STUDIES ON PERSONALIZED HCI

XU MENGDI

NATIONAL UNIVERSITY OF SINGAPORE

2013

STUDIES ON PERSONALIZED HCI

XU MENGDI

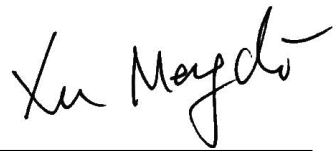
(B.Eng. (Electronic Engineering and Information Science), USTC)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE
2013

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



XU MENGDI
1 August 2013

Acknowledgments

I would like to express my gratitude to many people for the help they have given me throughout my Ph.D. study.

First and foremost, my thanks go to my supervisor, Associate Professor Yan Shuicheng, for his guidance and constant encouragement. He is always nice and encourages me through the difficult times. Without his patience and valuable suggestions, this dissertation would not have been finished.

I would like to extend my thanks to my senior Dr. Ni Bingbing for his constructive suggestions during the development of this dissertation.

I would also like to thank all the members in Learning and Vision Group. I had learned a lot from the discussions with them and enjoyed working with them.

My last thanks go to my family for always being by my side when I needed them and supporting me all these years without complaint.

Contents

Summary

In this work, four personalized HCI applications: Dynamic Captioning, Image Re-emotionalizing, Learning to Photograph, and Touch Saliency were studied to improve human experience when using computer, and touch mobile devices.

The aim of the first work was to propose a new scheme, dynamic captioning, to help hearing impaired audience understand video. Due to the loss of video information, people suffering from hearing impairment have difficulty understanding video content. Although captioning technology can help hearing impaired audience to a certain degree, the existing captioning techniques are far from satisfactory in assisting them to enjoy videos. To implement the proposed scheme, a rich set of multimedia technologies, such as face detection and recognition, text-speech alignment, visual saliency analysis, etc., were explored. Different from other methods which put scripts at the static position, the proposed method puts scripts at suitable position to help hearing impaired audience better identify the speaker. In addition, the scripts were highlighted word-by-word and the variation of voice volume was illustrated. In this way, the audience could better track the scripts and perceive the moods which are conveyed by the volume variation. This technology was implemented on 20 video clips and a user study with 60 real hearing impaired users was conducted. The user study results demonstrate the usefulness and effectiveness of the proposed video accessibility enhancement scheme.

The aim of the second work, image re-emotionalizing, was to synthesize user specified emotional affection onto arbitrary input images. Considering that the image affection generation process is subjective and complex, a learning-based framework which discovers emotion-related knowledge from a set of emotion annotated images was proposed. For each emotion-specific scene subgroup, emotion-specific generative models were constructed using color features of the image superpixels. Then, a piece-wise linear

transformation was conducted by aligning the feature distribution of the given image to the constructed statistical model of the given emotion-specific scene subgroup. Finally, a Bayesian framework was developed by further incorporation of a prior term which enforces the spatial smoothness and edge preservation for the derived transformation, and maximum a posteriori (MAP) was sought using standard nonlinear optimization method. User study results demonstrate that the proposed framework can yield effective and natural effects.

In the third work, learning to photograph, an intelligent photography system which recommends the most user-favored view rectangle for arbitrary camera input was presented. Due to the subjectivity of human's aesthetics judgment and large variations of image contents, automating the view recommendation process is difficult. The proposed framework discovers the underlying aesthetic photographic compositional structures from a large set of user-favored online sharing photos and utilizes this implicitly shared knowledge among the professional photographers for aesthetically optimal view recommendation. In particular, an *Omni-Range Context* method which explicitly encodes the spatial and geometric distributions of various visual elements in the photo as well as co-occurrence characteristics of visual element pairs by using generative mixture models was proposed. Searching the optimal view rectangle was then formulated as maximum a posteriori (MAP) by imposing the trained prior distributions along with additional photographic constraints. Comprehensive user studies well demonstrate the effectiveness of the proposed framework for aesthetically optimal view recommendation.

In the last work, touch saliency, a new touch saliency concept was proposed to study the relationship between human touch behavior and image saliency. The touch data was collected and touch fixation maps were generated for the images. The comparison study demonstrates that touch saliency map is highly correlated with human visual saliency map for the same stimuli. However, compared to the latter, the touch data collection process is much more flexible and requires no cooperation from users. Moreover, a saliency prediction framework which considers segment information, namely middle-level category features, under the recently proposed Multi-task Sparsity Pursuit (MTSP) saliency prediction model was studied. This touch saliency study opens up a new research direction of saliency study by harnessing human touch information on popular multi-touch smart mobile devices.

List of Tables

List of Figures

Chapter 1

Introduction

Human-computer interaction (HCI) studies the interaction between user and computers. It is a combination of computer science, user behavior study and other multimedia studies which could also extend to interaction between user and other devices, such as mobile phone, iPod, etc.

Human-tool interaction study has a long history. Applying science and technology to improve work efficiency became prevailing about a century ago. Workers were trained to operate machines, and researches aimed to reduce training time and eliminate errors were conducted [1]. With the popularity of computer, HCI emerged in the early 1980s and expanded rapidly for three decades, incorporating diverse concepts and approaches. In the last decade, due to the rapid development of portable techniques, a lot of new HCI researches have been conducted and more universal interfaces and equipment for disabled and aged people have been designed. The interaction between portable devices and users attracts researchers' attention. These portable devices include portable computers, portable phones and cameras, etc. [2].

The aim of personalization is to provide special contents for individuals based on user's implicit behavior and preferences. Different from the general case, personalized HCI could provide better services for computer or mobile device users by studying specific interaction between computer and user. Personalized HCI has been well applied in different forms of media, such as computer [3–5], mobile devices [6], TV [7–9], etc. Popularized social network websites, such as Facebook, Twitter, etc., can use personal

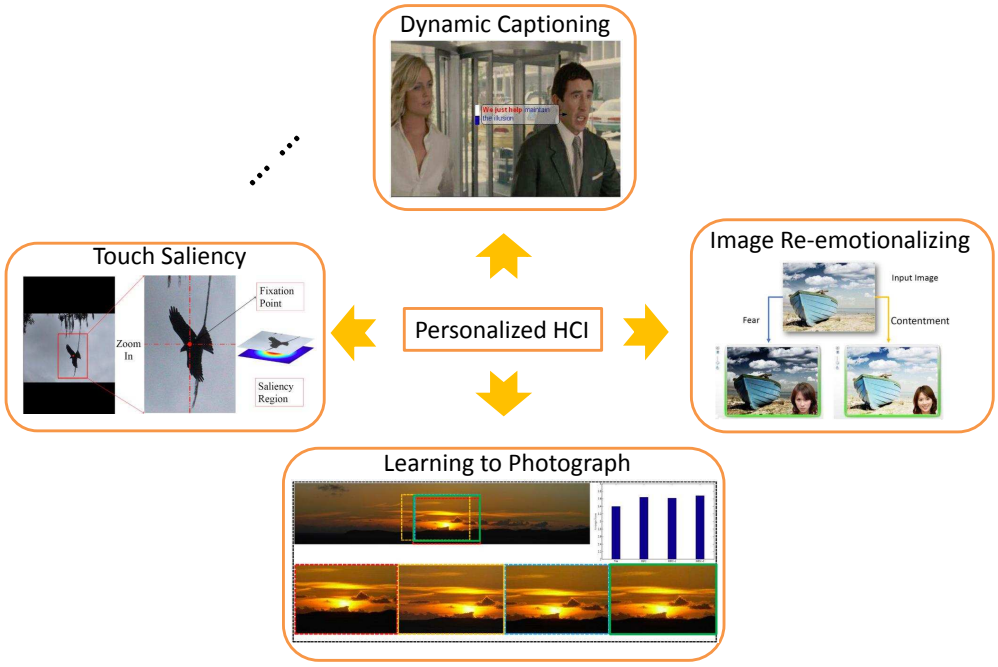


Figure 1.1: This dissertation discusses four personalized HCI applications: dynamic captioning, image re-emotionalizing, learning to photograph and touch saliency.

data to provide better services. Usually, to solve the personalization problem, researchers first collect data from websites and other resources to get the global information, then provide contents and services to an individual by considering both the global information and user specific information.

