:

$\mathbb{H}1_{final}$ Users consider the task of creating personal photo stories to be *mentally* laborious and time consuming, and this effort is *as high as* their concern to manually create a good photo story.

14.3 Implications for design

The findings presented herein support a few guidelines that might help designers and multimedia technology experts to build social storytelling solutions, including:

Focus on face aesthetics. Seven participants (58%) complained about face aesthetics in the System stories (participants 1, 2, 5, 7, 9, 11 and 12). While some of the reasons might be easily covered in future automatic algorithms (*e.g.* eyes closed detection), others are harder (*e.g.* detection of "weird" smile, goofy face, fat face, *etc.*). Face aesthetics was considered relevant not only for the main *character*, but also for the peers: *"When it comes to people, I draw the line."* (participant 9). We believe participants were concerned with their aesthetics mostly because of the main goal of the task: create a story to share on their social network. Given that self-promotion is one of the main reasons for sharing multimedia content in social networks [25], users will definitely appreciate automatic approaches that highlight photos in which they look better.

Reduce information overload automatically, but support the user's creativity with story customization. Even though the proposed system has proven to be effective, sharing photo stories is a social activity, as exemplified by a comment from participant 2: *"You know, some things you want to share only with your friends, not your family"*. Photo story personalization is key, specially because no one else knows the event captured by the photos better than the users themselves. The validation of hypotheses $\mathbb{H}2$, $\mathbb{H}3$, and $\mathbb{H}4$ confirms that users would benefit from the proposed approach, but would want to control the storytelling task, which confirms the results we presented in Chapter 13. This is supported by the fact that, from the nine participants that would upload the System story to their OSN, eight would either remove, reorder or swap photos from the generated stories. Note that we considered *personal* photo stories where our participants were the main *characters*. That might have been the reason for the extra motivation to edit the stories. Future work shall evaluate the combination of automatic and manual approaches towards increasing productivity, reducing information overload and supporting the user's creativity.

Combine automatic approaches with collaborative storytelling. The participants preference for the System story instead of the Professional story was strongly associated with their preference to reuse the System draft instead of creating the story from scratch. In other words, those who preferred the Professional story also preferred to create the story by themselves instead of reusing any of the generated stories. Hence, our findings suggest that there is a subset of users that do not seem to benefit from automatic solutions, but are more likely to benefit from human-generated ones. Therefore, we expect automatic multimedia storytelling solutions to benefit from a collaborative component in order to better fulfill the users' needs. Previous work has already tackled the problem from a purely collaborative perspective [61, 102, 104], but our results suggest that the combination of automatic and collaborative approaches might lead to a more appropriate balance between increased productivity, information overload reduction and subjective satisfaction with the final story.

Part III

System Implementation

CHAPTER 15

Introduction to our system implementation

In order to implement the described algorithms, a front-end/back-end architecture was devised. The front-end, *i.e.*, main user interface, has beed designed in a very slim format, *i.e.*, it does not require high computation on the output device; while the back-end is running on a muti-core server with high resources at its disposal. Front-end and back-end communicate through an XML Web services API, using Representational State Transfer (REST) [45] interfaces. This architecture enables the system to potentially run on a wide variety of clients, ranging from PCs to tablets to cell-phones.

We implemented the algorithms and data structures presented in Chapter 13, as a system for automatic image selection for photo book creation, along with an intuitive user interface for fine tuning of the selection results. This system is the property of the Hewlett-Packard company, and therefore only a high level description is presented in this dissertation.

The images are selected based on their relevance, while preserving a good coverage of the event –time and people– as presented in Chapter 13. The selected images are laid out and presented to the user through an Adobe Flex user interface [3], which allows them to select images and swap them by semantically related ones in an intuitive manner. The final result is output to a PDF file, which was used to print the Chapter 13 User Study photo-book samples on an HP LaserJet 5000 printer in *Photo Mode*.

The implemented system, which we named *SoftBacks* and was originally published in [C14], selects the images and generates a soft proof of the final photobook –see Fig. 16.1– which in turn is interactive, and allows the users to replace images by semantically related images in an intuitive way.

CHAPTER 16

Front-end description

SoftBacks interaction follows a simple flow: setting image selection criteria, choosing important face clusters, and interacting with the generated realistic soft proof.

When setting the selection criteria the end users specify two parameters:

1. the selection percentage with respect to the whole collection, and
2. the percentage of characters images,
3. and, finally, the users also select the images clusters of the main characters the story should be about.

Once the user has set those criteria, they are presented with a list of ranked face clusters, where they can select the characters they want to favor in the photo book. Finally, the automatic selection is performed, and the results are laid out as seen in Fig. 16.1 and 16.2, using the Adobe FLEX FLEX-Book component [3].

Figure 16.1: User Interface of the interactive *SoftBacks* [C14] application, showing the automatic layout. This is the automatic layout used in our user studies in Chapter 13, which was output to PDF format, and printed on an HP LaserJet 5000 in *Photo Mode*.

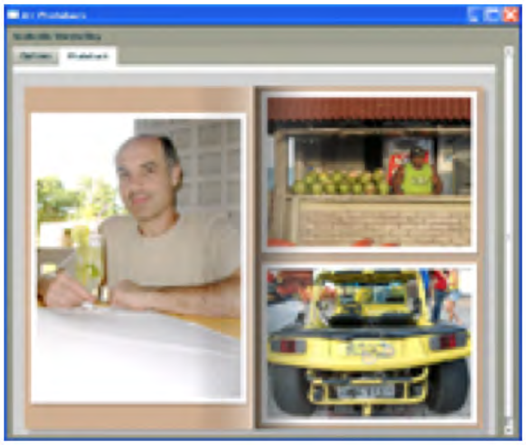


Figure 16.2: User Interface of the interactive *SoftBacks* [C14] application, showing user interaction by turning the pages of the photo book.

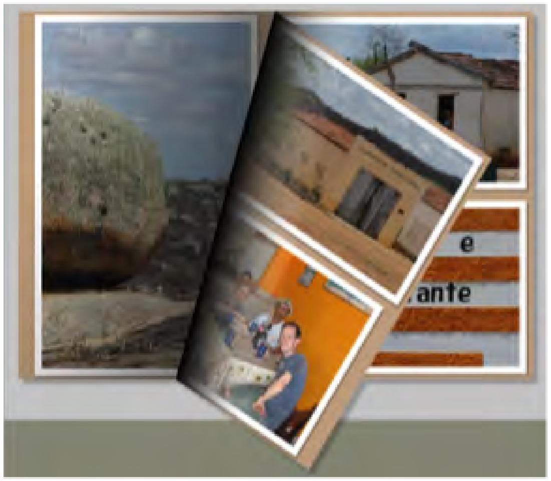
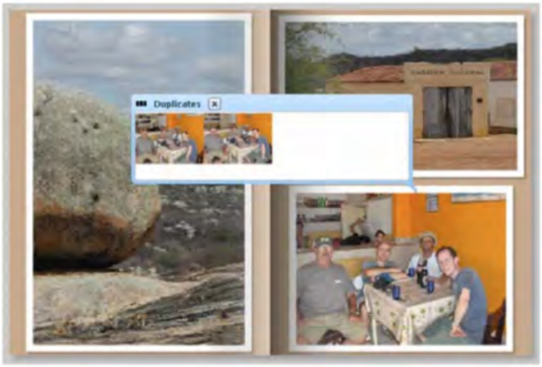


Figure 16.3: User Interface of the interactive *SoftBacks* [C14] application, showing the near-duplicate functionality, where once a photo is clicked upon, the associated near-duplicates are queried from the back-end through the web services API, and are presented to the user for a possible swap.



The user can use that interface to browse and inspect the final photo-book. In this way, the soft-proof book becomes the user interface. We have implemented an intuitive interaction mode in which the users can click on an image and request alternatives from the back-end repository database. In the current implementation, the user is presented with all the images that belong to the same similarity –or near duplicate– cluster as the selected image, allowing for an image swap, as depicted in Fig. 16.3.

CHAPTER 17

Back-end description

SoftBacks has been implemented using a multi-tier web architecture (presentation tier, application tier, and data tier). In addition, we added one special parallel tier that does not commonly appear in other web architectures, an execution tier; this was needed because image analysis algorithms, as well as image clustering algorithms, have characteristics that make it problematic to run in the web application tier (*i.e.*, resulting in unresponsive user interfaces due to processing needs).

The central component to our system is the relational database (MySQL) in the data tier, which hosts a core schema with tables to hold media files, metadata extracted through image analysis, tags, and results of clustering, searching, and ranking. The architecture has a back-end plug-in mechanism for integrating algorithms into the execution tier. There is also a service-oriented XML-web services API for extending the system and integrating web-client functionality, like the Adobe FLEX user interface in Fig. 16.1, 16.2 and 16.3.

17.1 Plug-in's implemented in our system

We now list all the plug-in's that were implemented within our system's back-end, in order to accomplish the selection algorithm presented in Chapter 13.

17.1.1 Analysis plug-in's

These are the plug-in's that implement the metadata extraction at image ingestion time into our system described in Chapter 13. Some of them are sophisticated analysis algorithms, like the image aesthetic appeal measure, and others are more straightforward header information extraction, like the EXIF [6], header extraction.

1. **EXIF header**: capture time information is used for time clustering –Section 3.1– and similarity measures –Section 3.2. Size and aspect ratio of the image are used for layout purposes.

2. **Image aesthetic appeal**: as calculated by Eq. 5.23, in –Section 5.5. This is used to rank images based on their aesthetic appeal.

3. **Face detection**: as described in [118, 62], using the implementation included in the OpenCV package [19]. This allows the system to know the region of the image where to extract the face aesthetics low level features, as described in Chapter 6.

4. **Face feature extraction**: as described in [129]. These face features are later used in the face clustering plug-in below.

5. **Region based color histogram**: as described in Section 3.2.2, which is used in order to cluster images based on a region color similarity measure as described in [C73].

6. **Smile detection**: as described in [124].

17.1.2 Clustering plug-in's

1. **Time clustering**: the algorithm described in [51] is used to cluster the images in a multi-resolution way, as described in Section 3.1.1.

2. **Face clustering**: the algorithm described in [129] uses the face features extracted above, in order to perform face clustering.

3. **Similarity clustering**: the algorithm described in Section 3.2.2, is used with an adaptive threshold at each hierarchy level, as described in Chapter 13.

17.1.3 Ranking plug-in's

1. **Aesthetic appeal image ranking**: the images in a subset of the collection are ranked based on their image aesthetic appeal value, as calculated by Eq. 5.23, in Section 5.5.

CHAPTER 18

System output examples

As described above, we implemented a system that performs photo-book auto-population from a photo collection, along with an intuitive user interface to fine tune the selection results. In this chapter we present a few output results of the whole system, starting from a collection of over 800 photographs, and resulting in a photo book.

Figure 18.1: Template based layout, where the preferred photo is chosen to be the most relevant photo as described in Chapter 13, the other three images are the most relevant second and third and, finally, the fourth photo is the smallest one.

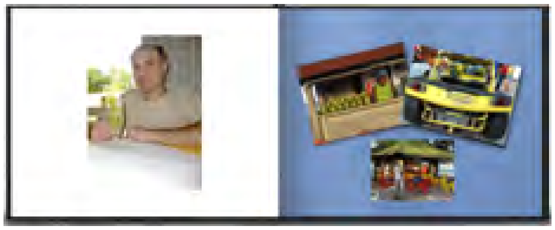


Fig. 18.1 presents a template based layout. These layouts are quite commonly used by photo book software and services, since they are designed by professional designers, along with clip art and nicely harmonized colors. The user only needs to select the images to be placed into each of the empty holes in the template. In this case the algorithm in Chapter 13 was run and the four most relevant images were assigned to the four empty holes, assigning the most relevant image to the largest hole, and the least relevant image of the four to the smallest hole.

Fig. 18.2 presents an automatic layout, where a set of images is laid out on a specified aspect ratio sheet of paper, optimizing coverage [13]. In this case the top ten most relevant photos are assigned to the layout engine. In this case we experimented with the possibility to blow up the images with smiling characters to the full size of the page, as seen in Fig. 18.3.

Finally, Fig. 18.4 presents 6 actual photo books used in our user study in Chapter 13, for which the cover photograph was selected automatically as the most aesthetically appealing landscape oriented photograph –*i.e.*, *width* > *height*– of the whole collection.

Figure 18.2: Automatic layout, as described in [13], with the top 10 relevant photos, in no particular order.

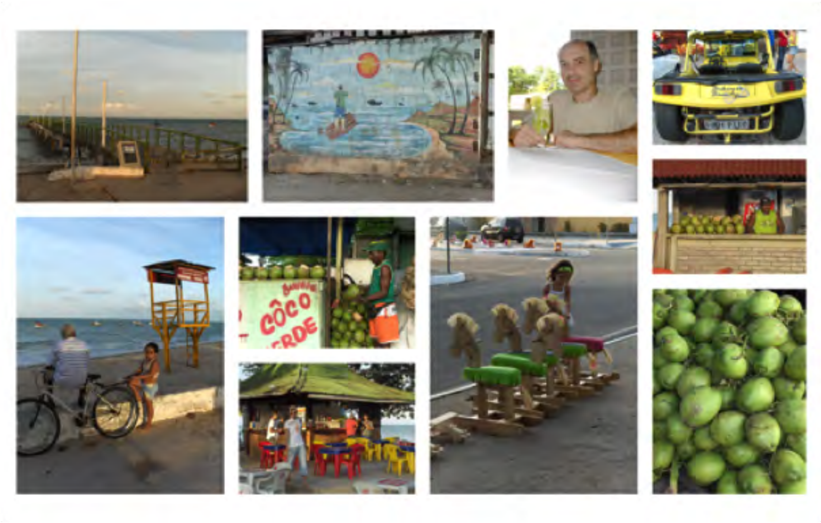


Figure 18.3: From the images that have been selected based on their relevance –see Chapter 13, and using the layout in [13], we give preference to the images with detected smiles in them, and blow them up to the full size of the page.

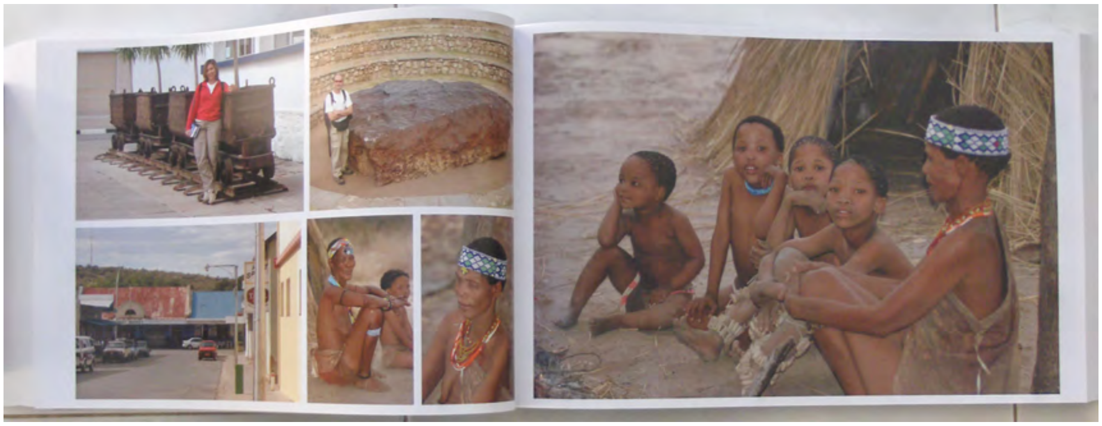
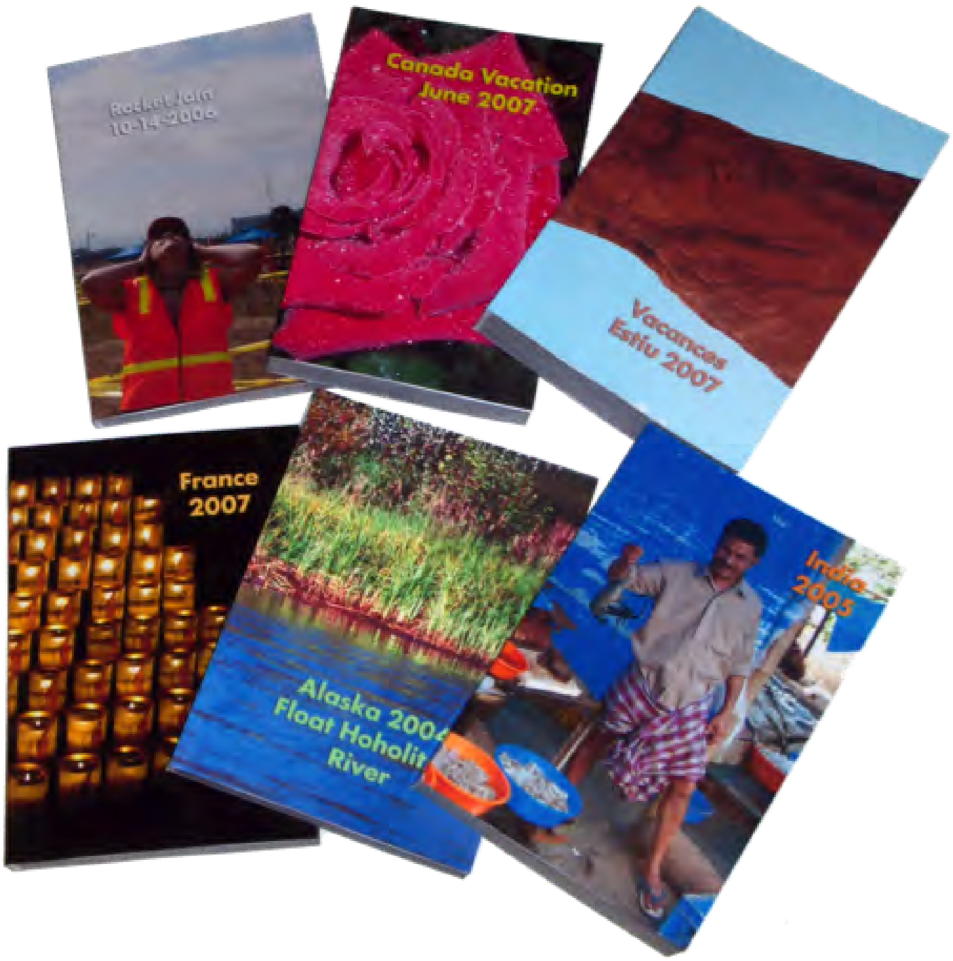


Figure 18.4: Top aesthetic images in 6 collections used in the user study in Chapter 13, that were used as the actual photo-book cover.



CHAPTER 19

Conclusions and future work

In this dissertation we have covered three main parts, namely, Part I describing media aesthetic models, Part II describing photo storytelling algorithms, and Part III describing a practical implementation of a photo storytelling system. We would like to remind the reader that intellectual property rights associated with this dissertation are owned in part by *Hewlett-Packard Development Company, L.P.*, as well as in part by *Telefonica I+D Company*; see Appendix D for patent details.

In the Introduction, we set ourselves four main goals. We consider that we have covered each of those points in Parts I, II and III of this dissertation, and we hereby briefly describe the main conclusions for each of them:

1. Build media aesthetics computational models that correlate well with human perception in order to help users manage their photos.

   We described in detail the different models that we came up with for regression based image aesthetics [C9,C11,C13], which turned out to be the most useful method for photo storytelling, thanks to the ability to rank photos based on their aesthetic appeal. Regression models for face aesthetics [C11] turned out to be extremely helpful at the time to choose the photos of the main characters to include in the final story summary. Other models were also investigated, like classification based model for both images and videos. Finally, one of the regression based models for image aesthetics was used in a user study for image search re-ranking [C9].

2. Create new image collection structures that allow to target specific summarization counts.

   We developed a hierarchical scalable representation –HSR– introduced in [C10], that, by jointly clustering in time and similarity, allow the collection of images to be ordered in a relevance order, allowing one single scalable representation to be used in order to generate summaries of any length.

3. Create new photo selection algorithms that sample the original photo collection in a meaningful way from a multimedia storytelling point of view.

Test these algorithms in user studies in which the users provide their photo collections, and judge how well the proposed algorithms perform.

We investigated two different avenues to automatic photo storytelling. In the first one, only the images in the photo collection to summarize were analyzed in order to come up with the final summary [C10]. Meanwhile, in the second approach [C2,C3], we also analyze the images in the user's online social network –OSN–, and use that as social context to identify who in the story to summarize should have more presence in the final summary. In this second approach we also extract the number of images in the online social network with faces in them, and use that as a user's trait, and mirror that ratio in the final summary. Both approaches were tested in user studies, in which end users were judging the collection summarization accomplished by the proposed algorithms.

4. Build an automatic image selection system that helps users reduce the overall workload to generate a story to share with friends and family.

   We implemented an automatic photo storytelling system[1] [C14], which was used to test the algorithm presented in Chapter 13.

We therefore, consider we have accomplished the goals we set forth in the introduction. We will now briefly introduce new research directions for future work.

19.1 Future work

We see both the areas of media aesthetics, as well as the area of multimedia storytelling, as having a great future. We list here a number of research directions aimed at improving both research areas.

In the area of face aesthetics much can be done in order to improve the regression based model we presented. For instance new features that would be relevant are face pose, gaze direction, recognizing other face expressions, image composition taking into account face position, etc.

In the area of aesthetics based image re-ranking new experiments including a large scale user study with the participants personal image libraries, where their friends and relatives will be asked to search for images in their collection, is a possible avenue to further analyze the effect of aesthetics in personal media re-ranking.

[1] This system is the property of the Hewlett-Packard company.

It was clear in our user studies that category dependent image aesthetic models can have a great improvement over general ones, rendering this area as a very fruitful one for future aesthetics research[2].

With regards to video aesthetics future work includes exploring temporal models to characterize video aesthetics, investigating personalization techniques and shedding light on which features of our aesthetics model may be universal *vs.* person-dependent, and assessing the influence of audio in aesthetic ratings so as to form a complete measure of audio-visual aesthetics. A user study analyzing the impact of an aesthetics based video re-ranking scheme would be of great value, also.

In the area of photo storytelling, new users' traits can be extracted by analyzing other features from their on-line social network, either textual data –*i.e.*, user comments– as well as image data. These traits can, in turn, further personalize, and therefore improve, the final photo summaries to be shared on-line. We feel that this is going to be a great research area in the years to come.

And, finally, collaborative photo storytelling is quite a new area of research, but we feel it will also be of interest in the near future, since it will tackle the very nature of photo storytelling, which is rarely an individual task, rather, it is usually a collaborative effort. Therefore, new algorithms that can combine multiple photo collections belonging to different users, and allowing the final summary to adapt to each of them in a specific way, will be of great interest.

[2] The author has recently submitted for publication new research in this specific area of category dependent image aesthetics.

Glossary

act is a major section of a play (dramaturgy), in which all story elements are related to one another; in our application this can be seen as a relatively large group of photos representing a well defined period in time. 124, 142

aesthetics is the field that deals with the human appreciation of beauty, in which they study the psychological responses to beauty and artistic experiences. Therefore, a media aesthetics model tries to automatically predict a media object's aesthetic value, *i.e.*, how beautiful would it be perceived by humans. 3

character is a person portrayed in an artistic piece, such as a drama or novel . 57, 125, 133, 135, 136, 166

event can be defined as a significant occurrence or happening, or as a social gathering or activity. In both cases, if the user is capturing memories in the form of either photos or videos, there will be a collection of media that will represent that event. 3, 19, 114, 124

human centric is the reasoning behind building specific functionality into computers with a focus on the long-term effects those systems will have on humans. xxiii, 2, 30, 31

media object is an object of any of the possible multimedia types, like image, video or audio. In this dissertation we focus on images and videos. 3, 11, 12, 25, 26, 76

perfect bound is the adhesive binding of a book or magazine that allows it to open flat (180 degrees). Used mainly for paperbacks, its durability lies often in the quality and amount of adhesive used. 124, 138

photo storytelling is the activity of telling stories using photos as part of the resources used by the storyteller to accomplish his/her goal. In the more traditional setting, photo storytelling was performed using a photo album around which a group of people would listen to one or more storytellers explain past experiences or events [32]. 3

saddle stitched is a method of securing loose printed pages with staples down the middle of a folded sheaf of papers. Many booklets are saddled-stitched. 124

scene is a subsection of an *act* (dramaturgy), during which the set is unchanged. 143

shot is a single video sequence captured by one single camera without interruption (cinematography). 143

treatment is a process or intervention administered to experimental units in a controlled experiment. For example, a doctor treats a patient with a skin condition with two different creams to see which is most effective (treatment A vs. treatment B). 98–100, 103, 105–107, 109

WIPO is the World Intellectual Property Organization (WIPO), which is a specialized agency of the United Nations. It is dedicated to developing a balanced and accessible international intellectual property (IP) system, which rewards creativity, stimulates innovation and contributes to economic development while safeguarding the public interest . 210

Acronyms

HCI Human Computer Interaction. 118, 119

HSR Hierarchical Scalable Representation. 122, 136, 174

HVS Human Visual System. 35, 37

OSN On-line Social Network. 115, 121, 143, 144, 146, 147, 151–157, 161, 175

REST Representational State Transfer. 164

APPENDIX A

Highlight detection

Digital cameras have a limited dynamic range to sense the brightness of incoming light. If the incoming light is too bright at some pixel position, thus exceeding the camera dynamic range, it will not be correctly recorded, resulting in what is known as color clipping [91]. These produce non-linearities, that generate high energy output from our filter-bank implementation (see Fig. 5.6). This is a real-life problem in consumer photography.

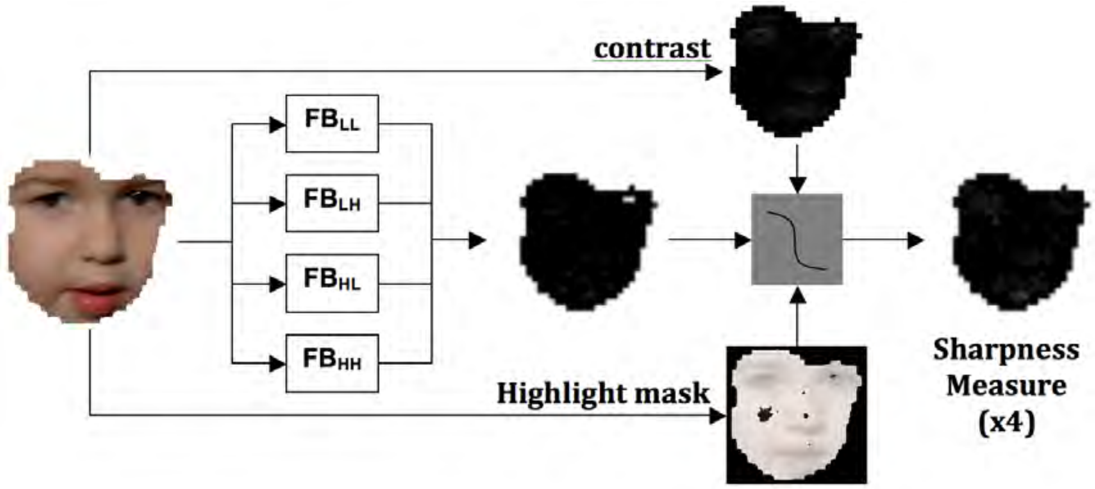
Humans are capable of perceiving the correct colors of objects in an image, even in the presence of shading and highlights. This property, known as color constancy, indicates the capacity of isolating illumination effects when perceiving colors from objects [70]. Conversely, humans are also capable to filter out any highlights, in order to judge the real sharpness of an image. Any accurate algorithm for sharpness measurement will need to incorporate a way to ignore the hard-clipped highlight areas in order to calculate a sharpness metric that will correlate better with the HVS.

In this work, two main sources of clipping were found specially problematic when calculating the sharpness metric, namely,

1. In out-of-focus areas, each point of light becomes a disc (*i.e.*, circle of confusion). The sharpness measurement algorithm may detect that circle as a sharp object.
2. Highlight specularity is a bright and highly saturated region in an image, produced by mirror-like reflections from glossy surfaces. If these reflections are of light sources (or high luminance objects), which may be either in focus or out of focus (see above), this may fool the sharpness measurement algorithm. Motion blur may worsen the situation, since a highlight from a point light source will generate a sharp high contrast line that follows the trajectory (see Fig. A.1). Specular highlights usually occur on glossy surfaces, and are specially problematic on human faces (human eyes, moist human skin, eye-glasses, etc.).

We trained a naïve, but very fast, highlight detector on a set of 200 images with highlights that were giving problems. On each specific region the 0.5% pixels with highest luminance are marked as highlight, which generates a highlight mask. This mask is morphologically dilated by a structure element of size 3 × 3 so as to add the borders of the highlight to the mask, since it is these borders that are generating those high frequency aliases that should be avoided. The sharpness measures falling under the highlight mask will be ignored in the final sharpness measure map (see Fig. A.1). In a region with few or no clipped pixels, the algorithm will erroneously detect highlights caused by diffused reflection. These diffused reflections occur always in flat (*i.e.*, low frequency) areas of the image, not impairing the sharpness calculation. See Fig. A.1 in which certain diffused reflections on the cheek and nose have been included in the highlight mask.

Figure A.1: This is a motion blurred face, with a highlight generated by a point light source (flash strobe), producing a sharp line highlight in the subject's left eye. A highlight detection stage has been added, so that pixels in the highlight areas are ignored from the sharpness calculation. Note the final sharpness measure has been amplified by 4 to show detail, since the face is quite blurry to start with.



An alternative to this method, is to modify the contrast correction function introduced above in section 5.2.1, and attenuate even more the high contrast pixels. We trained our system and identified the best exponent factor to deal with the clipping high contrast is -0.0336 in δ' (see Eq. A.1), as opposed to the original factor of -0.024 in the original formulation.

$$\delta'(contrast(j,k)) = \begin{cases} -0.0042 \cdot contrast(j,k) + 1 & \text{if } 0 \leq contrast(j,k) \leq 50 \\ 0.8 \cdot e^{-0.0336 \cdot (contrast(j,k)-50)} & \text{if } 50 < contrast(j,k) \leq 200 \end{cases} \tag{A.1}$$

In this case, the sharpness measure is given by $S'(i,j)$, in Eq. A.2.

$$\begin{aligned} S'(j,k) = \delta'(contrast(j,k)) \cdot \\ FB_{LL}(j,k) \ \cdot \ FB_{LH}(j,k) \cdot \\ FB_{HL}(j,k) \ \cdot \ FB_{HH}(j,k) \end{aligned} \quad \text{(A.2)}$$

In a way, this alternative method is equivalent to not taking the maximum value of sharpness in the regionand instead taking a lower quantile, while setting $\delta(contrast(j,k)) = 1$. For instance in our video aesthetics work –see Chapter 8, the sharpness feature that was the most discriminative for video aesthetic classification was the 3^{rd} quartile of the sharpness measure.

APPENDIX B

Sharpness density calculation

Sharpness density (SD) is defined as the percentage of an image region that has energy content in mid and high frequencies –*i.e.*, not only low frequencies. This is a very convenient measure in order to identify which regions have a reliable sharpness measure, since regions with low SD may actually be composed of multiple regions that have been merged in the segmentation process, some of them flat; or the whole region is too flat to have any reliable sharpness reading. Section 5.3.1 already showed how the sharpness density is useful to fuse the sharpness, contrast and colorfulness for each of the image regions.

This has been implemented as a variation of the matting algorithm presented in [75] where the matting result is intersected with the image regions to calculate the SD for each region. In this implementation the output of the filter bank in Fig. 5.2 is modified additively by the lowest band pass filter (FB_{LL}) in order to give relevance to all regions that may depict an object, or part of it, but may be very out of focus. The very high contrast edges are also toned down by the δ function of Eq. 5.2. This yields a modified sharpness metric MS in Eq. B.1.

$$MS(j,k) = \delta(contrast(j,k)) \cdot \Big(FB_{HH}(j,k) \cdot FB_{LH}(j,k) \cdot FB_{HL}(j,k) + 4\Big) \cdot FB_{LL}(j,k) \tag{B.1}$$

This modified sharpness is then normalized (NMS) between 0 and 255, in order to be able to detect any object in the scene no matter how out of focus it may be[1]. A bilateral filter [112] is then applied to the NMS (see Eq. B.2). The output is a weighted average of the input by means of an edge-stopping function. The weight of a pixel depends on this edge-stopping function in the NMS measure domain, which decreases the weight of pixels with large NMS differences.

$$FNMS(j,k) = \frac{1}{K} \sum_{x,y \,\epsilon\, \eta(j,k)} exp\left(-\frac{(x-j)^2 + (y-k)^2}{2{\sigma_s}^2}\right) \cdot exp\left(\frac{(NMS(x,y) - NMS(j,k))^2}{2{\sigma_i}^2}\right) \cdot NMS(x,y) \tag{B.2}$$

[1] The aesthetic appeal of an out of focus object is usually higher that that of a completely flat image.

where K is a normalization factor, and $\eta(x, y)$ denotes the window of smoothing area that is centered at pixel (x, y). In the current implementation $\eta(x, y)$ is a $5x5$ pixel square, $\sigma_i = 10$ and $\sigma_s = 1$. A series of mathematical morphology filters are applied to FNMS in order to homogenize the areas with objects in the scene, so that they do not overflow into adjacent objects, see Eq. B.3.

$$MFNMS = erosion_{se3}(opening_{se2}(closing_{se1}(FNMS))) \tag{B.3}$$

Where the structuring elements are the circles of diameter: $se1 = 7$, $se2 = 3$ and $se3 = 5$ pixels.

Pixels with $MFNMS$ equal to 255 are certain to belong to an object with a certain edge/texture content. The sharpness density is calculated on a region by region basis as the fraction of pixels belonging to objects, see Eq. B.4.

$$SD_i = \frac{\#PixelsInRegion(i)|_{MFNMS(j,k)=255}}{\#PixelsInRegion(i)} \tag{B.4}$$

APPENDIX C

Attached related publications by the author

One last consideration that needs to be taken into account is the final media presentation to be experienced by the end user. The document in which the images will be embedded may introduce a set of aesthetic conditions which may only be fulfilled by a small set of images from the initial collection.

The theory of aesthetics can also be applied to the final output document to be experienced by the end user, for instance taking into account the layout of pictures along with other graphical object on web pages or printed documents [53], or taking into account the color and geometry temporal evolution in slideshows or videos [18].

In this appendix three publications present an analysis of the former case, *i.e.*, for printed documents, in four specific situations:

1. aesthetic appeal of the image with the document background color. This specific approach was used in the Hewlett-Packard Photosmart Digital Camera series;

2. aesthetic appeal of the image with document taking into account color harmony with the rest of the document;

3. aesthetic appeal of the image with document, taking into account the visual balance with the rest of the document; and finally;

4. aesthetic appeal of the image with document, taking into account color harmony and visual balance with the rest of the document. A look is given at how to use these measures in a search and retrieval situation, *i.e.*, having a specific document to which the user wishes to attach an image, query the image database in order to find the most aesthetic matches.

These algorithms have been protected by patents, see [C39,C68,C69,C70].

C.1 Automatic color scheme picker for document templates based on image analysis and dual problem

Referenced as [C17].