This dissertation studies four personalized HCI applications, *i.e.* dynamic captioning, image re-emotionalizing, learning to photograph and touch saliency, which can be used to improve user experience. Figure 1.1 shows the examples of the discussed personalized HCI applications. Dynamic captioning puts the caption beside the face of the speaker and highlights the subtitles according to the speech, which is more specific than statistic caption; image re-emotionalizing synthesizes image with specific user input emotion; learning to photograph aims to help amateur photographers produce professional photographs; touch saliency refers to as the saliency which collected by touch behavior, which could be used in personalized advertisement recommendation by collecting touch data for individual user.

In the rest of this chapter, a brief overview of the proposed applications, *i.e.* dynamic captioning, image re-emotionalizing, learning to photograph and touch saliency will be

introduced. More details of the previous and on-going researches on these systems will be discussed in Chapter 2.

1.1 Personalized HCI

1.1.1 Dynamic Captioning

Watching video is a common entertainment for normal people nowadays. However, for millions of people who are suffering from hearing impairment problem, it is hard to fully understand the video due to the loss of audio information. Typically there are two approaches to helping these audience better access the videos, they are "*direct access*" and "*assistive approach*". The former approach provides access as part of the previously developed system [10]. However, just providing access could not provide sufficient information for hearing impaired audience. Thus most attempts focus on the latter approach which provides extra information to help hearing impaired audience access the video. Captioning is the most widely-applied assistive technique. The traditional static captioning puts the subtitles at the bottom of the screen, which is not convenient for foreigners and people with hear-impairment problem to understand the video. Research [11] showed that caption can provide much information for hearing impairment audience. Thus it may be possible to solve the video understanding problem with captioning study.

Different from the static captioning, the proposed dynamic captioning puts the subtitle just near speaker's face and synchronously highlighting the subtitles during the playing of videos, thus might provides more information for hearing impairment audience.

1.1.2 Image Re-emotionalizing

Images may affect people into different emotions. For example, a photo taken in a rainy day looking at a dark street will usually give one a feeling of sadness; while a picture of a sunshine beach will mostly make people delighted. Affective image analysis has

been studied for decades; however, most studies focus on affective image retrieval or classification. Bianchi [12] proposed a system called K-DIME to retrieve emotional images. And recently, Machajdik *et al.* [13] proposed an International Affective Picture System (IAPS) to classify affective images. Based on the large number of affective image studies, it may be possible to apply affective image study to other multimedia work, such as social networks.

The proposed re-emotionalizing system, which synthesizes image with user specific emotion, is an exemplar combination. It might improve user experience during online chatting, online video, etc.

1.1.3 Learning to Photograph

Shooting a photograph has never been as easy as nowadays. Most of the cameras are built with much functionality, such as automatic focus; face detection etc., to assist the amateur photographers. In order to develop more *intelligent* functionalities for digital cameras, among which automatically finding an aesthetic view rectangle from the wide camera input view has recently begun to attract research attention [14, 15]. Existing study [15] considers using well-grounded composition guidelines, including *Rule of Thirds*, *Visual Balance*, *Diagonal Dominance*, *Size of Region*, etc., to evaluate the composition aesthetics. However, these rules may not be useful when the scene is too complex. Besides, the judgment of photo aesthetics is subjective and involves sentiments and personal taste [16].

Considering that the existing photo sharing websites provide millions of sharing photos, a large portion of which are taken by professional photographers and rated by users, we propose a learning based framework for photo composition modeling based on omni-range context to search the optimal view rectangle. The proposed scheme may help amateur photographers produce professional photos.

1.1.4 Touch Saliency

Visual saliency, which refers to the preferential fixation on the meaningful region in a scene, has been extensively studied in vision and psychology literature (*e.g.*, [17–19]),

due to its various applications in image segmentation [20], multimedia [21,22] and image retargeting [23, 24], etc. Several fixation datasets [20, 25, 26] have been built to evaluate the saliency prediction models. Usually, these datasets are collected by eye trackers. During visual data collection process, human subjects are asked to look at the stimuli image and eye fixation data would be recorded. The eye trackers are always complicated for operation and the data collection process usually makes human subjects tired. Thus it is necessary to find an alternative way to collect attention data. The widely used mobile touch devices show the possibility of collecting large fixation dataset through touch instead of just look.

In the touch saliency study, touch fixations are collected using touch devices and touch fixation maps are generated. The touch interest region is called *touch saliency* to distinguish it from visual saliency. The personalized touch behavior may be recorded to provide further services, such as product recommendation and automatic image enlarging.

1.2 Thesis Focus and Contributions

This thesis focuses on four personalized HCI works: dynamic captioning, image re-emotionalizing, learning to photograph, and touch saliency. The main aim of these works is to improve user experience for computer or mobile device users. The specific contributions of the discussed works are summarized as follows:

1) **Dynamic Captioning:** A video accessibility enhancement scheme for hearing impaired audience is proposed. To the best of our knowledge, this is the first integrated solution to facilitate hearing impaired users in video access. The proposed scheme involves the combination of a variety of technologies as well as novel methodologies. For example, the script-face mapping is an important topic per se and the proposed algorithm can be applied to many other applications. An in-depth user study is conducted to compare different captioning paradigms with real hearing impaired audience. Several conclusions and analysis also shed light on further research in this direction.

2) **Image Re-emotionalizing:** A novel learning based system by applying a brand new color transfer method to synthesize image based on user specific input emotion is developed. Moreover, an emotion-specific image dataset is constructed by collecting Internet images and annotating them with emotion tags by Amazon's Mechanic Turk [27]. The user study results and conclusions of image re-emotionalizing may show that it is possible to transfer image emotion by using multimedia approaches. The analysis could be useful for further study on this research area.

3) **Learning to Photograph:** A learning based framework for photo composition modeling based on omni-range context is developed. Moreover, a probabilistic inference framework for aesthetically optimal view rectangle recommendation is proposed. The scheme is implemented by combination of the mined omni-range context prior and other photographic constraints. Also, a large image dataset of professional landscape photos is constructed. The constructed dataset can be further made publicly available for encouraging this research direction. Comprehensive user studies well demonstrate the effectiveness of the proposed framework for aesthetically optimal view recommendation.

4) **Touch Saliency:** First an iPhone interface is built to collect human touch fixation points and generate a fixation map. Then the differences between visual saliency and touch saliency are analyzed by comparing the performances of some state-of-the-art methods. Further discussions show that it is efficient to get human attention information with the help of mobile devices. After that, middle-level category features which represent the segment information is proposed. The proposed middle-level category feature could be utilized into the Multi-task Sparse Pursuit (MTSP) saliency detection model. Evaluation results show that the proposed model outperforms other state-of-the-art unsupervised models.

1.3 Organization of the Thesis

The organization of this thesis is as follows. Chapter 2 reviews the previous works on these four subtopics. The technical details of dynamic captioning, image re-emotionalizing,

learning to photograph and touch saliency are introduced in Chapter 3, 4, 5 and 6, respectively. The conclusions and future work will be shown in Chapter 7.

Chapter 2

Literature Review

This chapter summarizes the research works on personalized Human-Computer Interaction (HCI) and four specific studies: dynamic captioning, image re-emotionalizing, learning to photograph, and touch saliency.

2.1 Personalize User Experience

Personalized HCI can be used in many multimedia related research areas to collect personal information, generate personalized models and provide personalized services. Many personalized tasks and problems have been investigated due to their potential commercial value. Early in 1990s, researchers started to study the personalized web experience problem. Rob *et al.* proposed a Web Browser Intelligence (WBI) system to personalize user's web experience by joining user's personal information with global information [3]. There are also researchers who focus on personalizing web experience for mobile users [6]. Recently, with the increasing online surfing and shopping, personalized recommendation systems have been widely studied. In 2003, Amazon proposed a real-time recommendation algorithm to personalize the store for each customer [4]. Andriy Shepitsen *et al.* studied personalized recommendation in social tagging systems in 2008 [5]. Websites like Facebook and Google provide better service to customers by using account information. This dissertation covers most aspects of personalized human

computer interaction which related to the functionalities of audio (dynamic captioning), visual (image re-emotionalizing and learning to photograph), and touch (touch saliency).