Automatic color scheme picker for document templates based on image analysis and dual problem

Pere Obrador

Hewlett-Packard Laboratories
Palo Alto, CA, USA
pere.obrador@hp.com

ABSTRACT

This paper presents two complementary methods to help in the area of document creation where the document includes color templates (banners, clipart, logos, etc.) as well as photographs. The problems that are being addressed are:

- given a photograph that a document needs to be built around, extract a good palette of colors that harmonize with the selected photograph, which may be used to generate the color template;
The images are segmented with a color based morphological approach, which identifies regions with a dominant color. Based on the morphology of such "color" regions, and the other color objects in the template the scheme will pick a set of possible color harmonies (affine, complementary, split complementary, triadic) for such color elements within the document based on the combined morphology image-document. If the image is changed in the future the color scheme could be changed automatically.

- given a document color template, identify from a collection of images the best set that will harmonize with it.
The document color template is analyzed in the same way as above, and the results are used to query an image database in order to pick a set of images that will harmonize the best with such a color scheme.

1. INTRODUCTION

Color harmony sets guidelines on how to create effective color combinations. Many attempts have been made, through many historical periods, to create recipes for color harmony. It is, however, not possible to make a list of rules to describe the harmonious or disharmonious visual image. Complementary contrast, whatever the subject, is not a requirement for a harmonious color image. "Ton-sur-ton" or analogous color scheme (where all colors are related to one color hue in slightly different shades or tints) color use doesn't guarantee harmony either. Only the human eye can judge the final artistic result[1].

The color schemes used the most in harmonization are[1]:
Analogous scheme: uses any three consecutive hues or any of their tints and shades on the color wheel
Complementary scheme: uses direct opposites on the color wheel
Clash scheme: combines a color with the hue to the right or left of its complement on the color wheel
Monochromatic scheme: uses one hue in combination with any or all of its tints and shades
Split complementary scheme: consists of a hue and the two hues on either side of its complement

The algorithms presented in this paper try to help in the generation of color documents that include banners, logos and photographs. The images and documents are analyzed in order to suggest possible color palettes or images that will harmonize well with the existing document/photograph.

In order to develop an algorithm that will help in color harmonization it is of critical importance to identify areas with a homogenous dominant color in which there is little color activity, and also of high importance is to identify smaller regions with high color activity (i.e., even if the color region is very small within that high color activity region, it may still be of high importance at harmonization time if the chroma of such region is significantly different from the rest of regions' chroma). Such an algorithm was developed in [3], and it will be briefly described in the following section below.

Since this paper is 100% color-related, I would encourage you to check the color version at the HP-Labs web-site (http://www.hpl.hp.com/techreports/).

2. ALGORITHM DESCRIPTION

This paper describes the color patch extraction algorithm[3], as well as a few algorithms for color harmonization: a) starting from a photograph suggest a color palette to go with it; b) starting with a document, retrieve the best images from a database that will harmonize with it.

2.1. COLOR PATCH EXTRACTION

The image, either the document or the photograph, will be converted to an abstract representation that will ease the task of harmonization as shown in figures 4 and 6. This representation includes a reduced number of color patches, their centroid location on the page/photograph, the size of each patch (in number of pixels), and the connection of each patch with any of the four borders of the image.

The image is first quantized to a set of color bins (which may be generated from the image itself for best results). The quantized image is then manipulated in order to obtain a reduced number of regions that can be used for harmonization purposes. The fewer patches and the larger they are the best for harmonization purposes, since the human visual system is most sensitive to large areas of color. These large regions are the ones that will be taken into close consideration in the algorithms below in the next two sections.

No perfect scene object segmentation[4,5], is intended in the color patch extraction process, since different objects may be quantized to the same color, the result may be a merged color patch.

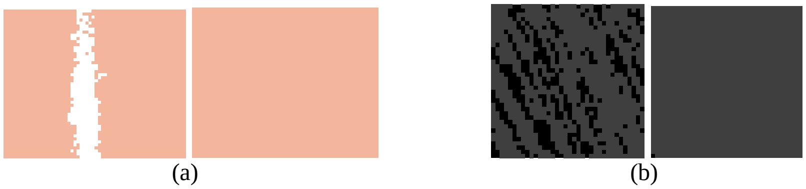
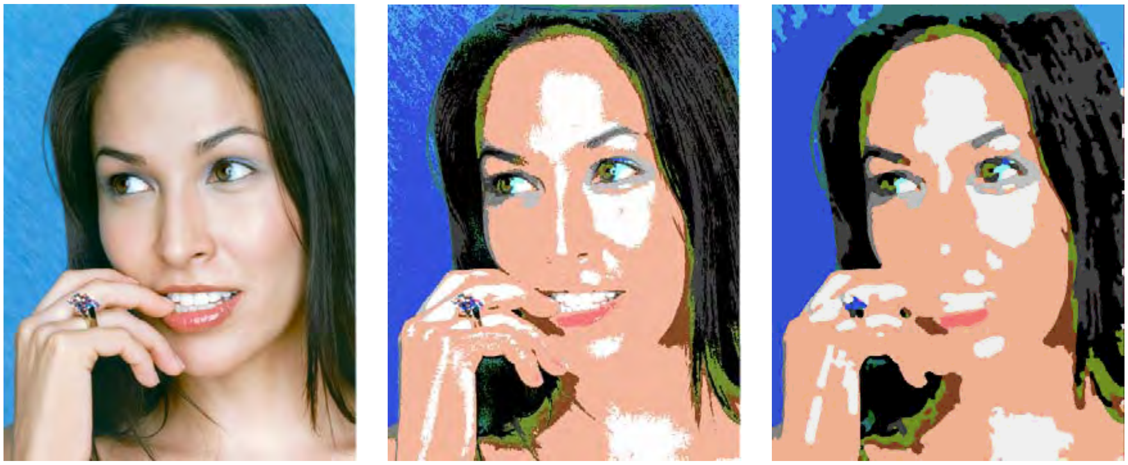


Figure 1. Examples of color patches (from the sample image "girl" in figure 2) and their underlying color patches, a) at a low resolutiolarger scale, b) at a lower scale.

Another aspect that should be noted is the multi-resolution nature of such a problem. Underlying color patches at different scales may look very different indeed. i.e., what may look as an underlying color patch at a very small scale, might just look as a non-underlying color at a very large scale (see figure 1a, where the vertical white stripe, reflection on nose in figure 2b, is removed at the larger scale in favor of the underlying color; on the other hand, at a smaller scale it is definitely an underlying color within its region as seen in figure 2c). And the other way around, large enough regions (given the scale/resolution) with high color activity will not have a clear underlying color patch and should be left alone.

Having these requirements in mind, a technique was developed[3], in order to extract the underlying color patches in an image, quantized with a predetermined quantization table (palette): a parallel symmetrical a*lternating sequential filter* scheme which allows for color patch extraction, while maintaining edges and detail regions, and also, a maximum likelihood scheme to fix edge jitter[6,7], in color morphological filters applied to sparsely quantized images. This algorithm is multi-resolution by nature, and it can be devised having in mind the scale of the color patches that need be preserved. This filter is implemented in stages (*alternating sequential*), starting with the smaller scale, and ending with the larger scale. 4 stages give good results for the color harmonization application.



(a) (b) (c)

Figure 2. a) original “girl” image, b) quantized to 25 color bins, c) 4-stages color patch extraction

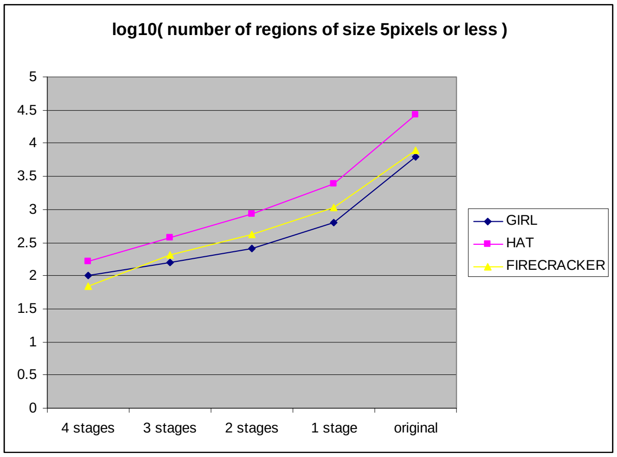
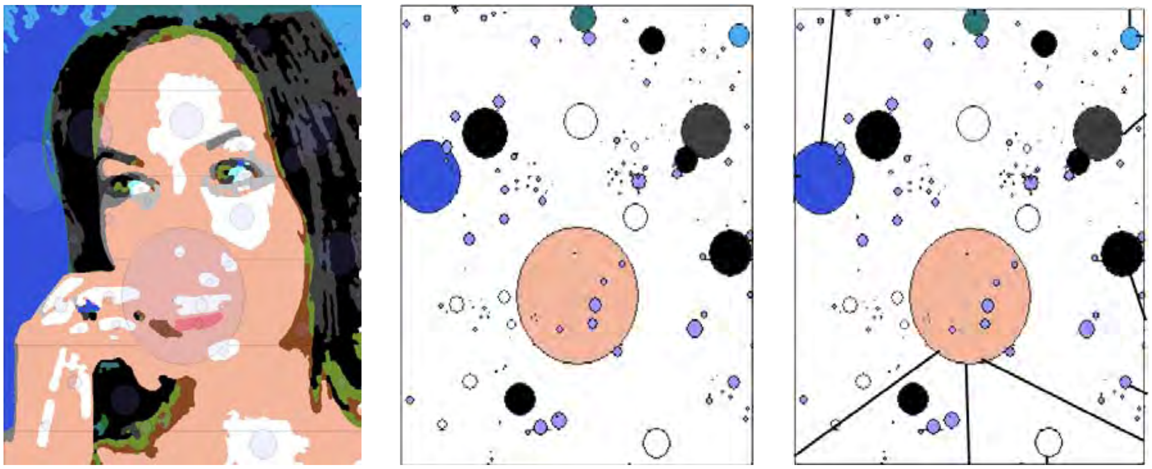


Figure 3. overall number of color patches of size 5pixels or less after processing on the image in figure 2, and 2 other images^3



(a) (b) (c)

Figure 4. a) “girl” image with superimposed color patch abstraction, b) “girl” color patch abstraction, c) larger patches touching borders abstraction.

Figure 2 shows an original image, the quantization, and the color patch extraction result. In the color patch image it can be observed that the detail is well preserved in areas of high color activity (i.e., eyes, ring), and at the same, the overall number of patches (regions) has been reduced considerably (see figure 3 to see how the number of color patches is reduced with the number of processing stages).

Once the color patches are extracted, the abstract representation of figure 4 can be built. The centroids, as well as the average color (in Lab color space) for each patch, and the morphology of the image with respect to the borders (number of pixels of each patch touching which border) are stored in such representation. The importance of such representation will become apparent in the sections below.

2.2. COLOR SCHEME PICKER BASED ON PHOTOGRAPH

This algorithm allows to automatically generate a palette of colors from a photograph, which the graphic artist or designer can use directly knowing that it will harmonize with said photograph.

By extracting the Lab color averages of each of the color patches the palette for logos/banners/text can be easily generated. It can also be done by a weighted average of the color in all the regions from the same color bin, or a combination of the two. This is quite straight forward from the abstract representation presented in section 2.1 (see figure 7a.

One important thing to keep in mind is text readability, since not all color combinations are easily read. Once the palette is generated, the text possible colors are reduced as presented in [8], based on the background color that is selected.

If the graphic artist would like to place the photograph on a full-bleed background for instance, this means that all the photograph will be immersed in that background color[1,2] (i.e., all borders of the photograph will be touching that color), with strong implications to the color harmony of the result. In such situations, the color palette for such background is also reduced by the following algorithms (in fact after obtaining the image abstraction as in figure 4, any of the color schemes presented in the introduction could be used to select a background color):

Analogous background #1 (largest):

- pick the non-gray color bin in which all regions are barely touching the borders (*sizeFactor* >=minSizeFactor)
- pick the largest (in number of pixels) such color bin
- the background hue will be the hue of the average of such color bin

Analogous background #2 (unusual):

- pick the non-gray color bin in which all regions are barely touching the borders (*sizeFactor* >= minSizeFactor)
- pick the colorimetrically most different (**Lab** space, **ab** distance) such color bin
- the background hue will be the hue of the average of such color bin

Complementary background:

- pick the non-gray color bin in which all regions are mostly touching the borders (*sizeFactor*< minSizeFactor)
- pick the largest (in number of pixels) such color bin
- the background hue will be the complementary hue of the average of such color bin

The algorithm is not considering gray-scale color bins, since these colors harmonize well with any color, and for this reason are not considered in this study.

Where *sizeFactor* is calculated for each color bin as:

$$\text{sizeFactor}_i = \frac{\sqrt{\sum_{j\ patch \in patch_i} totalPixels_j}}{1 + \sum_{j\ patch \in patch_i} borderPixels_j} \qquad \textbf{equation } 1.$$

And gives a good estimation of the relation between the size of a patch and how much it is in contact with the border. The more pixels touching the border, the smaller *sizeFactor* becomes, and will not be chosen in the analogous algorithms above, which makes sense, because the background color should not merge with any part of the photograph so that it stands out. If the color is touching the border (and it is the largest such patch), a complementary background will bring out the photograph.

So all color bins with *sizeFactor* >=minSizeFactor (barely or not touching border) are considered for analogous schemes, and color bins with *sizeFactor*< minSizeFactor (mostly touching border), are considered for complementary color schemes. See figure 7b for an example of analogous #1 (largest), where the full bleed background has been generated automatically, after the average hue of the circled regions presented below in figure 5c.

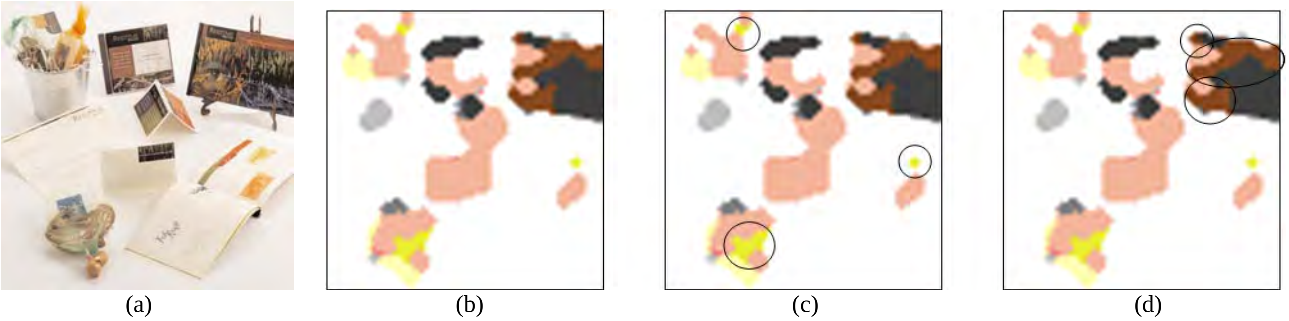


Figure 5. Image in figure I with a) original "stationary" image, b) color patches being extracted, c) regions belonging to the largest color bin with *sizeFactor*>=minSizeFactor, d) regions belonging to the largest color bin with *sizeFactor*>=minSizeFactor, for bottom and left borders only

On more level of complexity arises if the graphic artist is willing to add new color objects that are intersecting the photograph. For instance, it could be intersecting the photograph on the left and bottom borders (as shown in figure 7c). In this situation the algorithm can still be the same as the described above in this section, but modifying the formula for *sizeFactor,* where only the borders that will be intersected are considered as borders in equation 1. Figure 5d shows how the result for the analogous #1 (largest) has changed to another color bin (larger than the selected in figure 5c) which is touching the right border. One last consideration is that this new color object should harmonize with the background of the image (figure 7c shows that in this case the result is harmonious); in the case it would not harmonize, the algorithm could fall back to the following (smaller) analogous color bin, and so on.

2.3. PHOTOGRAPH RETRIEVAL BASED ON DOCUMENT QUERY

In this section the dual problem from the one presented in the section above is presented. In this case, given a document with space for a photograph, the system queries a database retrieving the best suited images for such document based on the morphology of document and photograph, and an analogous color scheme. Other color schemes are obviously possible, as the ones presented in the introduction, for instance.

The main idea is to generate abstract representations both for the document and the image to be tested for similarity. The morphology of each of them will be used to see how close in space the color patches are to each other. The color similarity and the size of the patches.

Figure 6 shows an example of all the process in order to obtain such abstract representation. The original image (Figure 6a), is quantized with the same color bins as the images in the data base, resulting in figure 6b. This is further processed by the 4-stage color patch extraction algorithm resulting in figure 6c, and the abstract representation of figure 6d. Up to here, it is exactly the same process as described in section 2.1.

One of the important things to keep in mind when processing color documents is that anti-aliasing borders usually map to intermediate color bins, resulting in color artifacts when trying to detect color similarity.
After heavy experimentation, for better results it is important to keep the document abstraction to a few color bins only. In the presented algorithm the largest color bins covering at least a certain percentage (*areaPercentage*) of the color (non-gray) areas are kept in the abstract representation (figure 6e).

Now, both document and images in the database have the abstract representation (document: figure 6e, sample image: Figure 4b). Such representations can be used in order to calculate how well the document and sample images fit in a color harmonious way.

The approach to calculating how well the document and image fit was initially investigated as a modified gravitational function, in which the *harmony measure* between two color patches would be directly proportional to the product of color patch areas, and inversely proportional to the square of their distance, both euclidean in document space, and also euclidean colorimetric in Lab space. In this specific rendition of the algorithm the closer the color patches are (from sample image to document) the higher the *harmony measure*, but this could be changed at will, and could actually be left to the graphic artist to determine the layout of the photograph with respect to the document.