2.2 Dynamic Captioning

Hearing impairment refers to conditions in which individuals are fully or partially unable to detect or perceive at least some frequencies of sounds. Efforts on accommodating hearing impaired people in accessing videos can be traced back to 1970s when closed captioning was demonstrated at the First National Conference on Television in Nashville, Tennessee [28]. For television, captions are encoded into Line 21 of the vertical blanking interval in NTSC programming, while teletext (a television information retrieval service developed in the United Kingdom in the early 1970s and closed caption can be transmitted in the teletext signal) is used in captioning transmit and storage in Phase Alternate Line and Sequential Color with Memory. For movie, probably the best-known closed caption in theatres is the Rear Window Captioning System from the National Center for Accessible Media. Other captioning technologies for movie include hand-held displays similar to Personal Digital Assistant, eyeglasses fitted with a prism over one lens and projected bitmap captions. More recently, efforts have also been made to build accessibility features for digital cinemas.

Despite many captioning standards and technologies have been made, the analysis of the impact of captioning on hearing impaired audience is fairly scarce. The earliest study on investigating caption perception of hearing impaired audience shows that adjusting captions to suitable linguistic level and reading rate is able to significantly improve the information gain from captions [29]. Braveman and Hertzog [30] analyzed the language level but not the rate of captioning that affect deaf users' comprehension. Jelinek *et al.* [31] investigated the difference of video caption comprehension between hearing impaired and normal students. Garrison *et al.* [32] studied how working memory affected the language comprehension of deaf students. Gulliver and Ghinea [11, 33] investigated the impact of captions on the perception of video clips for hearing impaired audience. They concluded that much information can be gained from caption, but the information from other sources such as visual content and video text will be significantly reduced, *i.e.*, the caption has no significant effect on the average level of assimilated information

across all sources. This indicates that it is not easy for the special audience to track, perceive and learn from the caption efficiently.

Therefore, the existing captioning technology is still far from satisfactory in assisting hearing impaired audience. Recently, the closed captioning on YouTube is a meaningful exploration towards helping hear impaired users access web videos. There also exists software, such as Captioneer[1], that is able to support manual editing of captions or even add several attractive effects. However, they still cannot fully address the aforementioned problems and manual editing is also not an ideal solution due to the high labor cost. In this work, an automatic approach to intelligently present caption is investigated. It puts scripts in suitable regions, aligns them with speech and also illustrates the variation of voice volume. The user study with hearing impaired audience has demonstrated the effectiveness of this approach.

2.3 Image Re-emotionalizing

Several works have been done for image color transformation [34–37]. In [36], Reinhard *et al.* presented a system that transfers color by example via aligning the mean and standard deviation of the color channels in both input and reference images. However, user input is required to perform the preferred color transformation. Other works focused on non-photorealistic rendering (*i.e.*, image stylization) which communicates the main context of an image and explores the rendering effect of the scene with the artistic styles, such as painting [38, 39], cartoon [40] etc. Typically, the target exemplar style image is selected manually [34].

The proposed work is distinctive with these works: first, most of the previous works focused on only color transformation without any semantic knowledge transfer, however, the proposed work directly synthesizes affective property onto arbitrary images, which is hardly investigated throughout literature; second, the proposed system is fully automatic which requires no human interactions, however, most of the previous methods require either users' manual selection of certain painting parameters [38] or users' specification of specific example images [34].

[1] http://www.tsstech.org/captioneer.html

2.4 Learning to Photograph

2.4.1 Image Re-targeting

One related research direction to the proposed work is image re-targeting. The goal of re-targeting is to downsize the input image by preserving the recognizability of important image features, for displaying images on small screens such as mobile phones.

Visual attention models (VA) are commonly utilized for localizing the region of interest [41, 42]. Methods in this category employ a nero-morphic model that simulates which elements of a visual scene are likely to attract the attention of human observers. Given an image or video sequence, the model computes a saliency map, which topographically encodes for conspicuity (or *saliency*) at every location in the visual input by convolving the image with a series of special filters and encoding the responses at each pixel location. Image regions with maximum saliency values are selected.

Setlur *et al.* [43] presented a non-photorealistic algorithm for re-targeting large images to small size displays so that important objects in the images are still recognizable. They segment an image into regions, identify important regions, resize the remaining image, and re-insert the important regions. Avidan *et al.* [44] proposed an image operator called seam carving that can change the size of an image by gracefully carving-out or inserting pixels in different parts of the image. By applying the operator in both directions the new image can be re-targeted to a new size. The seam-carving operator is extended to video re-targeting and media re-targeting [45,46]. Wolf *et al.* [47] also presented a video re-targeting solution based on solving a linear set of equations. An optimized scale-and-stretch approach was proposed by Wang *et al.* [48] for image resizing. In [49–51], the relative positions of objects are modified using patch based methods. Recently, Guo *et al.* [52] formulated image re-targeting as a mesh parameterization problem that aims to find a homomorphous mesh with the desired size of the target display. Their method also achieves emphasizing important objects while retaining the surrounding context.

The philosophy of the above mentioned image re-targeting methods, however, is distinctive with the proposed task. Instead of recommending an aesthetically optimal view rectangle, these methods only concern whether the obtained sub-images can preserve the information to maximum extent, which might not be related to photo aesthetics.

This experiment compares the view rectangle obtained by visual attention model (VA) based method and the proposed method.

2.4.2 Image Quality Assessment

The inverse problem to the studied task is known as *image quality assessment*. Both low-level and high-level image features are adopted by computer vision researchers for the purpose of image quality assessment. Luo *et al.* [53] proposed a photo quality assessment method which first extracts the subject region from a photo, and then formulates a number of high level semantic features for photo quality classification. Ke *et al.* [54] designed a high level sematic feature to match the people's perception of photo quality. Information theory is also used for evaluating the photo quality. Sheikh *et al.* [55] provided an information fidelity criterion for image quality assessment by modeling the statistics of natural scenes. Visual attention model is widely used by perception researchers, *e.g.*, Sun *et al.* [42] presented a framework for both generalized and personalized photo assessment by integrating computational visual attention model with computational principles. Recently, Luo *et al.* [56] proposed a content-based photo quality assessment method using regional and global features. Marchesotti *et al.* [57] proposed to use generic image descriptors to assess aesthetic quality.

2.4.3 Learning from Web

Learning from web-based resources (*e.g.*, images/videos) for multimedia applications has been an emerging research direction for the society. Zheng *et al.* [58] successfully leveraged the vast amount of multimedia data on the web to build a world-scale landmark recognition engine which organizes, models and recognizes the landmarks on the scale of the entire planet Earth. Ji *et al.* [59] reported a famous city landmarks discovery and personalized tourist suggestion system by mining the images automatically crawled from online sharing personal blogs. Hao *et al.* [60] proposed to mine location-representative knowledge from a large collection of travelogues, they proposed a probabilistic topic model which has applications on destination recommendation and travelogue summarization. Recently, Ni *et al.* [61, 62] presented a web image mining framework for training a universal face-pattern based age estimator.

An early version of this work has been published in ACM Multimedia 2009 [63]. In [63], spatial context is also modeled using generative Gaussian mixture models. However, geometric properties of visual elements and pairs are not explicitly considered. In this extended work, a comprehensive model is developed for encoding both spatial and geometric context of visual elements for mining professional photo compositional rules.

2.5 Touch Saliency

In order to study which region is interesting to human when they look at a scene, researchers create the human visual saliency map using eye fixation data. The eye fixation points are always collected by eye tracking devices. Many fixation datasets are presented ([20, 25, 26, 64–67]) for studying human visual saliency. Judd *et al.* [26] built a dataset of natural indoor and outdoor images. In this dataset, 15 participants are asked to freely explore 1,003 images. Meanwhile they learn a saliency model based on low, middle and high-level image features. Bruce dataset [25] is another widely used eye tracking dataset. It includes 120 color images of indoor and outdoor scenes. Different from other datasets, a recent NUSEF dataset [20] which contains images with several semantics (*i.e.* expressive face, nude, action, reptile, etc.) is constructed. In this dataset each image is viewed by about 25 persons.

The datasets mentioned above are constructed by collecting human fixation points using expensive eye tracking devices. This data collecting process requires tedious calibration and validation, and thus it is difficult to get a large dataset. In this work, a new fixation-point collection method which is based on touch devices instead of eye tracking devices is proposed. In the experiment, users are asked to freely view images on touch-screen mobile devices. Saliency information is collected from finger trajectories during zooming-in, zooming-out and other operations.