After psycho-visual tests, the formula was modified to equation 2, where the influence of the color distance was greatly increased. The size of the areas influence has been reduced to the square root of the product of the areas, the euclidean distance in document space has also been reduced to the distance itself (non-sqared), and the euclidean distance in Lab color space was augmented to be the distance to the fourth power.

$$harmony\_measure_{ij} = bin\_factor * \frac{\sqrt{patchSize_i * patchSize_j}}{(1 + centroidDist_{ij}) * (1 + (colorDist_{ij})^4)}$$ **equation** 2.

with

$$centroidDist_{ij} = \sqrt{(centroidX_i - centroidX_j)^2 + (centroidY_i - centroidY_j)^2}$$

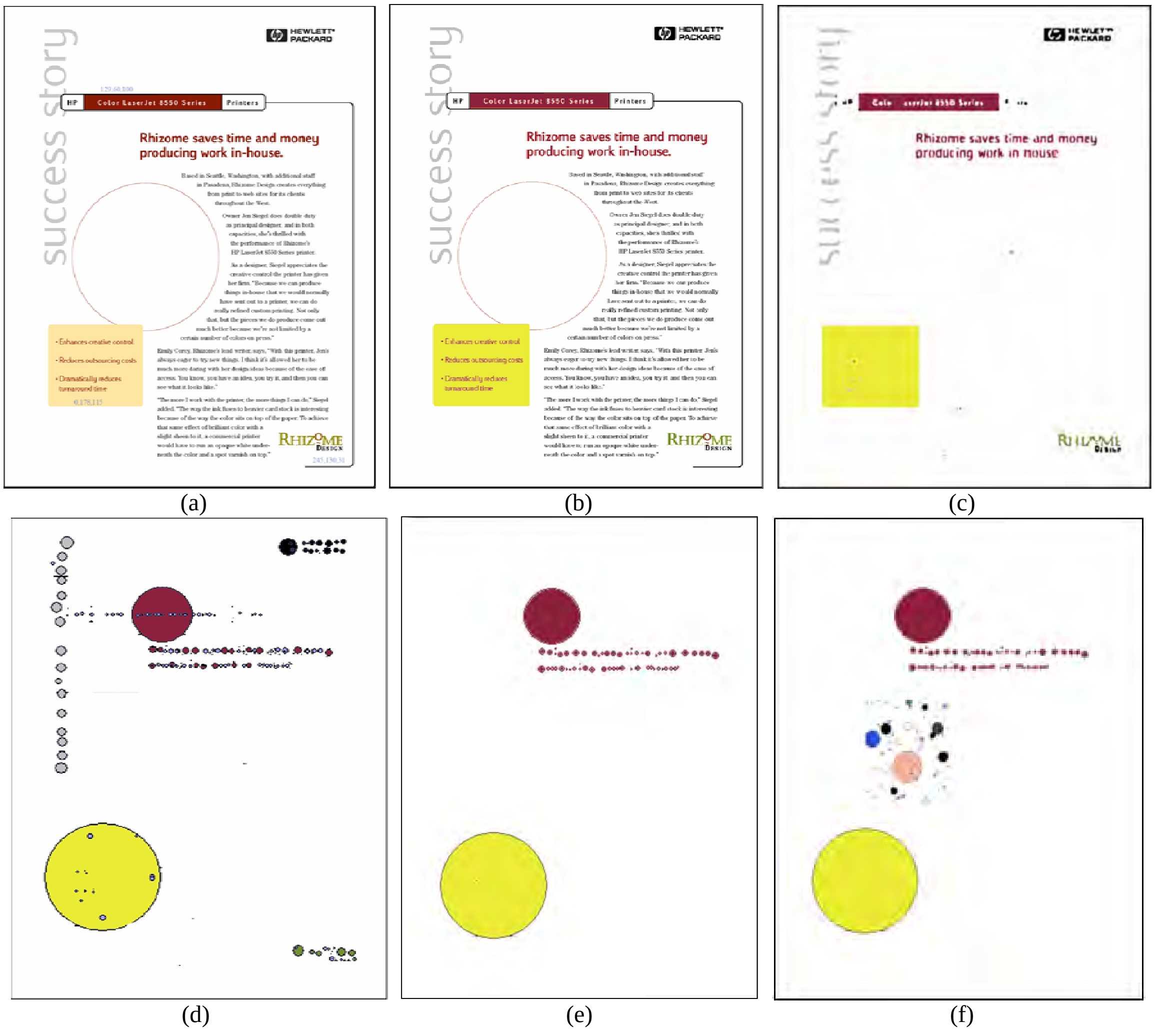
and

$$colorDist_{ij} = \sqrt{(averageL_i - averageL_j)^2 + (averageA_i - averageA_j)^2 + (averageB_i - averageB_j)^2}$$

The way the system is implemented is by placing the sample image abstraction in the reserved space for the photograph on the document (figure 6f). For each color patch in the document, the *harmony measure* between this patch and each of the color patches in the sample image are calculated as shown in equation 2, and then added together resulting in the *final_harmony_measure* (equation 3).

$$final\_harmony\_measure_m = \sum_{i \in document} \sum_{j \in image} harmony\_measure_{i,j}$$ **equation** 3.

Each image *m* in the database ends up with a *final_harmony_measure*. The larger this quantity, the better it will harmonize with the document, based on the rules stated above.


success story
HEWLETT PACKARD
HP Color LaserJet 8550 Series Printers
Rhizome saves time and money producing work in-house.
RHIZOME DESIGN

(a) (b) (c)

(d) (e) (f)

Figure 6. a) original document with a circular space for photograph, b) quantized to 25 color bin, used colors are the centroids of the color bin, c) color patch extraction, d) color patch abstraction, e) 90% of color area starting from larger color bin (avoiding color artifacts due to aliasing), f) color patch abstraction incorporating a sample image ("girl").

3. RESULTS

These algorithms have been tested with a color quantization scheme with 25 fixed bins (non image dependent). As mentioned above, it would be best to requantize based on each of the starting images for the color scheme picker base on a photograph, but it is unrealistic for the photograph retrieval based on document query, since all the images in the database would have to be re-processed for every change in the color scheme of the document.

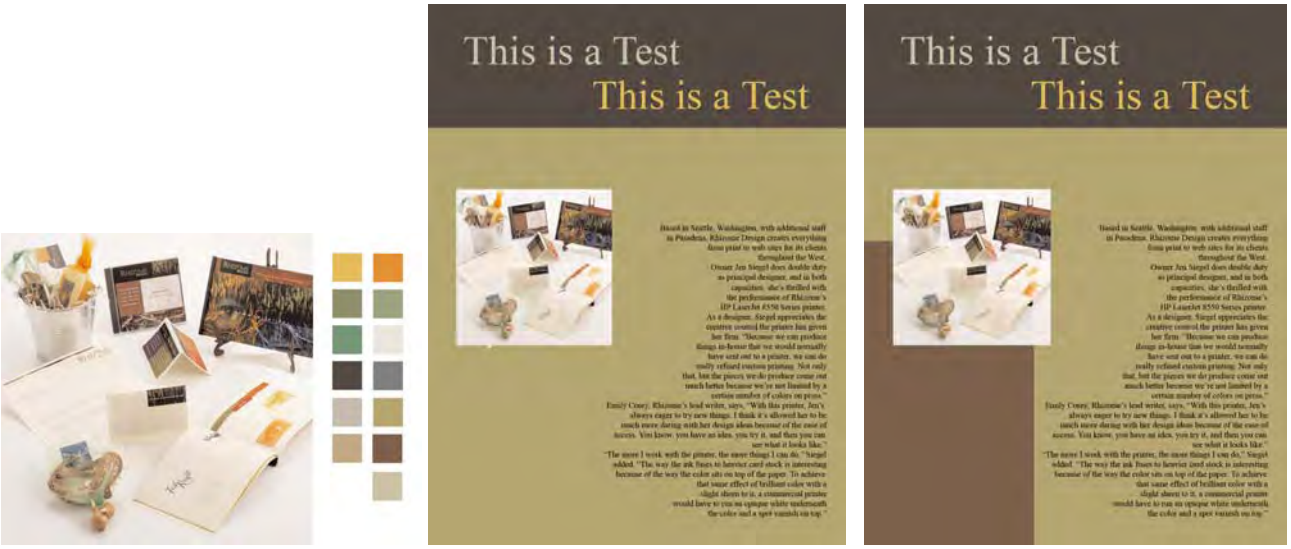
The color patch extraction is performed with a 4 stage filter[3].

In order to optimize for speed, with reduced impact in the final results, the photo abstraction size was kept very small (64*48 pixels), while the document abstraction is kept pretty detailed (1024*800) due to the fine details in such images.

The *sizeFactor* is set to 4, which is equivalent to an area of 16 pixels not touching the border in a 64x48 pixels abstraction image representation.

The *areaPercentage* is set to *90%.*

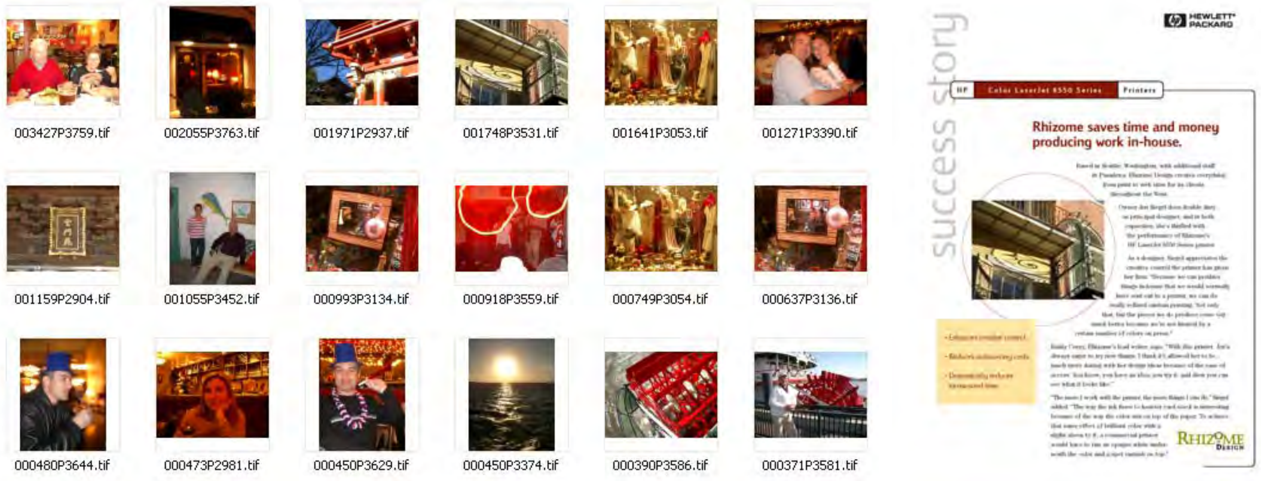
Figure 7 shows, starting from the photograph on the left, an automatically generated analogous color palette (a), along with a analogous (largest) full bleed background automatic selection (b), and finally an extra rectangle added for design style which overlaps the photograph on the left and bottom borders, for which the color is also calculated automatically with an analogous (largest) color scheme. See section 2.2 for details.



(a) (b) (c)

Figure 7. a) Palette automatically extracted from the image, b) automatic background full bleed color, analogous #1 (largest) color scheme, c) adding a rectangle overlapping with image (color selected automatically with analogous #1 (largest) color scheme using only the bottom and left borders to check *sizeFactor*).

Figures 8 and 9 show an automatic retrieval of images from a 900 photograph database (images taken with a digital camera over a period of a few months with no prior selection whatsoever). The retrieved collections are ordered from highest *final_harmony_measure* to lowest (file name first 6 characters).



a) b)

Figure 8. a) Results retrieved after querying the database with the document on the right (distance measure appears as first 6 digits in file name), b) example of the document with one relevant photograph from the retrieved list

Figure 8b shows a document template with red (top) and beige (bottom) banners. The results are psycho-visually very relevant from the color harmonization standpoint. The selected image in figure 8b (ranked 4th in the retrieval list) is very relevant, since it also has the red areas on the top-right, and the beige areas in the bottom-left. The graphic artist can safely select this image, since this is the harmonization rule embedded in the algorithm.

Figure 9b shows a document template with green (top) and violet (bottom) banners. This is a very unusual combination in the real world, but the algorithm still manages to retrieve very relevant results. The selected image in figure 9b (ranked 1st in the retrieval list) is quite relevant, since it also has the green areas on the top-right and middle of photograph, and the violet areas in the bottom. The graphic artist can safely select this image, since this is the harmonization rule embedded in the algorithm.

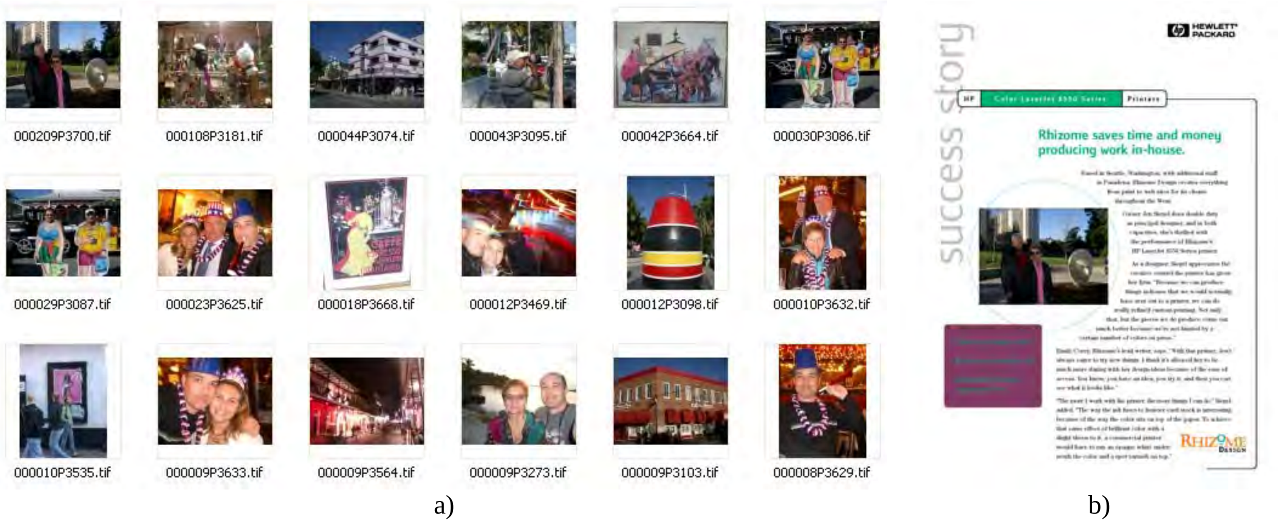


Figure 9. a) Results retrieved after querying the database with the document on the right (distance measure appears as first 6 digits in file name), b) example of the document with one relevant photograph from the retrieved list

4. CONCLUSIONS AND FUTURE WORK

A color patch extraction algorithm has been introduced, which allows for color harmonization algorithms. Two of these have been presented.

a) color scheme picker based on photograph, where given a photograph a color palette and possible background colors are presented automatically following 3 color schemes (analogous largest, analogous unusual and complementary)

The logical way to extend this work is to try more sophisticated color schemes (not only analogous and complementary), as the ones presented in the introduction.

b) photography retrieval based on document query, where given a document with space for a photograph, it queries a database retrieving the best suited images for such document based on the morphology of document and photograph, and an analogous color scheme.

This algorithm can also be extended by trying more sophisticated color schemes (not only analogous), as the ones presented in the introduction. Also experimenting with new morphology relationships between the document and photographs, and also extending the theory of overlapping color objects with photographs presented in section 2.2 to this algorithm, which would allow retrieving photographs placed on non-white document regions.

Check the color version at the HP-Labs web-site (http://www.hpl.hp.com/techreports/).

REFERENCES

1. B.M. Whelan, “Color Harmony 2”, Rockport Publishers, Massachusetts, 1997.
2. P. Bonnici and L. Proud, “Designing with photographs”, Rotovision SA, 1998.
3. P. Obrador, “Multiresolution Color Patch Extraction”, Electronic Imaging, VCIP 2006, San Jose.
4. Y. Deng; B.S. Manjunath; H. Shin; “Color image segmentation”, Computer Vision and Pattern Recognition, 1999.
5. J. J. Corso and G. D. Hager, “Coherent regions for concise and stable image description”, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference Vol 2, 20-26 June 2005 Page(s):184 – 190 IEEE Computer Society Conference on. Volume 2
6. E.R. Dougherty and R.A. Lotufo, “Hands-on morphological image processing”, SPIE press, 2003.
7. M. L. Comer and E. J. Delp, “Morphological operations for color image processing,” *J. Electron. Imag.*, vol. 8, pp. 279–289, July 1999.
8. “A color selection tool for the readability of textual information on Web pages”, Silvia Zuffi, Giordano Beretta, Carla Brambilla, HP Laboratories Palo Alto, Tech Report HPL-2005-216, November 28, 2005.

C.2 Content selection based on compositional image quality

Referenced as [C16].

Content selection based on compositional image quality

Pere Obrador

Hewlett-Packard Laboratories
Palo Alto, CA, USA
pere.obrador@hp.com

ABSTRACT

Digital publishing workflows usually have the need for composition and balance within the document, where certain photographs will have to be chosen according to the overall layout of the document it is going to be placed in. i.e., the composition within the photograph will have a relationship/balance with the rest of the document layout.

This paper presents a novel image retrieval method, in which the document where the image is to be inserted is used as query. The algorithm calculates a balance measure between the document and each of the images in the collection, retrieving the ones that have a higher balance score. The image visual weight map, used in the balance calculation, has been successfully approximated by a new image quality map that takes into consideration sharpness, contrast and chroma.