Many saliency prediction models [17, 18, 26, 68, 69] have been proposed for saliency study. Recently several saliency benchmarks [70, 71] are built to better evaluate the performances of different saliency prediction models. Ali *et al.* proposed a framework to measure the existing methods [70]. They analyzed the benchmark datasets and provided a quantitative comparison of the state-of-the-art saliency detection models. One difference between the saliency detection methods is the mechanism to measure saliency.

Some saliency prediction methods [17, 72] are based on neurobiological [73] or psychological findings. Typically, the feature maps are computed for the features, then normalized and combined to form the final saliency map [17]. Alternatively, some saliency detection methods formulate the visual attention function in a computational framework. Most recently, many recent saliency detection studies emphasize utilizing graph model [74], maximum information sampling [18], subspace analysis [75], frequency domain measurement [76] and global contrast [77].

Saliency detection performance highly depends on the choice of feature space. Since it is very hard to find a single feature that works well on various images, how to combine multiple features efficiently is an important problem [17, 78–80]. Recently, Lang *et al*. [81] proposed a sparsity pursuit framework for saliency detection and generalized this framework to integrate high-level priors. This work formulates the saliency detection process as a convex optimization problem. Given an input image, first partition it into non-overlapping patches of $p \times q$ pixels. These patches are used as basic elements here. If the patch is judged as salient, the pixels within this patch are salient. The higher the value is, the more salient the region is. Let $X_1, X_2, \cdots, X_K$ be K feature matrices for K types of features, and a multi-task generalization of Low rank representation (MTSP) seeks a sparse matrix E (which measures saliency) by solving the following convex optimization problem:

$$\begin{aligned} \min_{\substack{Z_1,\cdots,Z_K \\ E_1,\cdots,E_K}} \quad & \sum_{i=1}^{K} \|Z_i\|_* + \lambda \|E\|_{2,1} \\ \text{s.t.} \quad & X_i = X_i Z_i + E_i, i = 1, \cdots, K, \end{aligned} \tag{2.1}$$

where $E = [E_1; E_2; \cdots; E_K]$, and E_k denotes the k-th feature. The sparse matrix E measures saliency. Evaluation results show that MTSP manages to combine different features efficiently. Thus this model is used as the basic model to study the performance of middle-level category features in the second part of this work.

There are some studies on improving saliency detection results by adding object or face detection results. Cerf *et al.* [82] combined face detection and low-level features to improve the Itti model. Judd *et al.* [26] studied efficient low, mid and high-level features and trained an SVM model for the combined features. In their work, the face and object detection results are used as the high-level feature, and the mid-level feature is obtained from a horizon line detector. The proposed middle-level category features are different

from the previous methods since the SVM estimations of the segment instead of the face or object detection results is used. Besides, a more general object concept instead of the commonly used face or car is studied.

Chapter 3

Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment

3.1 Introduction

Video is an important information carrier that presents visual and audio content in live form. With rapid advances of capturing and storage devices, networks and compression techniques, videos are growing in an explosive rate and play an increasing important role in peoples' daily life. However, there are millions of people that are suffering from hearing impairment. They are fully or partially unable to perceive sound. It is estimated that there are more than 66 million people with hearing impairment, of which about 41% cannot hear any speech at all and 59% are able to hear only if words are shouted around their ears [83]. This disability brings them great difficulty in comprehending video content as audio information is lost.

There are two typical approaches to helping these special audience better access videos. The first is "*direct access*", which provides access as part of the previously developed system [10]. However, merely providing access is not sufficient for hearing impaired audience. Therefore, most attempts have focused on the second approach, *i.e.*,

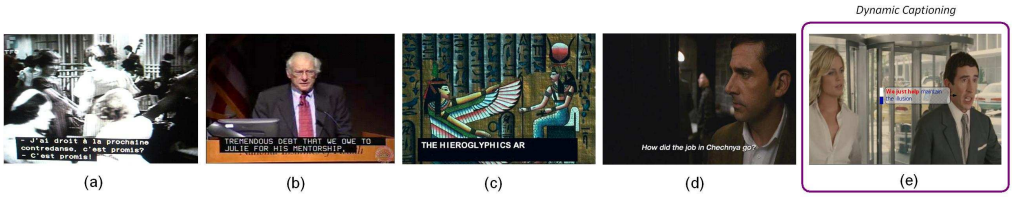


Figure 3.1: Examples of different captioning styles: (a) scroll-up captioning; (b) pop-up captioning; (c) pain-on captioning; (d) cinematic captioning; and (e) dynamic captioning. The first four techniques can be categorized as static captioning, and different from them, dynamic captioning in (e) benefits hearing impaired audience by presenting scripts in suitable regions, synchronously highlighting them word-by-word and illustrating the variation of voice volume.

the so-called assistive approach. For a large family of videos that have associated scripts[1] , such as movies, television programs and documentary, captioning is the most widely-applied assistive technique. By synchronously illustrating the scripts during the playing of videos, hearing impaired audience can obtain the necessary information from texts.

Generally, captioning can be categorized into open captions and closed captions according to whether it is able to be activated by users; and it can also be presented in a variety of styles (several examples can be found in Figure 3.1). However, the existing captioning methods most resemble each other as the scripts are simply demonstrated in a fixed region and they are illustrated statically. Although hearing impaired audience can get certain information from the scripts, they still encounter difficulty in the following aspects:

(1) Confusion on the speaking characters. When multiple characters are involved in a scene, hearing impaired audience need to judge from which person the scripts come, and this adds their difficulty of content understanding and also degrades their experience of video enjoyment.

(2) The tracking of captioning. In video playing, there is no hint on the duration of each piece of script. As speaking pace can vary significantly, the duration of the script presentation will also vary over a wide range. This brings hearing impaired

[1]These videos are also called *multimedia videos*. Although there are also many videos that have no script information, as mentioned in Section 3.4, the proposed scheme can be extended to deal with general videos by further exploring speech recognition and speaker identification technologies. Actually this work is just the primary step towards helping hearing impaired people better access video content.

audience difficulty in the tracking of scripts. For example, they may miss a part of a sentence when the character is speaking rapidly.

(3) The loss of volume information. The variation of volume conveys important information about the emotion [84, 85]. For example, the sound of a character will be loud if he/she becomes happy or angry. However, such information is lost in the existing captioning technology.

Therefore, the existing captioning approach is far from satisfactory in assisting hearing impaired audience. A recent study reporting that the conventional captioning approach can hardly add significant information for hearing impaired audience's perception [11, 33]. One major reason is that the audience can hardly track the scripts and match them with visual content rapidly.

In this work, a novel approach named *dynamic captioning* is proposed to enhance the accessibility of videos for hearing impairment. Compared with the existing captioning methods which are categorized as static captioning, dynamic captioning is able to help hearing impaired users match the scripts with the corresponding characters. Dynamic captioning is also able to synchronize the scripts word-by-word with the speech as well as highlight the variation of voice volume. In this way, the aforementioned three problems can be addressed. Figure 3.1(e) gives an example of the proposed dynamic captioning.

The dynamic captioning is accomplished by exploring a diverse set of technologies, including face detection and recognition, lip motion analysis, visual saliency analysis, etc. The scheme mainly contains three components: script location, script-speech alignment, and voice volume estimation. Script location determines the region in which scripts will be presented. It first performs a script-face matching to establish the speaking face for each piece of scripts (*i.e.*, establish the person from whom the scripts are coming) based on face detection and recognition techniques. It then selects a non-intrusive region around the face via visual saliency analysis in order to avoid the occlusion of important visual content. Script-speech alignment temporally matches each piece of script and the corresponding speech segment, and in this way the scripts can be highlighted word-by-word in synchrony with the speech. Voice volume estimation computes the magnitude of audio signal in a small local window, and visually demonstrates its variation within the scripts.

The main contributions of this work can be summarized as follows:

(1) A video accessibility enhancement scheme for hearing impaired audience is proposed. To the best of our knowledge, this is the first integrated solution to facilitate hearing impaired users in video access.

(2) The proposed scheme involves the combination of a variety of technologies as well as novel methodologies. For example, the script-face mapping is an important topic per se and the proposed algorithm can be applied to many other applications.

(3) An in-depth user study is conducted to compare different captioning paradigms with real hearing impaired audience. Several conclusions and analysis also shed light on further research in this direction.

The organization of the rest of this work is as follows. Section 3.2 introduces the system overview of video accessibility enhancement. Section 3.3 describes the components of the scheme in detail. Experimental results and user study are presented in Section 3.4. Finally, the conclusion of the proposed work is drawn in Section 3.5.

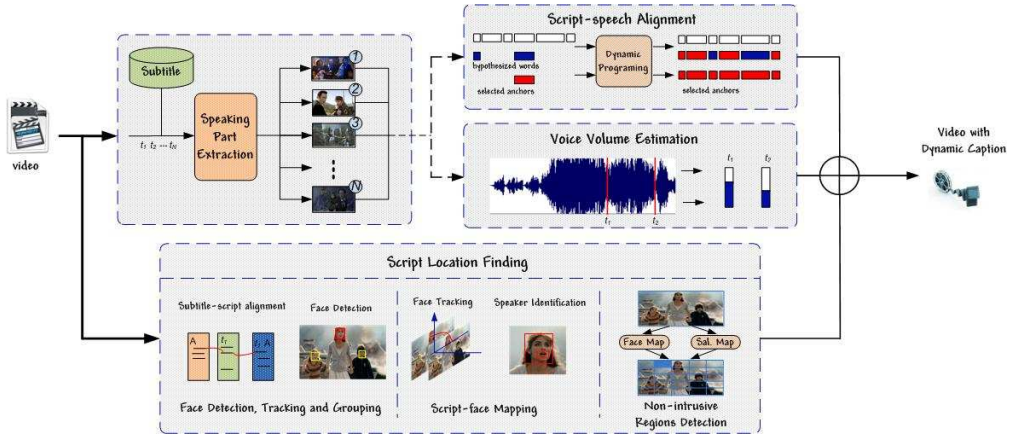


Figure 3.2: The schematic illustration of the accessibility enhancement.

3.2 Dynamic Captioning: System Overview

Figure 3.2 demonstrates the schematic illustration of the p video accessibility enhancement process. It mainly contains three components: script location, script-speech align-

ment and voice volume estimation. Given a video along with its script and subtitle file[2], speaking parts are firstly extracted according to the time information in subtitle. Then the character faces are mapped to the corresponding scripts with face detection and recognition techniques. After that, a non-intrusive region is detected around the face based on visual saliency analysis, in which the scripts are presented.

In parallel, the scripts are aligned with the audio track based on script-speech technology [86], and the starting and ending time of each word are recorded. Based on this information, the scripts are synchronously highlighted word-by-word along with the speech so that the hearing impaired audience can better track them. The voice volume estimation component estimates the local power of the audio signal, then visualize it near the scripts to help audience understand the emotion of the corresponding characters.

The proposed scheme thus generates a set of metadata in addition to the scripts, including the region information of each piece of script, the starting and ending time of each word and the voice volume information. In this work, an XML file is used to record these metadata. With these metadata, it is possible to develop an intelligent player to display videos with dynamic caption.

3.3 Dynamic Captioning: Technologies

This subsection introduces the three components in the proposed scheme in detail.

3.3.1 Script Location

This subsection describes the script location algorithm that finds suitable regions for presenting the scripts. As shown in Figure 3.2, it comprises three steps: (1) face detection, tracking and grouping; (2) script-face mapping; and (3) non-intrusive region detection.

[2]In some cases, "subtitle" may assume the audience can hear but cannot understand the language or accent but "caption" aims to describe to the hearing-impaired all significant audio content. In this study, the subtitle is restricted as the text that has time information and dialog. There also exists subtitle that contains richer content but it is not easily to acquire.

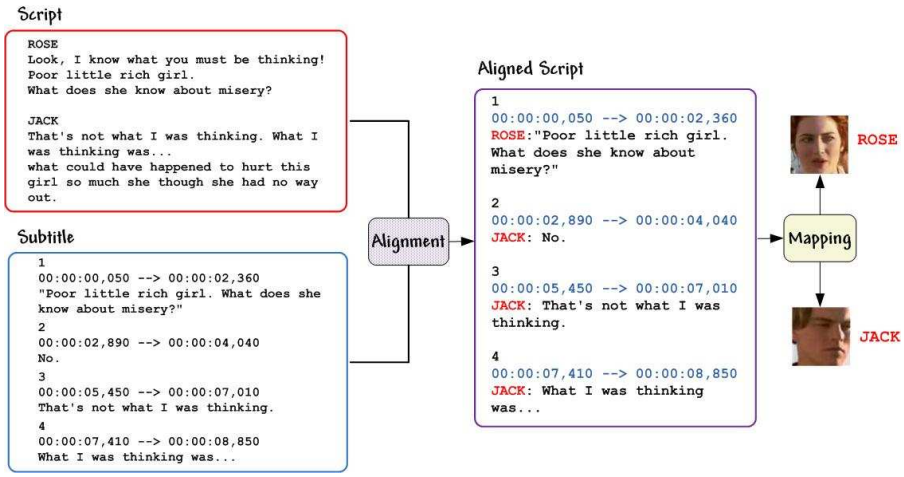


Figure 3.3: An example of the merge of subtitle and script files. The relationship between script, character identity and faces can be further established after script-face mapping.

Face Detection, Tracking and Grouping

In several cases, script file only contain speech content and speaker identity and there is another subtitle file that records the time information, as illustrated in Figure 3.3. Therefore, the speech content, speaker identity and time information from the subtitle and script need to be merged. Here a dynamic time warping method [87] is utilized to align subtitle and script. Figure 3.3 demonstrates an example of such alignment. Of course this step can be eliminated if there is only a script file encoding all the information.

Next a face detector is implemented to extract faces from the frames in the speaking parts. Here the face detection algorithm in [88] is adopted and several examples of detected faces can be found in Figure 3.4(a). As a video may contain thousands or even more detected faces, the continuously detected faces of a particular character are grouped as a face "track" with a robust foreground correspondence tracker [89]. The tracker mainly works as a follows. Given a pair of faces in adjacent frames, the size of overlapped area between the two bounding boxes of faces is estimated. If this value is greater than a given threshold, a matching is declared. This tracking procedure is also able to deal with the cases that faces are not continuously detected due to pose variation or expression change. In this way, the number can be significantly reduced (typically only hundreds of such tracks need to be dealt with). As a consequence, face track is adopted as the unit for labeling.

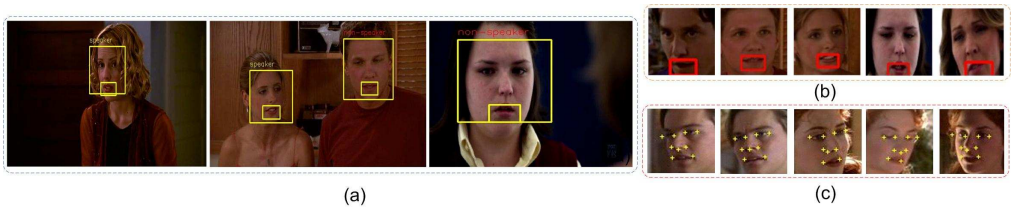


Figure 3.4: (a) and (b) illustrate the examples of detected faces and mouths, and (c) illustrates the facial feature points of several exemplary frames that are used in multi-task joint sparse face recognition.

Script-Face Mapping

This part introduces the script-face matching problem. The difficulty mainly lies on the following two facts: (1) in many cases there are more than one face within a frame (see the middle image in Figure 3.4(a)) and it is necessary to judge who is the speaker; and (2) even when there is only one face in the frame, he/she may not be the speaker and scripts come from another character (the third image in Figure 3.4(a) is an example and the girl in the frame is actually not speaking). To deal with these problems, a lip motion analysis [90] is adopted to establish whether the character is speaking when the frame contains only one face based on the fact that speaking is associated with distinctive lip movement.

The lip motion analysis is performed as follows. First a rectangular mouth region within each detected face region is detected using *Haar* feature based cascade mouth detector. Figure 3.4(b) illustrates several examples of mouth detection. Then the mean squared difference of the pixel values within the mouth region between each two continuous frames is computed. To keep translation invariance, the difference is calculated over a search region around the mouth region in the current frame and the minimal difference is used for decision. Two thresholds are set to establish three statuses, namely "speaking", "non-speaking" and "difficult to judge".