Keywords: No reference image quality, sharpness metric, document layout, document balance, photographic composition

1. INTRODUCTION

Creating a document from scratch is a hard task for non-experts, and even professional graphic artists will probably use some help when indexing large image databases from which to select an image to insert in the document. Effective ways to cluster and retrieve sets of relevant images to ease the document creation process is the main motivation behind this paper. This is a continuation of the work started in[1] where the proposed tool retrieves images from a database that will color harmonize with the document where the user wants to insert it/them.

Document layout balance has been researched for some time now[2,3]. This paper presents an attempt to help the user in indexing large image databases, by analyzing the document composition, and selecting appropriate images based on some simple balance and symmetry rules.

The presented approach is not trying to look at the photograph composition in itself; it rather tries to extract some fundamental image features that will help determine whether that photograph will work with the rest of the document.

A no-reference objective image quality map is presented. This takes into account sharpness, contrast and chroma features, each of them calculated on a region by region basis, eventually assigning a quality number to each of the regions in the image.

The presented image quality map is then used to calculate the visual weight map of such a photograph, and plug the result into a document balance measure.

This paper presents a method to help in the area of document creation where the user starts with a document, and needs to select a photograph from a large image collection, i.e., given a layout, automatically retrieve a photograph from an image database that will balance that selected layout, or, retrieve the top N photographs that would best balance that layout for the user to choose from.

2. PROPOSED METHOD

2.1. DOCUMENT BALANCE

Document balance is a very important aesthetic feature that graphic artists work towards in their creative process. It can either be achieved by symmetrical balance, which gives a feeling of permanence and stability, or by asymmetrical balance which creates interest[2]. Two main ways of defining balance are[3]: centered balance, where the center of visual weight is at the visual center of a page, and left-right balance, where the weight of an object on the left side of the page is matched by the weight of an object at the same vertical position on the right side of the page.

The visual weight of an object is defined as its area times its optical density, and this is also for photographs and graphics.

In this paper, a novel way to calculate a photograph's visual weight is presented, which can be used to better balance the document, in which image analysis and no-reference quality assessment are used to generate a quality map from which extract the visual weight.

2.2. IMAGE QUALITY MAP

As mentioned above, the visual weight of an object is defined as its area times its optical density, and in the literature[3] this has been applied to images too.

When the graphic artist inserts an image into a document[13], the center of gravity towards which he is expecting the document viewer to look at has less to do with the image optical density than, rather, the image composition and which areas of the image are in high quality for the human eye to linger and examine. This is specially true in high quality photography and design magazines, billboards, even websites, where low depth of field images are heavily used.

An image quality map is presented below, where the areas with higher quality will be assumed to have higher visual weight when balancing such image with a document.

Image segmentation[4] is performed on the input image (see figure 2.b), so that each region is assigned a certain quality level[5]. The rational behind this is that if all the pixels in a region belong to the same object, it can be assumed, in a normal situation, that all these pixels will have similar objective quality. This will generate a map with multiple regions, each of them having its own quality.

The quality map is defined below as the combination of three different maps (sharpness, contrast and chroma) as shown in figure 1.

The focused saliency map[6] (FSM) is the first step towards the final sharpness map used in this approach. Most of the energy in the FSM corresponds to the object in focus, while a large amount of the energy of the out of focus region is removed efficiently. This provides for an excellent starting point to effectively segment the low depth of field images mentioned above, and it also does a quite good job at detecting different levels of sharpness (without being an accurate sharpness measure as explained below).

In recent subjective visual tests[7] on Just Noticeable Blurriness, it was found that the human visual system has different responses to blur /sharpness at different contrast levels. The resulting non-linear function of the contrast (figure 2.c) can be factored into the sharpness measure to be used, and in this approach it is applied to the FSM (figure 2.d). This results in a much more homogeneous FSM on the sharp object.

In order to separate the in focus regions from the out of focus regions, a bilateral filter is applied on the FSM (figure 2.e), and further morphologically filtered (figure 2.f) in order to generate an in focus mark map that can be intersected with the segmented image. For each of the resulting regions an approximate sharpness value can be derived from the contrast corrected FSM (figure 2.d). A more accurate sharpness map is obtained by combining the contrast corrected

FSM with a more accurate sharpness measure[8,9], where each of these additional measures are also corrected with the non-linear contrast function.

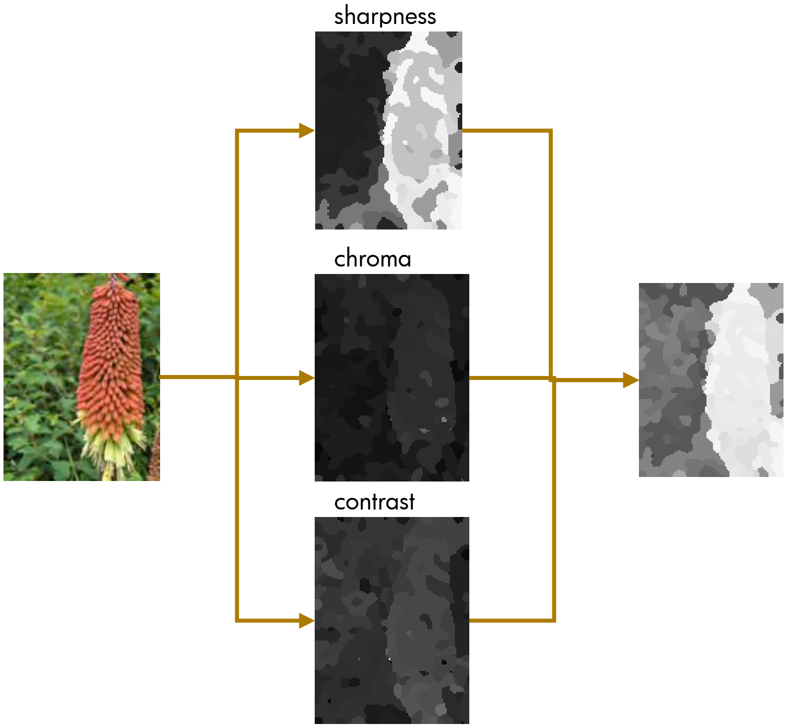


Figure 1. Combination of the sharpness map, contrast map and chroma map into the final quality map.

In low depth of field situations, the described sharpness map is enough for the presented approach to work well. In other situations some more assumptions need to be made in order to guess where the document viewer's attention will be set. Visual attention or saliency models are being used extensively in the literature[10], which do a good job at guessing exactly that. In this region based approach, though, it is not straightforward to extend the saliency models. Instead, a simpler local contrast measure is implemented, which assumes a certain image size and viewing distance. This is calculated on a region by region basis, and a contrast map (figure 1) is generated.

A third and final map is generated by calculating the local chroma content in each of the regions. This map is created after[11], where the relation between perceptual image quality and naturalness was investigated by varying the colorfulness and hue of color images of natural scenes, and concluding that human observers prefer more colorful images.

Both the contrast and chroma map complement the sharpness map, but never dominate it, i.e., if a region is out of focus, there is so much contrast and chroma can add to the final quality map (see figure 1).

2.3. DOCUMENT BALANCE BY MEANS OF THE QUALITY MAP

Balancing photographs and documents has been tried in[1], where color harmony was accomplished by looking at the relationship between image regions and document regions, both colorimetrically and spatially, i.e., region with harmonizing colors can be weighed higher if they are close together or opposing in the document, and hence balancing one another.

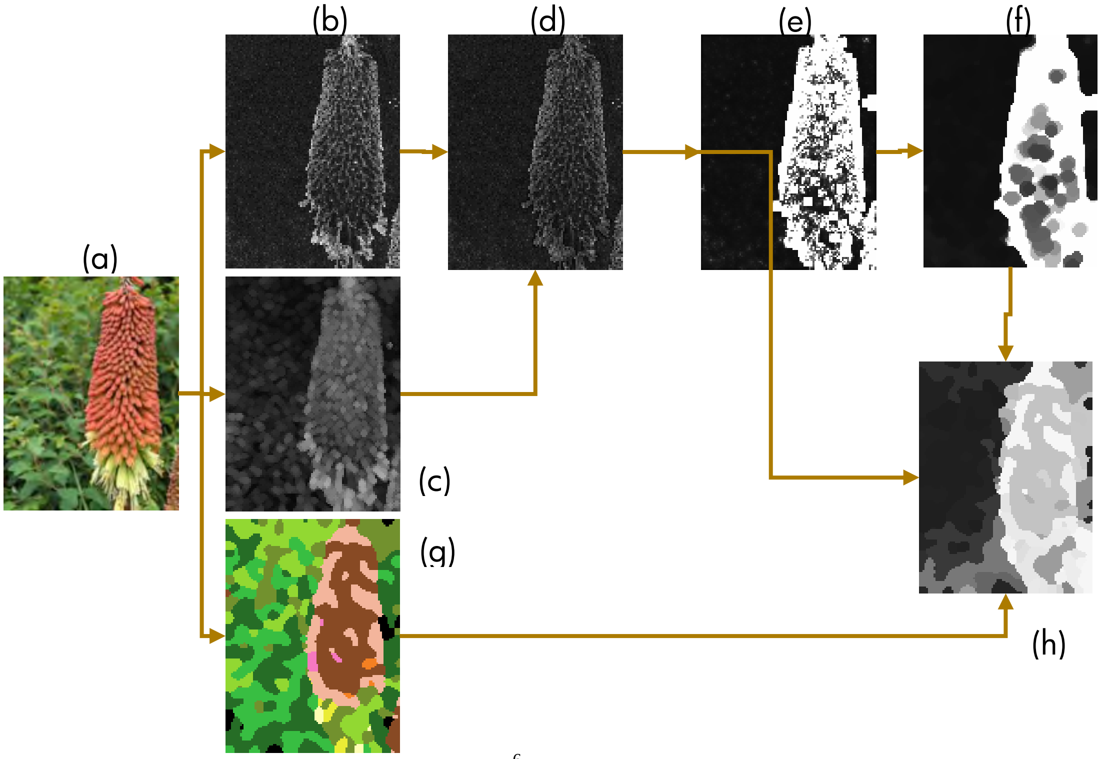


Figure 2. (a) original image, (b) FSM as described in[6] , (c) contrast map, (d) contrast corrected FSM, (e) bilateral filtering, (f) morphological processing, (g) region decomposition), (h) final sharpness map.

Color balance is not enough, as stated above, since the main object of interest (in high quality) is the area that should be weighed higher in the image visual weight map. The current approach combines both techniques when querying an image database, i.e., both the balance query and the color harmony query (see figure 6).

In order to perform the query, a simple model for the image quality map is presented below, which can be extracted and stored with each of the images as metadata for fast indexing.

2.3.1. Image quality map abstraction

Once the image quality map is obtained (see section 2.2 above), it has to be abstracted into an easy to use model for fast querying and retrieval. The current approach thresholds the quality map (see figure 3), and the resulting region/s are approximated by an ellipse.

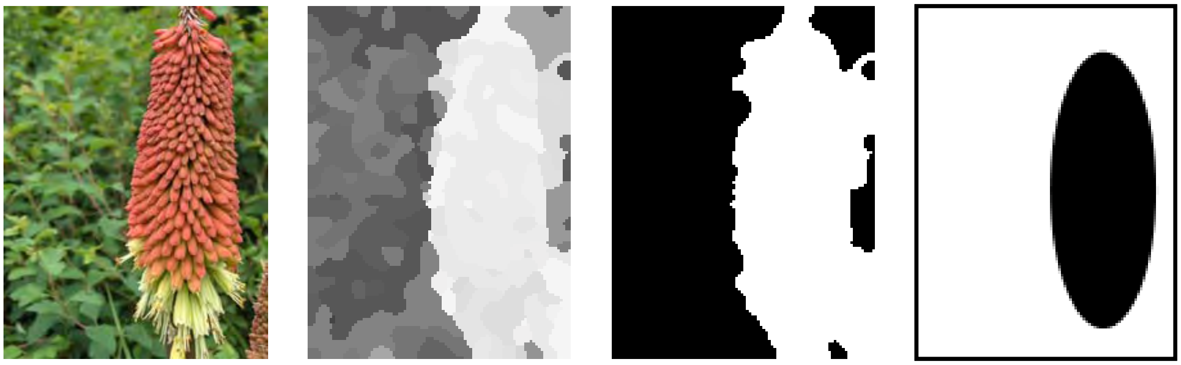
Two different thresholds have been experimented with. First, a fixed threshold for all images, and second, an adaptive threshold to the image quality map content. Each has advantages and disadvantages.

In the case that the image collection is known to have only high quality pictures, having an adaptive threshold makes most sense, since it is known beforehand that at least there is a high quality region in each image. This takes care of some artistic soft focus photographs, and/or abstract photography.

In the case of a consumer photo collection, such assumption cannot be made since excellent photos coexist with very bad ones. Therefore a fixed threshold is necessary, meaning that those worse shots will never be retrieved in a balance query.

The resulting thresholded image is then approximated by an ellipse, (centroid plus spread or axes). Best results are obtained if a quality map weighted centroid is used.

Both centroids and spreads are expressed in percentage of width and height. This allows for querying both landscape and portrait photographs with little added complexity.

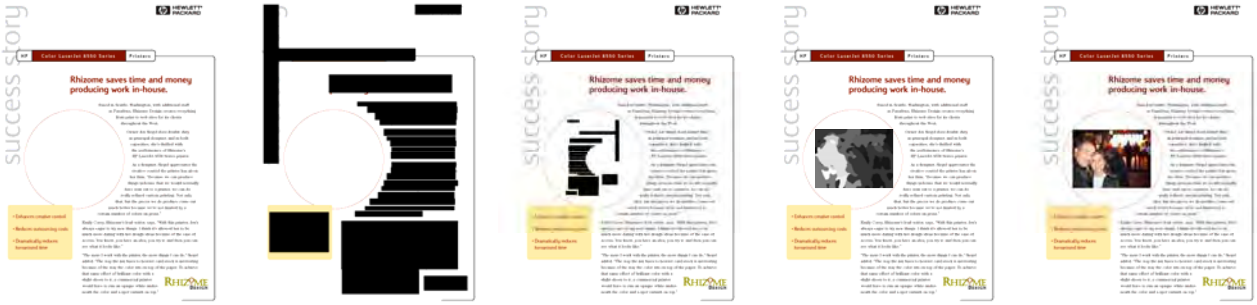


(a) (b) (c) (d) (e)

Figure 3, (a) original, (b) quality map, (c) thresholded quality map, (d) map abstract representation by an ellipse

2.3.2. Document visual weight map abstraction

In order to perform a fast query, a simple abstraction of the document's visual weight map is needed[12]. For this reason, the centroids and size in pixels for each of the objects in the document are calculated (see figure 4.b). All these are then combined to find the centroid for the whole document visual weight map, and its spread.



(a) (b) (c) (d) (e)

Figure 4. (a) original document, (b) extracted document visual weight map consisting of multiple document objects, (c) mirrored visual weight map in photo area, (d) retrieved image quality map from the collection, (e) retrieved image.

For all objects in the document visual weight map:

$$centroidX_{documentWeight} = 100\frac{\sum_j x_j M_j}{documentWidth\sum_j M_j}; \quad centroidY_{documentWeight} = 100\frac{\sum_j y_j M_j}{documentHeight\sum_j M_j}$$

Where (x_j, y_j) is the centroid coordinates for document object *j*, and M_j is the number of pixels belonging to this document object *j*. Again, a percentage of width and height are calculated to ease the query process (query both portrait and landscape photographs).

2.3.3. Balance measure between image and document

The hypothesis that was made is to approximate the presented image quality map as the image visual weight map. This quality map is then compared with the document's visual weight map, and a measure of balance between the two is calculated.

The image collection needs to be queried based on a certain balance criteria that the graphic designer or user need to specify. In this implementation only two criteria are possible:

1. Left Right symmetrical balance: implemented as a horizontal symmetry of the document visual weight map (see figure 4). In this case the centroid for such query would be: $centroidX_{query} = (100 - centroidX_{documentWeight}); \quad centroidY_{query} = centroidY_{documentWeight}$
2. Centered symmetrical balance: implemented as a center symmetry of the document visual weight map: $centroidX_{query} = (100 - centroidX_{documentWeight}); \quad centroidY_{query} = (100 - centroidY_{documentWeight})$

This modified query centroid (figure 4.c) is the one that will be compared with the image quality map centroid. The spread of the document visual quality map is not changed.

In the experiments it was seen that the balance was reduced roughly inversely proportional to the square of the centroid distance between the mirrored document weight map and the image quality map. The spread of the respective maps was less relevant, and the measure was set to be inversely proportional to its difference (see below).

Balance measure between image and document query:

$$balance\_measure_i = \frac{regionQuality_i}{1 + centroidDist_i^{\,2} + sigmaDist_i} \qquad \text{equation 1.}$$

Where regionQuality is an optional term, and is basically a 2D integral of the desired map the user wants to weigh into the equation, i.e., if the user wants to weigh chroma in the high quality region, this term would add up all the chroma values in the chroma map (see figure 1) in a region under the ellipse abstracting the mirrored document's visual weight map. See figure 8 for an image retrieval example using regionQuality, where the chroma has been factored in. If the regionQuality term needs to be used, then the whole quality map needs to be stored as metadata, increasing the needed storage size as well as computation time;

The other terms in the above formula are:

$$centroidDist_i = \sqrt{(centroidX_i - centroidX_{query})^2 + (centroidY_i - centroidY_{query})^2}$$

and

$$sigmaDist_i = \sqrt{(sigmaX_i - sigmaX_{query})^2 + (sigmaY_i - sigmaY_{query})^2}$$

Where ($centroidX_i$, $centroidY_i$) are the coordinates of the high quality ellipse weighted centroid for image *i*, and $sigmaX_i$ and $sigmaY_i$ are the spread of such ellipse for image *i*; and $sigmaX_{query}$ and $sigmaY_{query}$ are directly proportional to the spread of the document visual weight map.