The following part considers the cases that a frame contains more than one face. The proposed approach is to first label faces with speaker identities and then match them with scripts accordingly (the script file contains the speaker identity information). Note that in cases that the frame only contains one face, the face could be labeled with speaker identity easily. Then the face tracks are labeled with identities based on such information. For example, if over half of the faces in a track are detected as speaking status and

the script shows that merely "EDWARD" is speaking in this period, then this track can be labeled as "EDWARD" with high confidence. The highly-confident labeled tracks are treated as training exemplars to predict other tracks that are unlabeled due to not containing enough established identities. Each unlabeled face track is simply represented as a set of history image feature vectors. One simple method for identification, as conducted in [87,91], is to directly calculate the feature distance between a testing face track and exemplar face tracks, and then assign testing face track to the nearest neighborhood. Another feasible method is to classify each history image independently via certain classification methods such as sparse representation based classification [92, 93], and then assign the face track to the class that achieves the highest frequency.

In this work, by regarding the identification of each history image in a testing face track as a task, The face track identification challenge can be formulated as a multi-task face recognition problem. This motivates us to apply the multi-task joint sparse representation model [94] to accomplish the task. The key advantage of multi-task learning is that it can efficiently make use of the complementary information embedded in different sub-tasks. The representation of face appearance is constructed by a part-based descriptor extracted around local facial features [87]. Here a generative model [95] is used to locate nine facial key-points in the detected face region, including the left and right corners of two eyes, the two nostrils and the tip of the nose and the left and right corners of the mouth. Figure 3.4 illustrates the detected key-points of several faces as examples. Then the 128-dim Sift descriptor is extracted from each key-point and concatenate them to form a 1152-dimensional face descriptor (SiftFD).

The employed face detection, tracking as well as speaker detection are able to offer a number of face tracks where the proposed identity is correct with high probability. For tracks which contain only a single identity, they can be treated as exemplars for labeling other tracks that contain no, or uncertain proposed identity. Each unlabeled face track is, nevertheless, simply represented as a set of history image vectors. For such history image in the track, the identification can be efficiently done via sparse representation classification [92].

The proposed multi-task joint sparse representation model works as follows. Suppose that there is a set of exemplar face tracks with M subjects. Denote $X = [X_1, \ldots, X_M]$ as the feature matrix in which the track $X_m \in \Re^{d \times p_m}$ is associated with the m-th subject consisting of p_m samples. Here d is the dimensionality of feature and $\sum_{m=1}^{M} p_m = p$ is

the total number of samples. Given a testing face as an ensemble of L history images $y^l \in \Re^d$, we consider a supervised L-task linear representation problem as follows:

$$y^l = \sum_{m=1}^{M} X_m w_m^l + \varepsilon^l, l = 1, ..., L \tag{3.1}$$

where $w_m^l \in \Re^{p_m}$ is a reconstruction coefficient vector associated with the $m-th$ subject, and ε^l is the residual term. Denote $w^l = [(w_1^l)^T, ..., (w_M^l)^T]^T$ as the representation coefficients for probe image feature y^l, and $w_m = [w_m^1, ..., w_m^l]$ as the representation coefficients from the m-th subject across different case images. For simplicity, W is denoted as $[w_m^l]_{m \times l}$. The proposed multi-task joint sparse representation model is formulated as the solution to the following multi-task least square regressions with $\ell_{1,2}$ mixed-norm regularization problem:

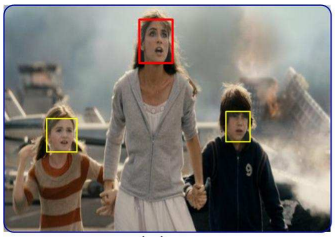
$$\min_W F(W) = \frac{1}{2} \sum_{l=1}^{L} \left\| y^l - \sum_{m=1}^{M} X_m w_m^l \right\|_2^2 + \lambda \sum_{m=1}^{M} \|w_m\|_2 \tag{3.2}$$

Here the Accelerated Proximal Gradient (APG) approach [96] is used to solve the optimization problem in Eqn. (3.2).

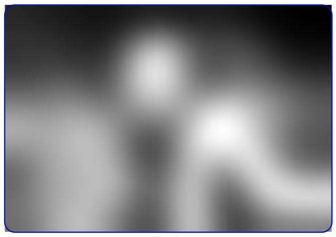
When the optimum $\hat{W} = [\hat{w}_m^l]_{m,l}$ is obtained, a testing image y^l can be approximated as $\hat{y}^l = X_m \hat{w}_m^l$. For classification, the decision is ruled in favor of the class with the lowest total reconstruction error accumulated over all the *L* tasks:

$$m^* = \arg\min_m \sum_{l=1}^{L} \left\| y^l - X_m \hat{w}_m^l \right\|_2^2 \tag{3.3}$$

After labeling each face track with speaker identity, it is possible to establish the speaking character even there are more than one face in a frame. Hitherto the mapping between scripts and faces is accomplished. It is worth mentioning that there also exist scripts that cannot be successfully mapped to faces, and in this work these scripts are directly displayed on the bottom of frames just like static captioning (off-screen voice is also processed in the same way).



(a)



(b)

0	1	0	1	0
0	0.8	0	0.8	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

(c)

Figure 3.5: An example of the saliency map and face weighting map for an image from the movie "2012". (a) is the original image; (b) illustrates the saliency map; and (c) shows the weighting map around the speaking face (although there are three faces in the frame, only the face indicated by the red box is speaking).

Non-intrusive Region Detection

Up to now, establishing the speaking of each piece of scripts is accomplished. As previously mentioned, the target of this work is to present the scripts near the speaking face such that hearing impaired audience can easily identify the character from whom the scripts come from. However, it is necessary to select a region that will not occlude important visual content and especially other faces. Therefore, a visual saliency analysis is performed to select the non-salient regions.

Given an Image I, the contrast of each pixel is an accumulated *Gaussian distance* between it and its neighbors:

$$c_{i,j} = \sum_{q \in \Theta} d(I_{i,j}, q) \tag{3.4}$$

where $I_{i,j}$ is the pixel position in I and Θ is the neighborhood of $I_{i,j}$. The contrasts $c_{i,j}$ thus form a saliency map [97, 98]. Figure 3.5(b) shows an example of the saliency map of the image in Figure 3.5(a) where the distance is measured in LUV color space. The brighter the pixel in the saliency map, the more important or salient it is.

For the detection of the non-intrusive regions, I is represented by a set of blocks $\mathcal{B} = \{b_i\}_{i=1}^{N_b}$ which are obtained by partitioning image I into $M \times M$ grids ($N_b = M^2$). Each grid corresponds to a block b_i and it gives a candidate region of caption insertion. For each block b_i, a saliency energy s_i $(0 \leq s_i \leq 1)$ is computed by averaging all the normalized energies of the pixels within b_i. As previously analyzed, the region should be selected around the speaking face. Therefore, a face weighting map $W = \{w_i\}_{i=1}^{N_b}$ is designed to weight the energy s_i, so that the caption will be restricted around the

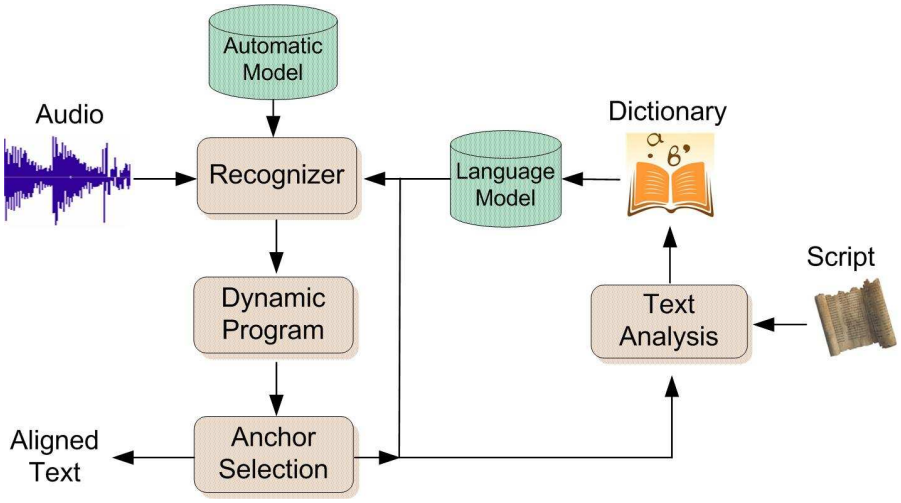


Figure 3.6: The schematic illustration of the script-speech alignment.

face. The face weighting map is generated by simply assigning the blocks around the speaker's face block constant weights and all other regions are assigned weight 0. More specifically, the weights of the left and right regions around the face region are set to 1, and the weights of the upper-left, bottom-left, upper-right and bottom right regions are set to 0.8. Figure 3.5(c) shows an example of the weighting map. Hence, the score for region selection is given by:

$$P(b_i) = w_i \times (1 - s_i) \tag{3.5}$$

The region with maximal score is finally established for caption insertion. In this work the parameter M is empirically set to 5, but it is also found that an adaptive setting of the parameter will be able to improve performance.