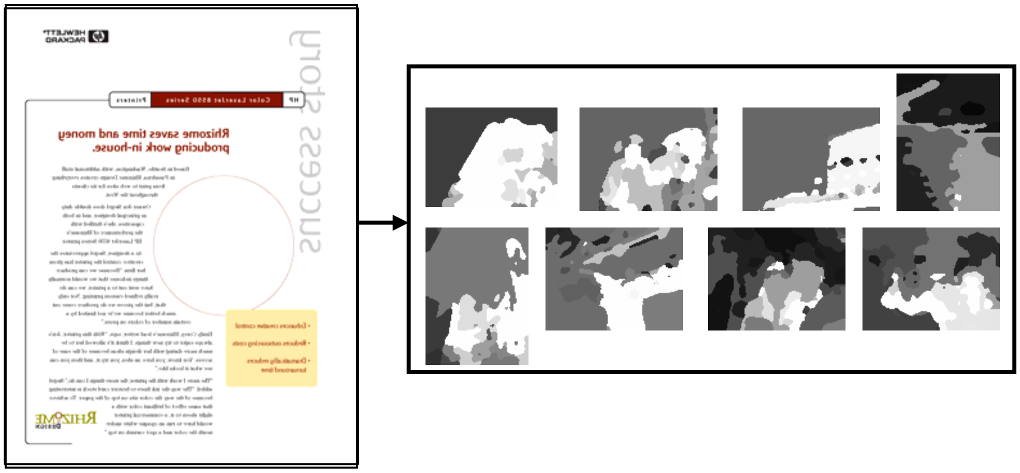
For a particular document, after abstraction, the centroid and spread of its visual weight map is used to index the image collection. The balance measure is calculated for every image in the collection, and either the image with the highest score, or the set of N images with the highest scores, are retrieved and presented to the user.

3. RESULTS

The experiments were performed on 850 personal images, i.e., consumer type images. For this reason, a fixed threshold was used for the quality map abstraction (section 2.3.1.).

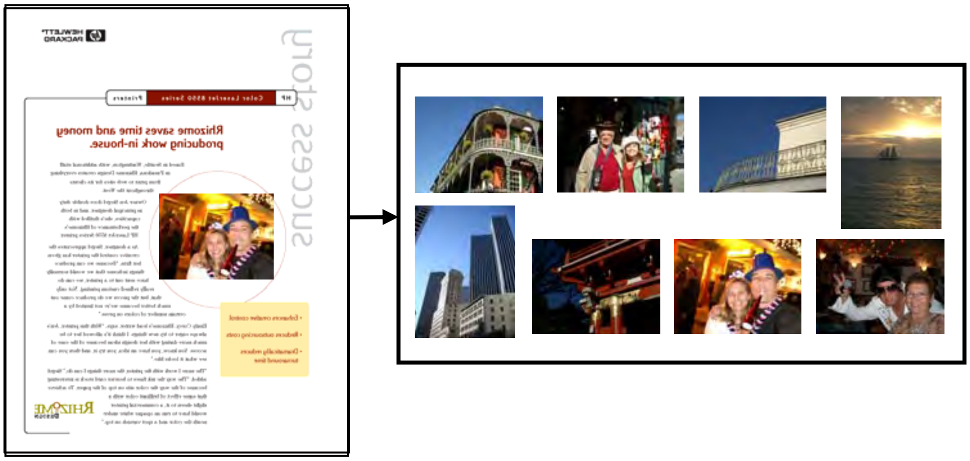
The results presented in this section were generated querying the collection with Left Right symmetrical balance: implemented as a horizontal symmetry of the document visual weight map.

Figure 5 presents the retrieved results using a certain document with its visual weight centered towards the left third. Notice the weight towards the right third of the image quality maps. Figure 6.b presents the actual top 8 retrieved images with such a document query. And figure 6.a presents the top retrieved image after applying the color harmonization query on the balance query results in figure 6.b. The algorithm managed to find pretty good result both for balance and color harmony within the relatively small 850 image collection.



(a) (b)

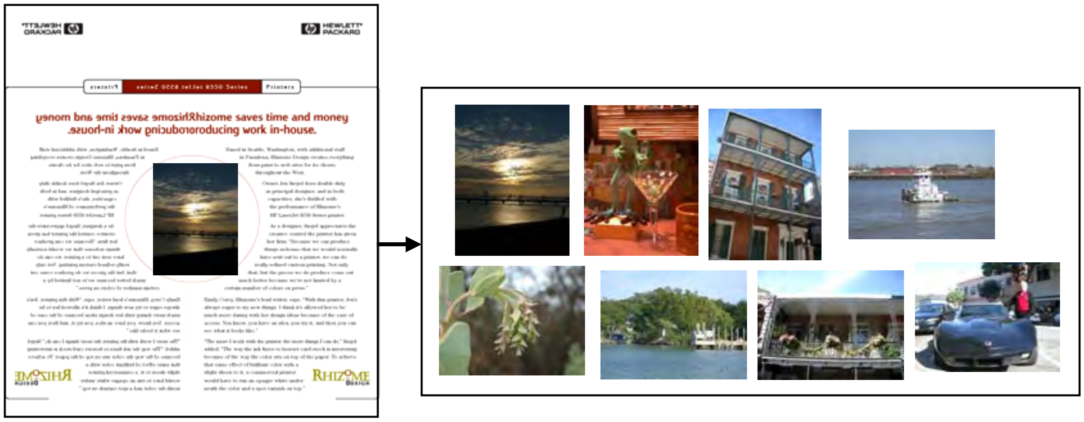
Figure 5. (a) document query, (b) top 8 quality maps retrieved from the collection



(a) (b)

Figure 6. combination of balance and color harmony queries. (a) top retrieved image after performing the color harmony query[1] on the results of the balance query. (b) Top 8 retrieved images with the balance query.

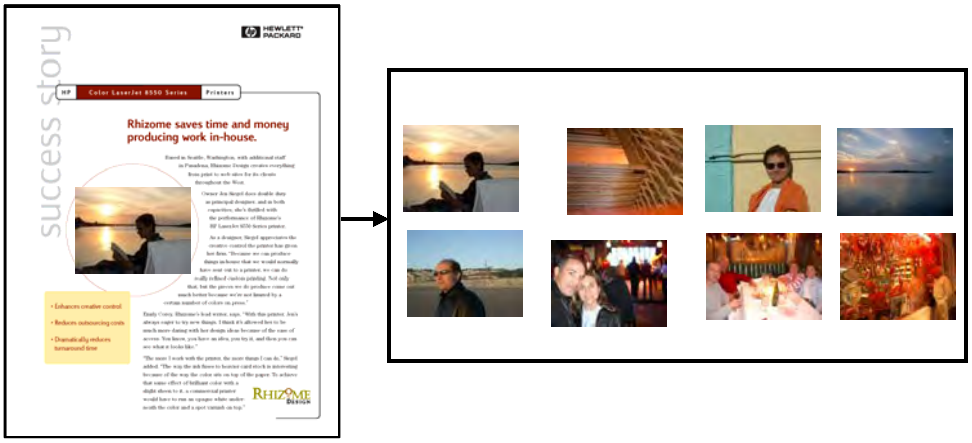
Figure 7 presents a balanced document, where the photograph has to lie in the center of it. In this situation it is expected to retrieve images with a quality map with its centroid close to the center of the image.



(a) (b)

Figure 7. (a) top retrieved image with the balance query. (b) Top 8 retrieved images with the balance query.

Figure 8 presents the query results with a document balanced towards the right third. The regionQuality factor in equation 1 was used, and it integrated the chroma map for this specific example. The results should have high quality towards the left third, and also high chroma content around that left third.


success story
HEWLETT PACKARD
HP
Color LaserJet 8550 Series
Printers
Rhizome saves time and money
producing work in-house.
• Enhances creative control
• Reduces outsourcing costs
• Dramatically reduces turnaround time
RHIZOME DESIGN

(a) (b)

Figure 8. balance query with *regionQuality* performed on the chroma map. (a) top retrieved image with the chroma map weighted balance query. (b) Top 8 retrieved images with the chroma map weighted balance query .

These results confirm our hypothesis, and therefore it is safe to use the image quality map as the image visual weight map when performing document balance analysis.

4. CONCLUSIONS AND FUTURE WORK

A new image retrieval method has been presented, in which the document where the image is to be inserted is used as query. The algorithm calculates a balance measure between the document and each of the images in the collection, retrieving the ones that have a higher balance score. The image visual weight map, used in the balance calculation, has been successfully approximated by a new image quality map that takes into consideration sharpness, contrast and chroma. This retrieval method has been successfully combined with the color harmony retrieval method presented in[1].

Future work needs to be done in adding extra visual saliency features to the image visual weight map, which may solve some of the problems encountered with evenly sharp images.

The color version of this paper: please refer to the HP-Labs web-site (http://www.hpl.hp.com/techreports/).

REFERENCES

1. P. Obrador, "Automatic color scheme picker for document templates based on image analysis and dual problem", EI126, proceedings of Digital Publishing Conference, 2006, San Jose.
2. H. Balinsky, "Evaluating interface aesthetics: a measure of symmetry", HP-Labs Technical report
3. S.J Harrington, J.F. Naveda, R.P. Jones, P. Roetling and N. Thakkar, "Aesthetic measures for automated document layout", ACM Symposium on document engineering, 2004.
4. P. Obrador, "Multiresolution Color Patch Extraction", Electronic Imaging, VCIP 2006, San Jose.
5. P. Obrador, "Method and system for image quality calculation", Filed Patent Docket HP 200503391-1.
6. H. Li and K.N. Ngan, "Unsupervised segmentation of defocused video based on matting model", ICIP 2006, Atlanta.
7. R. Ferzli and L. J. Karam, "A Human Visual System-Based Model for Blur/Sharpness Perception," 2nd International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, 2006.
8. Shaked, D. and Tastl, I. , "Sharpness measure: towards automatic image enhancement", ICIP 2005. 11-14 Sept. 2005, Volume: 1, On page(s): I- 937-40
9. S.H. Lim, J. Yen and P. Wu, "Detection of ill-focused Digital Photographs," in Proceedings of the HP TECHCON 2004, Orlando, FL, June 2004
10. L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE transactions on pattern analysis and machine intelligence", vol. 20, no.11: 1254-1259.
11. H. de Ridder, F. J. J. Blommaert and E. A. Fedorovskaya, "Naturalness and image quality: chroma and hue variation in color images of natural scenes", Proceedings of SPIE -- Volume 2411 Human Vision, Visual Processing, and Digital Display VI, April 1995, pp. 51-61
12. Chao, H., Fan, J., "Layout and Content Extraction for PDF Documents." In proceeding of IAPR Int. workshop on Document Analysis System, 2004.
13. P. Bonnici and L. Proud, "Designing with photographs", Rotovision SA, 1998.

C.3 Document layout and color driven image retrieval

Referenced as [C15].

Document Layout and Color Driven Image Retrieval

Pere Obrador
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, California 94304, USA
+1 650 857 42 70
pere.obrador@hp.com

ABSTRACT

This paper presents a contribution to image indexing applied to the document creation task. The presented method ranks a set of photographs based on how well they aesthetically *work* within a predefined document. Color harmony, document visual balance and image quality are taken into consideration. A user study conducted on people with a range of expertise in document creation helped gather the right visual features to consider by the algorithm. This shows some benefits for the traditional document creation task, as well as for the case of ever-changing web page banner colors and layout.

Categories and Subject Descriptors

H.3.1. [**Information Storage and Retrieval**]: Content Analysis and Indexing - *Indexing methods.*

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Image analysis and indexing, document balance, color harmony.

1. INTRODUCTION

When selecting an image to accompany a document, graphic artists/illustrators will usually follow three basic steps: (a) select the image based on content (i.e., semantic relevance to the article); (b) image quality; (c) document's layout [3], color scheme [5] and image composition [7]; (d) adjustments (e.g. color, crop) may be performed on the final image. In real life situations, graphic designers and journalists do not have all the time they need to find the photograph that would *best* aesthetically match a document (or perform those adjustments mentioned above); rather, they tend to make *acceptable* choices [3]. The work described below focuses on steps (b) and (c) above, where image quality, layout and color scheme are taken into consideration. Given a document (query), with a blank area to accommodate an image (i.e., the layout of the document is not altered), the goal of the presented algorithm is to rank a set of images based on their image quality and how well they visually balance the document's layout and color harmonize with the document's color scheme. A set of user studies were performed in order to learn what users value when aesthetically matching documents and images. The selection of features was driven by such findings, and the way they were combined is explained below. The results at the end of this paper show that these two features work in a somewhat orthogonal way, and the combination of the two produces very promising results.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

2. USER STUDIES

A set of documents were printed with a diverse set of images, shown to 8 subjects: expert photo-book creator, expert illustrator/publisher, expert in color science, two experienced photographers, and three other users. Their feedback has been condensed into the following list of findings:

F1) The images should have little clutter, with well defined homogeneous regions (also described in [4]).
F2) Left-right symmetrical visual balance [6] is preferred as opposed to center symmetry for document balancing.
F3) Analogous color harmonies [5] are preferred, with large homogeneous color patches representing such colors. One main color, with one accent color seemed to be preferred.
F4) Slight color tone differences between regions are singled out.
F5) High contrasts between color patches in the document and analogous color patches in the image are undesirable. i.e., having the analogous colors close together is favorable.
F6) Users will reject images if one of the features, either visual balance or color harmony, is below a certain threshold, no matter how good the other feature may be. This threshold seems to depend on the level of expertise for each user.
F7) Chosen images have to be above a certain quality threshold.

3. ALGORITHM DESCRIPTION

From these findings, a combination of quality assessment, visual balance and color harmony seems to be the right approach to solve this problem. Each of these features has been used in the literature individually, but when combined the results are greatly improved (suggested by F6), as will be shown below.

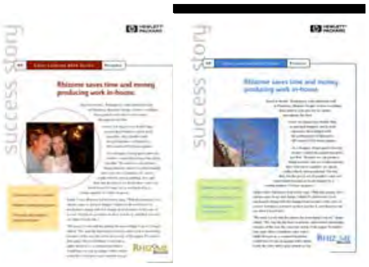


Figure 1. Top results.

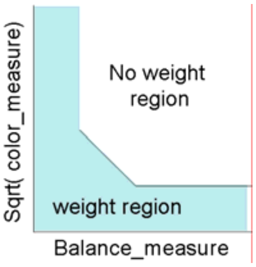


Figure 2. Weighted hull.

3.1 Computing the Color Harmony Measure

Color harmony image indexing has been proposed where certain attributes are derived of such harmony, and used in keyword search [9] or in query by drawing [2]. In [5] the author presented an image indexing technique based on how well the image color harmonizes with the document color layout where it must be inserted; this is proportional to the color patch size (*patchSize*), it can be used in analogous harmony tuned to be very sensitive to color changes (*colorDist*), and it can be tuned to favor closer patches of analogous colors (*centroidDist*). See [5] for details.

$$color\_measure = \sum_{i \in document} \sum_{j \in image} \frac{\sqrt{patchSize_i * patchSize_j}}{\left(1 + centroidDist_{ij}\right) * \left(1 + \left(colorDist_{ij}\right)^4\right)}$$

3.2 Computing the Visual Balance Measure

Visual balance is obtained by balancing the visual weight (i.e., left-right, F2) for all objects in the document. Objects are paragraphs, titles, banners, images, etc. For an image object, each of the regions within the image should be considered [4][6], since certain regions will have higher visual weight than others [1]. In [6] the author showed that the image visual weight map can be approximated by an image appeal map that takes into consideration sharpness, contrast and chroma; this map is thresholded and the resulting region (*visualWeight_region*) is used to measure how well this image balances the rest of the objects in the document. The visual balance measure from [6] was improved by incorporating the difference in region sizes (*width_height_dist*) as presented in [8] (see [6] for details):

$$balance\_measure = \frac{visualWeight\_regionQuality}{1 + 5 * visualWeightCentroidDist^2 + width\_height\_dist^2}$$

3.3 Combining Balance and Color Harmony

The color harmony measure was found to be less relevant than the visual balance measure (F1). Also the fact that when one of the features is below a threshold, the overall result is considered unacceptable (F6), yielded the following measure between image *i* and the query document:

$$combined\_measure\_m_i = balance\_measure_i * \sqrt{color\_measure_i}$$

This formula does not take care of the extreme cases (close to the axes) where one of the measures may be small and the other may be very large. Therefore, a hard-coded threshold (optimized from the gathered ground truth, see next section), was integrated in the algorithm, down-weighting the images whose coordinates would lie below the curve (Figure 2) (F6). In future work, a machine learning approach will be developed, in order to optimize these thresholds, based on a larger set of ground truth.

4. RESULTS AND CONCLUSION

A collection spanning 9 days was used (882 images overall), all images, good and bad were considered, and no image adjustments were performed. Two different documents (Figure 1), were printed with each photograph, and were tagged as ground truth by the same users as above, into three sets: *works*, *maybe* (weighted at 50% of *works* in the retrieval experiments), and *doesn't_work.* The proposed method was run on the collection set twice (once per query). For comparison purposes, the ranking based on color harmony only, and the ranking based on document balance only were also run individually. Figure 3 shows those three average precision-recall curves. The color harmony result on its own is the worst of the three. This is due to the fact that the method in [5] still allows for certain levels of non-homogeneity to be present with a high color harmony score, and as mentioned in our user studies (F1) and [4], this is an important factor when assessing the relevance of a photograph. The visual balance only result, instead, is reasonably good on its own since it favors images with a visual quality map concentrated in a particular area, which favors homogeneity. By combining the two algorithms: the nice homogeneity and symmetry of the visual balance complement the color harmony's lack of such; and the color harmony measure retrieves the well balanced images with the right color scheme at the top of the list.

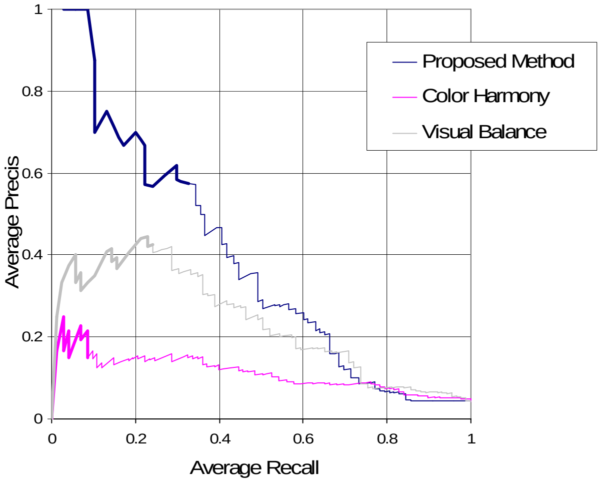


Figure 3. Average precision-recall graph. Thick line: top 20.

5. REFERENCES

[1] Bajcsy, R. Active Perception, Proceedings of the IEEE, vol. 76, no. 8, pp. 996-1005, 1988.

[2] Corridoni, J.M., Del Bimbo, A., De Magistris, S. Querying and retreiving pictorial data using semantics induced by colour quality and arrangement, Proc. Multimedia, 1996.

[3] Markkula, M.and Sormunen, E. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval,* 1:259-285, 2000.

[4] Martinet, J., Chiaramella, Y. and Mulhem, P. A model for weighting image objects in home photographs. In ACM CIKM'05, pages 760-767, Bremen, Germany, 2005.

[5] Obrador, P. Automatic color scheme picker for document templates based on image analysis and dual problem, in Proc. SPIE, vol. 6076, San Jose, CA, 2006.

[6] Obrador, P., Content Selection based on Compositional Image Quality, in Proc. SPIE, vol. 6500, San Jose, CA 2007.

[7] Savakis, A., Etz, S., and Loui A. Evaluation of image appeal in consumer photography, in *Proc. SPIE* vol. 3959, 2000.

[8] Smith, R. J. and Chang, S.-F. Integrated Spatial and Feature Image Query, *Multimedia Systems,* 7(2):129--140, 1999.

[9] Vasile, A., Bender, W.R. Image query based on color harmony, in Proc. SPIE Vol. 4299, San Jose, CA, 2001

APPENDIX D

Full list of published documents by the author

This is a comprehensive list of all published documents by the author, including conference papers, journals, granted patents, published filed patent applications –*i.e.*, not yet granted, and filed patent applications –*i.e.*, not yet published. In the Bibliography none of these publications appear.

All the patents have been filed in the United States of America, except for [C79], which has been filed only internationally at the WIPO, [C81] only in the European Union, and [82,83] only in Spain. Several of the patents filed in the USA have also been filed in Japan, in the EU, and at the WIPO simultaneously.

D.1 Academic publications

[C1] A. Ciancio, A. L. N. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64 – 75, 2011.

[C2] P. Obrador, R. de Oliveira, and N. Oliver. Supporting personal photo storytelling for social albums. In *Proceedings of the international conference on Multimedia*, MM '10, pages 561–570, New York, NY, USA, 2010. ACM.

[C3] P. Obrador, R. de Oliveira, and N. Oliver. Audience dependent photo collection summarization. In *Proceedings of the international conference on Multimedia*, Grand Challenge'10, New York, NY, USA, 2010. ACM.

[C4] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *International Conference on Image Processing*, pages 3185 – 3188. IEEE, 2010.

[C5] A. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. *Computer Vision, ECCV 2010*, 6315:1–14, 2010.

[C6] A. Ciancio, A. L. N. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *Electronics Letters*, 45(23):1162 – 1163, 2009.

[C7] X. Anguera, P. Obrador, T. Adamek, D. Marimon, and N. Oliver. Telefonica Research Content-Based Copy Detection TRECVID Submission. In *NIST Trecvid 2009 Workshop notebook paper.* 2009

[C8] X. Anguera, P. Obrador, and N. Oliver. Multimodal video copy detection of social media. In *WSM '09: Proc. of the* 1^{st} *SIGMM workshop on Social media.* ACM, 2009.

[C9] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver. The role of tags and image aesthetics in social image search. In *WSM '09: Proc. of the* 1^{st} *SIGMM workshop on Social media*, pages 65–72. ACM, 2009.

[C10] P. Obrador and N. Moroney. Automatic image selection by means of a hierarchical scalable collection representation. In *Proc. SPIE, Vol. 7257.* IS&T/SPIE, 2009.

[C11] P. Obrador and N. Moroney. Low level features for image appeal measurement. In *SPIE, Electronic Imaging, Image Quality and System Performance VI*, volume 7242, pages 72420T–1–72420T–12. IS&T/SPIE, 2009.

[C12] N. Moroney, P. Obrador, and G. Beretta. Lexical image processing. In *Proceedings of the 16th IS&T/SID Color Imaging Conference*, pages 268–273, 2008.

[C13] P. Obrador. Region based image appeal metric for consumer photos. In *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pages 696–701, 2008.

[C14] P. Obrador, N. Moroney, I. MacDowell, and E. OBrien-Strain. Softbacks: single click photo selection for photo-book creation. In *Symposium on document engineering.* ACM, 2008.

[C15] P. Obrador. Document layout and color driven image retrieval. In *SIGIR.* ACM, 2007.

[C16] P. Obrador. Content selection based on compositional image quality. In *Electronic Imaging vol. 6500.* IS&T/SPIE, 2007.

[C17] P. Obrador. Automatic color scheme picker for document templates based on image analysis and dual problem. In *Proceedings of SPIE.* IS&T/SPIE, 2006.

[C18] P. Obrador. Multiresolution color patch extraction. In *Proceedings of SPIE*, volume 6077, pages 788–799. IS&T/SPIE, 2006.

[C19] G. Vondran, H. Chao, X. Lin, P. Joshi, D. Beyer, and P. Obrador. Automated Campaign System. In *SPIE Digital Publishing Conference, part of the Electronic Imaging Symposium*, San Jose, California. IS&T/SPIE, 2006.

[C20] P. Wu and P. Obrador. Personal Video Manager: Managing and Mining Home Video Collections. In *International Symposium on Visual Communications & Image Processing.* Beijing, China. SPIE/IEEE, 2005.

[C21] Q. Lin, T. Zhang, M. Chen, Y. Deng, and P. Obrador. Mining Home Video for Photos. In *HP-Labs technical report HPL-2004-80.* 2004.

[C22] U. Gargi and P. Obrador. Multimedia links and metadata channels for distributed shared browsing and authoring. In *Electronic Imaging 2004 Internet Imaging V Conference.* San Jose, California. IS&T/SPIE, 2004.

[C23] A. Ortega and P. Obrador Compound document compression. *Novatica*, sep-oct 1996. N 123, pp.29-34.

[C24] P. Obrador, Eye-tracked based Stereo Video Compression Scheme. In *First International Workshop on 3D Imaging.* Santorini. Greece. September 1995.

[C25] G. Caso, P. Obrador, and C.-C. J. Kuo, Fast Methods for Fractal Image Compression. In *International Symposium on Visual Communications & Image Processing.* Taipei, Taiwan. SPIE/IEEE, 1995.

[C26] P. Obrador, G. Caso and C.-C. J. Kuo, A Fractal based Method for Textured Image Compression. In *Symposium on Electronic Imaging: Science & Technology.* San Jose, California. IS&T/SPIE, 1995.

[C27] P. Obrador and F. Vallverdú, Vocoder LPC at 2400 bps running on TMS320C30 & DSP32C. In *URSI Simpósium de Telecomunicaciones.* Caceres, Spain, 9-1991.

[C28] P. Obrador, Vocoder LPC-10 with silence detector. Real time implementation. *Masters Thesis.* Universitat Politècnica de Catalunya. Barcelona. 1991.

D.2 Granted patents

[C29] V. Alfaro, P. Obrador, and J. Gonzalez. Enhancement technique for asymmetrical print resolution. Hewlett-Packard Company, U.S. Patent 7185963, March 2007.

[C30] R. P. Cazier, N. M. Moroney, M. D. Craig, and P. Obrador. Method and apparatus for the creation of cartoon effect videos. Hewlett-Packard Company, U.S. Patent 7466321, December 2008.

[C31] Y. Deng, Q. Lin, J. Fan, P. Obrador, and U. Gargi. Digital camera with panoramic image capture. Hewlett-Packard Company, U.S. Patent 7746404, June 2010.

[C32] P. Obrador. Video indexing based on viewers' behavior and emotion feedback. Hewlett-Packard Company, U.S. Patent 6585521, July 2003.

[C33] P. Obrador. Method and apparatus for low memory rendering. Hewlett-Packard Company, U.S. Patent 6847467, January 2005.

[C34] P. Obrador. Resolution dependent image compression. Hewlett-Packard Company, U.S. Patent 6968082, November 2005.

[C35] P. Obrador. Video indexing using high quality sound. Hewlett-Packard Company, U.S. Patent 6931201, August 2005.

[C36] P. Obrador. Method and system for processing images using histograms. Hewlett-Packard Company, U.S. Patent 7058220, June 2006.

[C37] P. Obrador. Presenting a collection of media objects. Hewlett-Packard Company, U.S. Patent 7149755, December 2006.

[C38] P. Obrador. Scalably presenting a collection of media objects. Hewlett-Packard Company, U.S. Patent 7131059, October 2006.

[C39] P. Obrador. Method and system for image border color selection. Hewlett-Packard Company, U.S. Patent 7424147, September 2008.

[C40] P. Obrador. Still image capturing of user-selected portions of image frames. Hewlett-Packard Company, U.S. Patent 7388605, June 2008.

[C41] P. Obrador. System and method of manual indexing of image data. Hewlett-Packard Company, U.S. Patent 7444068, October 2008.

[C42] P. Obrador. Systems and methods for sampling an image sensor. Hewlett-Packard Company, U.S. Patent 7483059, January 2009.

[C43] P. Obrador. Method and system for image quality calculation. Hewlett-Packard Company, U.S. Patent 7693304, April 2010.

[C44] P. Obrador, J. Gonzalez, and J. Noh. Efficient storage of dithered raster image data in a printer or the like. Hewlett-Packard Company, U.S. Patent 6137589, October 2000.

[C45] P. Obrador and N. M. Moroney. Image processing methods, image management systems, and articles of manufacture. Hewlett-Packard Company, U.S. Patent 7848577, December 2010.

[C46] P. Obrador and D. Tretter. Concurrent dual pipeline for acquisition, processing and transmission of digital video and high resolution digital still photographs. Hewlett-Packard Company, U.S. Patent 6961083, November 2005.

[C47] P. Obrador, D. Tretter, and A. D. Silverstein. Remote high resolution photography and video recording using a streaming video as a view-finder. Hewlett-Packard Company, U.S. Patent 7450157, November 2008.

[C48] P. Obrador and P. Wu. Image management through lexical representations. Hewlett-Packard Company, U.S. Patent 7755646, July 2010.

[C49] P. Obrador, P. Wu, and J. Yen. Image management. Hewlett-Packard Company, U.S. Patent 7860319, December 2010.

[C50] P. Obrador and T. Zhang. System and method for indexing videos based on speaker distinction. Hewlett-Packard Company, U.S. Patent 7184955, February 2007.

[C51] P. Wu and P. Obrador. System and method for representing digital media. Hewlett-Packard Company, U.S. Patent 7692562, April 2010.

[C52] J. Yen, M. Chen, R. Samadani, P. Obrador, and H. Luo. Providing optimized digital images. Hewlett-Packard Company, U.S. Patent 7595823, September 2009.

D.3 Published patent applications

[C53] H. Chao, M. Indrani, G. Vondran, X. Lin, P. M. Joshi, D. M. Beyer, B. C. Atkins, P. Obrador, and A. X. Zhang. Producing marketing items for a marketing campaign. Hewlett-Packard Company, U.S. Patent Application 20070022003, January 2007.

[C54] U. Gargi, P. Obrador, P. M. Walker, and P. Boerger. Disc content enhancement systems and methods. Hewlett-Packard Company, U.S. Patent Application 20060153533, July 2006.

[C55] U. Gargi, P. Wu, and P. Obrador. Distributed processing with metadata placeholders. Hewlett-Packard Company, U.S. Patent Application 20070112779, May 2007.

[C56] S. R. Girshick, P. Obrador, and T. Zhang. Automatically editing video data. Hewlett-Packard Company, U.S. Patent Application 20080019669, January 2008.

[C57] P. Obrador. Multi resolution printing. Hewlett-Packard Company, U.S. Patent Application 20020186383, December 2002.

[C58] P. Obrador. Delayed encoding based joint video and still image pipeline with still burst mode. Hewlett-Packard Company, U.S. Patent Application 20030169278, September 2003.

[C59] P. Obrador. Media object management. Hewlett-Packard Company, U.S. Patent Application 20030191776, October 2003.

[C60] P. Obrador. Multi-resolution boundary encoding applied to region based still image and video encoding. Hewlett-Packard Company, U.S. Patent

Application 20030002582, January 2003.

[C61] P. Obrador. System and method for efficiently managing video files. Hewlett-Packard Company, U.S. Patent Application 20030212993, November 2003.

[C62] P. Obrador. Video indexing using high resolution still images. Hewlett-Packard Company, U.S. Patent Application 20030118329, June 2003.

[C63] P. Obrador. Video transcoder based joint video and still image pipeline with still burst mode. Hewlett-Packard Company, U.S. Patent Application 20030169818, September 2003.

[C64] P. Obrador. Image capture systems and methods. Hewlett-Packard Company, U.S. Patent Application 20040090548, May 2004.

[C65] P. Obrador. Systems and methods of authoring a multimedia file. Hewlett-Packard Company, U.S. Patent Application 20040201609, October 2004.

[C66] P. Obrador. Sequential processing of video data. Hewlett-Packard Company, U.S. Patent Application 20060103736, May 2006.

[C67] P. Obrador. Imaging methods, imaging systems, and articles of manufacture. Hewlett-Packard Company, U.S. Patent Application 20070098238, May 2007.

[C68] P. Obrador. Method for selecting an image for insertion into a document. Hewlett-Packard Company, U.S. Patent Application 20080030752, February 2008.

[C69] P. Obrador. Compositional balance and color driven content retrieval. Hewlett-Packard Company, U.S. Patent Application 20090024580, January 2009.

[C70] P. Obrador. Compositional balance driven content retrieval. Hewlett-Packard Company, U.S. Patent Application 20090024579, January 2009.

[C71] P. Obrador and U. Gargi. Method and system for automatically selecting images from among multiple images. Hewlett-Packard Company, U.S. Patent Application 20060259863, November 2006.

[C72] P. Obrador and Q. Lin. Providing a visual indication of the content of a video by analyzing a likely user intent. Hewlett-Packard Company, U.S. Patent Application 20050231602, October 2005.

[C73] P. Obrador and N. M. Moroney. Image management methods, image management systems, and articles of manufacture. Hewlett-Packard Company, U.S. Patent Application 20080025647, January 2008.

[C74] P. Obrador and T. Zhang. Method and system for onboard camera video editing. Hewlett-Packard Company, U.S. Patent Application 20070283269, December 2007.

[C75] P. Obrador, T. Zhang, and S. R. Girshick. Producing output video from multiple media sources including multiple video sources. Hewlett-Packard Company, U.S. Patent Application 20080019661, January 2008.

[C76] A. D. Silverstein and P. Obrador. System and method for remote controlled photography. Hewlett-Packard Company, U.S. Patent Application 20040066457, April 2004.

[C77] S. Widdowson, C. B. Atkins, U. Gargi, and P. Obrador. Method and system for finding data objects within large data-object libraries. Hewlett-Packard Company, U.S. Patent Application 20070250499, October 2007.

[C78] P. Wu, U. Gargi, P. Obrador, P. M. Walker, and P. Boerger. Search file indicating languages associated with scenes. Hewlett-Packard Company, U.S. Patent Application 20060155680, July 2006.

[C79] P. Obrador. Automatic creation of a scalable relevance ordered representation of an image collection. Hewlett-Packard Company, W.I.P.O. Patent Application WO/2010/021625, February 2010.

D.4 Patent applications

[C80] P. Obrador, R. de Oliveira, and N. Oliver. Automatic Storytelling for Social Albums. Telefonica I+D, Patent Application Registered in the USA. US 13098801, 2011.

[C81] X. Anguera, P. Obrador, and N. Oliver. METHOD FOR DETECTING AUDIO AND VIDEO COPY IN MULTIMEDIA STREAMS. Telefonica I+D, Patent Application Registered in the European Union. PCT / EP2010 / 057588, 2010.

[C82] A. K. Moorthy, P. Obrador, and N. Oliver. Method for the classification of videos. Telefonica I+D, Patent Application Registered in Spain. P201031019, 2010.

[C83] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. Image aesthetics derived from image composition low level features. Telefonica I+D, Patent Application Registered in Spain. P201031332, 2010.

Bibliography [1]

[1] The author's publications do not appear in this list. Instead, they are listed in Appendix D [89].

References

[1] http://www.picasaweb.com.

[2] http://code.google.com/apis/picasaweb.

[3] http://www.quietlyscheming.com/blog/components/flexbook.