It is worth mentioning that although the non-intrusive region detection approach is effective, it cannot fully guarantee that informative visual content will not be occluded by caption. Thus in dynamic caption we choose to overlay the scripts with parent background such that audience can still recognize the content behind the caption.

3.3.2 Script-Speech Alignment

This subsection describes the script-speech alignment approach. As previously mentioned, based on this component the scripts can be synchronously highlighted word-by-

word and help impaired audience better track the scripts. Here we adopt a method based on recursive speech recognition with a shrinking dictionary and language model, which is analogous to the approach in [86]. Figure 3.6 demonstrates a schematic illustration of the proposed scheme. We use 39-dimensional MFCC features. The text analysis module processes the text file and the CMU pronouncing dictionary is used to translate each word into a phonetic sequence. For those words that are out of the dictionary, an automatic module introduced in [99] is introduced to process them. To reduce the computation cost, a simple bigram and trigram word model instead of a complete language model based on N-gram is built. Then SPHINX II [100], a speaker-independent speech recognition engine, is used to recognize the speech based on the previously generated language model and dictionary. When a complete hypothesis text string is produced for the whole audio stream, dynamic programming is employed to find the globally optimum alignment. The detailed process is as follows. We compare the scripts and the recognition results and the matched parts that contain more than N words are regarded as anchors. In this work N is set to 3 empirically.

Then the algorithm is iterated on each unmatched segment. In each iteration, the language model and dictionary are rebuilt to limit the list of active words and word sequence to those found in the script of this segment. This can speed up the recognition as only those words and their word pairs and triples that are available in the segment are searched. These steps are repeated on the unmatched segments until all the texts have been matched. The iteration also terminates if the recognizer is unable to find any additional words in the audio segment. The test on 20 video clips (the data are described in Section 3.4) shows that this approach is able to obtain accuracy, *i.e.*, the ratio of correctly aligned words, of above 90%.

3.3.3 Voice Volume Analysis

Existing studies reveal that the variation of voice volume conveys important information about human emotion [84, 101]. However, for hearing impaired audience, the volume information is fully lost. Therefore, in the proposed dynamic captioning scheme, the voice volume is symbolized and illustrated to help the special audience get more information.

The sound volume is estimated by computing the power of the audio signal in a small local window (the size of the window is set to $30ms$ in the proposed scheme).

After a normalization process, the estimated volume is visualized near the scripts with an ”indicator”. Figure 3.1(e) illustrates an example. The volume is indicated by the highlighted part of a strip and the size of the part will vary according to the estimated power.

3.4 Experiments

This subsection introduces experiments to evaluate the effectiveness and usefulness of the proposed scheme.

3.4.1 Evaluation of Script-Face Mapping

The experiments involve 20 clips from three movies, namely *”Titanic”*, *”Twilight”* and *”Up in the Air”*, and one teleplay, namely *”Friends”*. Table 3.1 presents the information about these clips.

Figure 3.7: Examples of the selected face tracks and exemplar faces. Five representative images for each track are presented in (a) and the selected exemplar faces with high confidence scores are illustrated in (b).

For script-face matching, a novel algorithm, namely multi-task joint sparse representation classification is proposed. We thus compare it against two existing methods: (1)

Table 3.1: The information about the video clips and the script-face mapping accuracy.

Movie Name	Clips	Frames	Face Tracks	Accuracy (%)		
				NN	SR	Our Algorithm
"Titanic"	C_1	2,864(1.99min)	22	73.33	72.26	**76.19**
	C_2	7,449(5.17min)	31	86.90	**90.90**	90.24
	C_3	2,868(1.99min)	18	80.89	87.47	**93.75**
	C_4	7,022(4.88min)	22	83.74	**91.51**	88.89
	C_5	9,801(6.80min)	43	**95.00**	**95.00**	87.50
"Twilight"	C_6	4,543(3.15min)	42	88.21	88.21	**89.29**
	C_7	8,193(5.69min)	47	**81.89**	80.60	81.45
	C_8	5,788(4.02min)	51	93.60	93.10	**95.89**
	C_9	6,317(1.95min)	47	75.30	81.90	**86.50**
	C_10	4,745(3.30min)	24	**96.15**	**96.15**	**96.15**
"Up in the Air"	C_11	9,707(6.74min)	22	**100**	**100**	**100**
	C_12	7,955(5.52min)	87	89.01	90.20	**92.52**
	C_13	3,852(2.67min)	31	91.7	**93.55**	92.98
	C_14	6,285(4.36min)	50	91.94	90.45	**92.20**
	C_15	6,533(4.54min)	44	92.75	93.33	**94.40**
"Friends"	C_16	4,549(3.16min)	38	74.26	74.07	**77.78**
	C_17	5,779(4.01min)	26	90.56	92.41	**100.0**
	C_18	3,621(2.51min)	39	78.08	78.61	**80.66**
	C_19	5,036(3.50min)	44	61.52	68.03	**71.90**
	C_20	4,748(3.30min)	31	74.07	86.41	**89.10**

nearest-neighbor (NN) classifier; and (2) the sparse representation (SR) classifier [92]. For each clip, the labeled exemplar faces with high confidence are used as the training set, and all the detected face tracks are regarded as the testing set. The parameter λ in Eqn. (3.2) is set to 0.1 throughout the experiment. Figure 3.7 shows several exemplar training faces and testing face tracks. The accuracies of script-face mapping achieved by the proposed algorithm and two existing methods are given in Table 3.1. It can be seen that the proposed algorithm outperforms the other two methods on 15 out of the 20 clips. Talbe 3.1 also shows that for most clips the recognition accuracy is above 80%. This is important for the proposed scheme, as putting scripts around an incorrect face will be misleading for hearing impaired audience. It can also be seen from Table 3.1 that the TV show gives better performance than movie. One possible reason is that the environment of TV show is well-controlled while that of the movie is wild.

3.4.2 User Study

There are 60 anonymous hearing impaired users participating in the study (21 male and 39 female). These participants come from Huangshan Branch, Anhui Special Education School, China. Their ages vary from 11 to 22. Most of them are pre-lingual deafness which means that they sustained hearing impairment prior to the acquisition of language and occur as a result of a congenital condition or through hearing loss in early infancy. Sign language is their first or preferred language. A small part of participants are post-lingual hearing impaired who occur as a result of disease, trauma or as a side-effect of medicine after the acquisition of language. In this study, two teachers from a deaf-mutes school helped us to communicate with the participants. Before the study, all the participants were required to carefully read the investigation questionnaire and made sure that they understood their roles in the experiment.

We compare the following three paradigms:

(1) No Caption (**NC**), *i.e.*, the hearing impaired participants were shown videos without caption.

(2) Static Caption (**SC**), *i.e.*, the hearing impaired participants were shown videos with static caption (here the cinematic captioning is adopted).

(3) Dynamic Caption (**DC**), *i.e.*, the hearing impaired participants were shown videos with dynamic caption.

All the participants are randomly divided into 3 groups (each group has 20 participants) to avoid the repeated playing of a video which will cause knowledge accumulation[3]. Therefore, each group merely evaluates one of the three paradigms for each video clip.

During the video playing process, participants were informed to stop and answer a number of questions which are related to the content of the movie clips after each showing. To sufficiently investigate the effectiveness of dynamic captioning, we first measure

[3]It is worth noting that it is not fair to let a user to directly compare two paradigms in this study, as the user will get accumulated knowledge if he/she watches the video for more than one time. Thus the participants are divided into groups, and fortunately there are fairly sufficient participants and statistical test (here one-way ANOVA test is adopted) can demonstrate the impact of users and the difference of paradigms.