[4] Facebook increases photo album limit to 200 pictures. Retrieved in 4/2010 from http://reface.me/news/ facebook-increases-photo-album-limit-to-200-pictures.

[5] Nasa task load index. Retrieved in 4/2010 from http://humansystems.arc.nasa.gov/groups/TLX.

[6] *Digital Still Camera Image File Format Standard, Japan Electronic Industry Development Association*. Tokyo, Japan, 1998.

[7] A. Adams. *The Negative: Exposure and Development*. Boston, 1948.

[8] B. Adams, D. Phung, and S. Venkatesh. Extraction of social context and application to personal multimedia exploration. In *Proc. of the 14th annual ACM international conference on Multimedia*, pages 987–996, 2006.

[9] A. Al-Maskari, P. Clough, and S. M. Users effectiveness and satisfaction for image retrieval. In *Workshop Information Retrieval*, 2006.

[10] X. Amatriain, J. M. Pujol, and N. Oliver. I Like It, I Like It Not. *Proceedings Int. Conf. UMAP'09*, 2009.

[11] D.-E. Amer, A. and A. Mitiche. Reliable and fast structure-oriented video noise estimation. In *ICIP*. IEEE, 2002.

[12] M. G. Ames and L. Manguy. Photoarcs: ludic tools for sharing photographs. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 615–618, New York, NY, USA, 2006. ACM.

[13] C. B. Atkins. Adaptive photo collection page layout. In *ICIP*, pages 2897–2900. IEEE, 2004.

[14] R. Bajcsy. Active perception. In *IEEE, vol. 76, no. 8*, pages 996–1005. IEEE, 1988.

[15] M. Balabanović, L. L. Chu, and G. J. Wolff. Storytelling with digital photographs. In *CHI '00: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 564–571, New York, NY, USA, 2000. ACM.

[16] J. G. Beerends and F. De Caluwe. The influence of video quality on perceived audio quality and vice versa. *Jnl. Aud. Engg. Soc.*, 47(5):355–362, 1999.

[17] F. Bentley, C. Metcalf, and G. Harboe. Personal vs. commercial content: the similarities between consumer use of photos and music. In *CHI '06: Proc. of the SIGCHI conference on Human Factors in computing systems*, pages 667–676. ACM, 2006.

[18] B. Block. *The Visual Story: Seeing the Structure of Film, TV, and New Media.* Focal Press, New York, 2001.

[19] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly, 2008.

[20] BT. 500-11: Methodology for the subjective assessment of the quality of television pictures.. *International Telecommuncation Union, Geneva, Switzerland*, 2002.

[21] D. Cai et al. Hierarchical clustering of www image search results using visual, textual and link information. In *ACM Multimedia.* ACM, 2004.

[22] A. Cavallaro and S. Winkler. Segmentation-driven perceptual quality metrics. In *ICIP*, pages 3543–3546. IEEE, 2004.

[23] C. Cerosaletti and A. Loui. Measuring the perceived aesthetic quality of photographic images. In *Intl. Workshop on Quality of Multimedia Experience*, 2009.

[24] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[25] M. Cherubini et al. Social tagging revamped: Supporting the users' need of self-promotion through persuasive techniques. In *CHI '10: Proc. ACM Int. Conf. on Human Factors in Computing Systems*, Atlanta, USA, April 2010. ACM press.

[26] J. Child. *Studio Photography: Essential Skills.* Focal Press, 4th edition, 2008.

[27] Y. Choi and E. M. Rasmussen. Users relevance criteria in image retrieval in american history. *Information Processing and Management*, 38, 5:695726, 2002.

[28] W.-T. Chu, L. Che-Cheng, and J.-Y. Yu. Advances in multimedia modeling: Travel photo and video summarization with cross-media correlation and mutual influence. *Book Series Lecture Notes in Computer Science*, pages 577–587, 2009.

[29] W.-T. Chu and C.-H. Lin. Automatic summarization of travel photos using near-duplication detection and feature filtering. In *Proc. of the* 17^{th} *ACM international conference on Multimedia*, pages 1129–1130. ACM, 2009.

[30] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y. Xu. Color harmonization. *Proc. ACM SIGGRAPH*, 25(3):624–630, 2006.

[31] M. Cooper et al. Temporal event clustering for digital photo collections. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(3):269–288, 2005.

[32] A. Crabtree, T. Rodden, and J. Mariani. Collaborating around collections: informing the continued development of photoware. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, pages 396–405, New York, NY, USA, 2004. ACM.

[33] J. Cui, F. Wen, and T. X. Real time google and live image search reranking. In *ACM Multimedia*. ACM, 2008.

[34] Z. Dai and Y. Wu. Where Are Focused Places of a Photo? *Lec. Notes. in Comp. Sci.*, 4781:73, 2007.

[35] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. *Lec. Notes. in Comp. Sci.*, 3953:288, 2006.

[36] R. Datta, J. Li, and J. Wang. Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition. *IEEE Intl. Conf. Image Proc.*, pages 105–108, 2008.

[37] A. Dewitz. Web-enabled print architectures. *Research Monorgraph of the Printing Industry Center at RIT*, PICRM-2008-06, 2008.

[38] M. P. Eckert and A. P. Bradley. Perceptual quality metrics applied to still image compression. *IEEE Signal Processing*, 70:177–200, 1998.

[39] M. FAIRCHILD. *Color Appearance Models.* Addison-Wesley, Reading, MA., 1998.

[40] E. Fedorovskaya, C. Neustaedter, and W. Hao. Image harmony for consumer images. In *IEEE International Conference on Image Processing, San Diego, California, USA*, 2008.

[41] E. Feisner. *Colour: how to use colour in art and design.* Laurence King Publishing, 2nd edition, 2006.

[42] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167181, 2004.

[43] R. Fergus, P. Perona, and Zisserman. A visual category filter for google images. In *ECCV*, 2004.

[44] R. Ferzli and L. J. Karam. A human visual system-based model for blur/sharpness perception. In *2nd Workshop on Video Processing and Quality Metrics*, 2006.

[45] R. T. Fielding. *Architectural styles and the design of network-based software architectures, Ph.D. dissertation.* University of California, Irvine, Irvine, California, 2000.

[46] M. Freeman. *The image.* William Collins Sons & Co Ltd, revised edition, 1990.

[47] D. Frohlich et al. Requirements for photoware. In *Proc. of the 2002 ACM conference on Computer supported cooperative work*, pages 166–175, 2002.

[48] U. Gargi. Consumer media capture: Time-based analysis and event clustering. *HP-Labs Tech Report*, 2003.

[49] A. Girgensohn, J. Adcock, M. Cooper, J. Foote, and L. Wilcox. Simplifying the management of large photo collections. In *In Proc. of INTERACT03, IOS*, pages 196–203. Press, 2003.

[50] A. Graham et al. Time as essence for photo browsing through personal digital libraries. In *Proc. of the second ACM/IEEE-CS Joint Conference on Digital libraries*, pages 326–335. ACM, 2002.

[51] D. Hall and J. Sturges. Temporal event analysis: Finding events in photo collections. *Research Disclosure Journal*, March:302 – 305, 2007.

[52] E. Hamilton. Jpeg file interchange format version 1.02. http://www.jpeg.org/public/jfif.pdf, 1992.

[53] S. Harrington, J. Naveda, R. Jones, P. Roetling, and N. Thakkar. Aesthetic measures for automated document layout. In *Symposium on document engineering*. ACM, 2004.

[54] D. Hasler et al. Measuring colorufulness in natural images. In *Electronic Imaging vol. 5007*. SPIE, 2003.

[55] D. Hasler and S. Susstrunk. Measuring colourfulness in natural images. *SPIE/IS&T Hum. Vis. Elec. Img.*, 5007:87–95, 2003.

[56] J. Hayes and L. Flower. *Identifying the organization of the writing process*. Lawrence Erlbaum associates, Hillsdale, New Jersey, 1980.

[57] N. V. House et al. From "what?" to "why?": The social uses of personal photos. In *CSCW*. ACM, 2004.

[58] W. H. Hsu, L. S. Kennedy, and Chang. Novel reranking methods for visual search. In *IEEE Multimedia*. IEEE, 2007.

[59] L. Itti et al. A model of saliency-based visual attention for rapid scene analysis. In *trans. on PAMI,vol.20, no.11*, pages 1254–59. IEEE, 2002.

[60] Y. Jing and S. Baluja. Pagerank for product image search. In *WWW*, 2008.

[61] M. Jones et al. "narrowcast yourself": designing for community storytelling in a rural indian context. In *Proc. of the 7th ACM conference on Designing interactive systems*, pages 369–378. ACM, 2008.

[62] M. J. Jones and P. Viola. Face recognition using boosted local features. *Tech. Report MERL TR-2003-25. Mitsubishi Electric Research Lab.*, 2003.

[63] D. Joshi, J. Z. Wang, and J. Li. The story picturing engine—a system for automatic text illustration. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):68–89, 2006.

[64] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, volume 6, pages 419–426, 2006.

[65] L. Kennedy, S.-F. Chang, and K. I. To search or to label?: predicting the performance of search-based automatic image classifiers. In *8th ACM Intl. workshop on Multimedia information retrieval*. ACM, 2006.

[66] L. Kennedy et al. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM Multimedia*. ACM, 2007.

[67] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. of the 17th international conference on World Wide Web*, pages 297–306. ACM, 2008.

[68] D. Kirk et al. Understanding photowork. In *CHI '06: Proc. of the SIGCHI conference on Human Factors in computing systems*, pages 761–770. ACM, 2006.

[69] P. Kisilev et al. Noise and signal activity maps for better imaging algorithms. In *ICIP*. IEEE, 2007.

[70] G. Klinker, S. Shafer, and T. Kanade. The measurement of highlights in color images. *Intern. J. Comput. Vision*, 2(1):7–32, 1988.

[71] A. Kuchinsky et al. Fotofile: a consumer multimedia organization and retrieval system. In *SIGCHI*. ACM, 1999.

[72] B. M. Landry and M. Guzdial. itell: supporting retrospective storytelling with digital photos. In *DIS '06: Proc. of the 6th conference on Designing Interactive systems*, pages 160–168, New York, NY, USA, 2006. ACM.

[73] C. Li and T. Chen. Aesthetic Visual Quality Assessment of Paintings. *IEEE Jnl. Sel. Top. Sig. Proc.*, 3(2):236–252, 2009.

[74] C. Li, A. C. Loui, and T. Chen. Towards aesthetics: a photo quality assessment and photo selection system. In *Proceedings of the international conference on Multimedia*, MM '10, pages 827–830, New York, NY, USA, 2010. ACM.

[75] H. Li and K. Ngan. Unsupervised segmentation of defocused video based on matting model. In *ICIP*. IEEE, 2006.

[76] S. H. Lim. Characterization of noise in digital photographs for image processing. In *SPIE vol. 6069*. SPIE, 2006.

[77] W. Lin, A. Hauptmann, and R. Jin. Web image retrieval re-ranking with a relevance model. In *IEEE/WIC*. IEEE, 2003.

[78] D. Lodge. *The art of fiction.* Secker & Warburg, London, U.K., 1992.

[79] A. Loui et al. Multidimensional image value assessment and rating for automated albuming and retrieval. In *IEEE Intl. Conf. Image Proc.*, pages 97–100, 2008.

[80] A. C. Loui and A. E. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Transactions on Multimedia*, pages 390–402, September 2003.

[81] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):910–110, 2004.

[82] J. Luo, A. E. Savakis, S. P. Etz, and A. Singhal. On the application of bayes networks to semantic understanding of consumer photographs. In *Proc. of ICIP*. IEEE, 2000.

[83] Y. Luo and X. Tang. Photo and Video Quality Evaluation: Focusing on the Subject. In *Proc. of the 10th European Conf. on Computer Vision: Part III*, page 399. Springer-Verlag, 2008.

[84] J. Martinet, Y. Chiaramella, and P. Mulhem. A model for weighting image objects in home photographs. In *CIKM'05*, pages 760–767. ACM, 2005.

[85] P. Messaris. *Visual persuasion: the role of images in advertising.* Sage Publications Inc., 1997.

[86] A. K. Moorthy and A. C. Bovik. Visual importance pooling for image quality assessment. *IEEE Jnl. Sel. Top. Sig. Proc.*, 3(2):193–201, April 2009.

[87] M. Naaman et al. Automatic organization for digital photographs with geographic coordinates. In *JCDL '04: Proc. of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 53–62. ACM, 2004.

[88] K. Nasrollahi and T. Moeslund. Face quality assessment system in videosequences. In *Workshop on Biometrics and Identity Management*, 2008.

[89] P. Obrador. All the publications by the author are listed in appendix D.

[90] T. N. Pappas and R. J. Safranek. Perceptual criteria for image quality evaluation. *Handbook of Image and Video Proc.*, 2000.

[91] B. Park and G. Desouza. Analysis of clipping effect in color images captured by ccd cameras. In *Sensors*, pages 292–295 vol 1. IEEE, 2004.

[92] E. Peli. Contrast in complex images. *J. Opt. Soc. Am.*, A/Vol. 7(No. 10/October):2032–2040, 1990.

[93] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. *Vis. Comm. and Imag., SPIE*, 5150:573582, 2003.

[94] J. C. Platt. Autoalbum: Clustering digital photographs using probabilistic model merging. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 96–100, 2000.

[95] J. C. Platt, M. Czerwinski, and B. A. Field. Phototoc: Automatic clustering for browsing personal photographs. *Microsoft Tech. Report*, 2002.

[96] P. Rice. *Professional Techniques for Black & White Digital Photography.* Amherst Media, Inc., 2005.

[97] I. Richardson. *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia.* John Wiley & Sons Inc, 2003.

[98] K. Rodden. *Evaluating similarity-based visualisations as interfaces for image browsing, Ph.D. dissertation.* Univeristy of Cambridge, Cambridge, U.K., 2002.

[99] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *CHI '03: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 409–416. ACM, 2003.

[100] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *WWW*, 2009.

[101] A. Savakis, S. Etz, and A. Loui. Evaluation of image appeal in consumer photography. In *Proc. of SPIE*, pages 111–121, 2000.

[102] L. Schäfer, C. Valle, and W. Prinz. Group storytelling for team awareness and entertainment. In *NordiCHI '04: Proc. of the third Nordic conference on Human-computer interaction*, pages 441–444. ACM, 2004.

[103] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[104] C. Shen et al. Sharing and building digital group histories. In *CSCW '02: Proc. of the 2002 conference on Computer supported cooperative work*, pages 324–333. ACM, 2002.

[105] E. Y.-T. Shen, H. Lieberman, and G. Davenport. What's next?: emergent storytelling from video collection. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 809–818, New York, NY, USA, 2009. ACM.

[106] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. *11th IEEE international Conference on Computer Vision*, pages 147–155, 2007.

[107] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349 – 1380, 2000.

[108] J. R. Smith and S.-F. Chang. Tools and techniques for color image retrieval. In *SPIE, Electronic Imaging, Storage & Retrieval for Image and Video Databases IV*, pages 426–437. IS&T/SPIE, 1996.

[109] J. R. Smith et al. Interactive search fusion methods for video database retrieval. In *ICIP*. IEEE, 2003.

[110] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[111] R. Thornhill and S. W. Gangestad. Facial attractiveness. *Trends in Cognitive Science*, pages 3:452–460, 1999.

[112] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998.

[113] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. *Lec. Notes. in Comp. Sci.*, pages 198–205, 2004.

[114] A. M. van Dijk, J. B. Martens, and A. B. Watson. Quality asessment of coded images using numerical category scaling. *SPIE Adv. Image Video Comm. Storage Tech.*, 2451:90–101, 1995.

[115] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 2000.

[116] L. R. Veloso et al. Smile detection combining pca and lbp neural network classifiers. submitted for journal publication. 2011.

[117] L. video quality assessment databases. http://live.ece.utexas.edu/research/quality/.

[118] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition*, pages 511–518. IEEE, 2001.

[119] Wang et al. Igroup: presenting web image search results in semantic clusters. In *SIGCHI*. ACM, 2007.

[120] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Tran. Image Proc.*, 13(4):600–612, 2004.

[121] Z. Wang, H. R. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of JPEG compressed images. *IEEE Intl. Conf. Image Proc.*, 1:477–480, 2002.

[122] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. *The Handbook of Video Databases: Design and Applications*, pages 1041–1078, 2003.

[123] B. Wayne. *http://www.businessinsider.com/is-youtube-doomed-2009-4*, 2009.

[124] X. wen Chen and T. Huang. Facial expression recognition: a clustering-based approach. *Pattern Recognition Letters*, 24:1295–1302, 2003.

[125] S. Whittaker, O. Bergman1, and P. Clough. Easy on that trigger dad: a study of long term family photo retrieval. *Journal Personal and Ubiquitous Computing*, 14,1:31–43, 2010.

[126] S. Winkler. Issues in vision modeling for perceptual video quality assessment. *IEEE Signal Processing*, 78:231–252, 1999.

[127] S. Winkler. Visual fidelity and perceived quality: Towards comprehensive metrics. In *Proc. of SPIE vol. 3959*. ACM, 2000.

[128] L.-K. Wong and K.-L. Low. Saliency-Enhanced Image Aesthetics Class Prediction. In *ICIP 2009, Cairo, Egypt*, 2009.

[129] J. Xiao and T. Zhang. Face bubble: Photo browsing with faces. In *Advanced Visual Interfaces*, 2008.

[130] W. Yan and M. Kankanhalli. Detection and removal of lighting & shaking artifacts in home videos. *Proc. ACM Conf. Mult.*, pages 107–116, 2002.

[131] L. Zhang et al. Enjoyphoto – a vertical image search engine for enjoying high-quality photos. In *ACM Multimedia*. ACM, 2006.