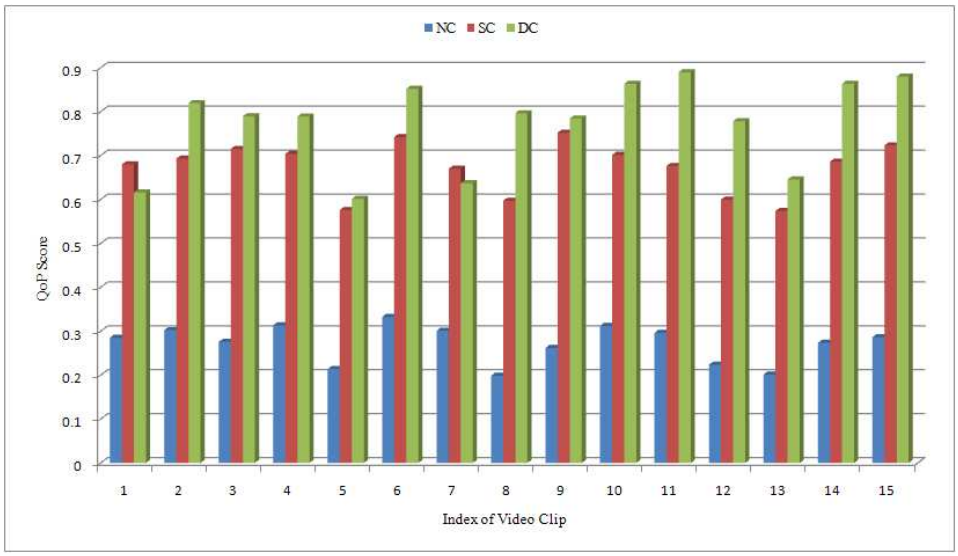


Figure 3.8: The QoP scores of: (1) no caption; (2) static caption; and (3) dynamic caption. The superiority of dynamic caption is very clear.

how much advantage the proposed scheme is able to gain on content comprehension and user impression, and then we further evaluate the components. Here content comprehension indicates the extent of understanding from the hearing impaired participants and user impression reflects whether the presentation of such dynamic caption is enjoyable and natural.

Evaluation of Full Scheme

1. Content Comprehension

As is known to all, some questions such as "how many characters are there in this movie clip" have a single definite answer. Thus it is possible to estimate the ratio of correctly answered questions for a pre-defined question set. In this study there are 50 questions for each movie clip. These questions are carefully designed to broadly cover the content in the video clip.

The questions can also be categorized according to the information source of their answers. For example, the question "who wore the sports clothes numbered 23?" can only be answered based on the video text information in the video, while "what's the name of hero" can merely be answered based on caption information. Therefore, it is

Table 3.2: The ANOVA test results on comparing DC and NC. The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

The factor of schemes		The factor of users	
F-statistic	p-value	F-statistic	p-value
86.75	2.47×10^{-11}	0.532	0.818

also possible to estimate the ratio of correctly answered questions that are related to different information sources. Thus the questions can be categorized as follows:

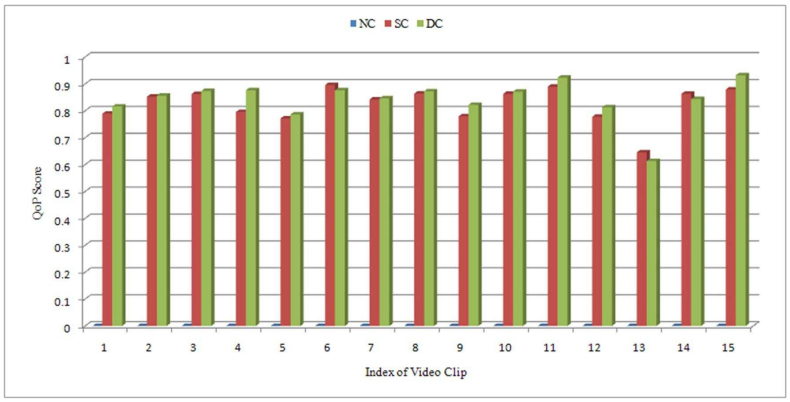
(1) Caption related: information from the captions only (34 questions in total).

(2) Video Text related: textual information contained in video but not in the caption (6 questions in total).

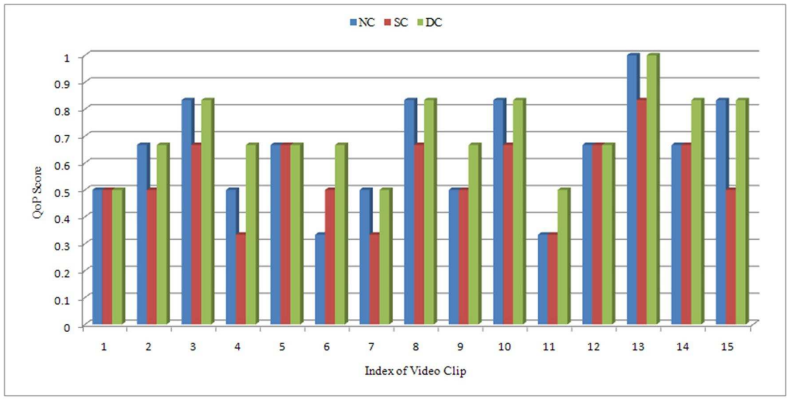
(3) Visual Content related: visual information contained in movie (10).

It can be seen that most questions (34 among 50) are related to caption. This is because caption is paramount to understanding the story of video. Hearing impaired participants were asked to answer the questions independently. The metric of Quality of Perception (QoP), which is defined as the ratio of the correctly answered questions in the full question set, is used for performance evaluation.

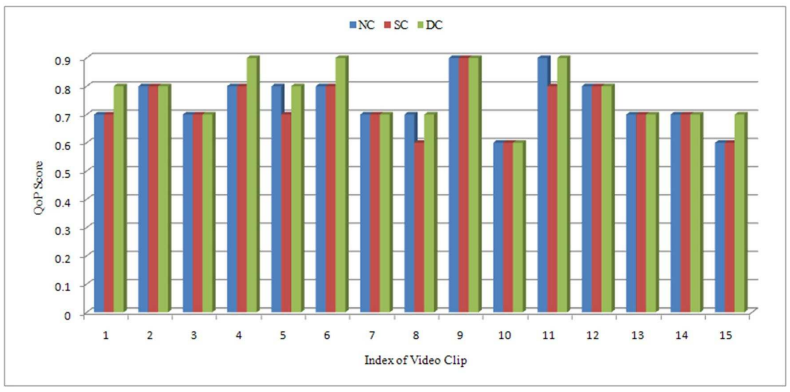
Figure 3.8 gives average QoP scores of each video clip (the scores are averaged over participants) with different captioning paradigms. Figure shows that both the static (SC) and the dynamic (DC) caption can greatly improve the comprehension level in comparison with the NC paradigm. It is worth noting that this does not contradict with the study in [11, 33] which reports that SC can hardly improve the information gain of hearing impaired audience, as in this study most questions are related to caption and thus audience cannot answer these questions without watching the captions. Next it can be seen that for most clips the DC paradigm outperforms SC. Only for clips C-1 and C-7 the DC paradigm performs slightly worse. This is mainly due to the relatively low script-face mapping accuracies (76.19% for clip C-1 and 81.45% for clip C-7). A one-way ANOVA test [102] is also performed, and the results are illustrated in Table 3.2 and Table 3.3. From the results it can be seen that the superiority of DC is statistically significant and the difference among users is statistically insignificant.



(a)



(b)



(c)

Figure 3.9: The QoP scores of: (1) no caption; (2) static caption; and (3) dynamic caption. Again, the superiority of dynamic caption is very clear. (a) (b) and (c) indicate different question sets. (a) caption relates questions; (b) video text related questions; and (c) visual content related questions.

Then the QoP scores for different question sets are estimated. Figure 3.9 illustrates the results. It can be seen that for questions that are related to caption (Figure 3.9(a)), the performance of NC is very poor as expected, and SC and DC are very close. This

Table 3.3: The ANOVA test results on comparing DC and SC. The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

The factor of schemes		The factor of users	
F-statistic	p-value	F-statistic	p-value
32.27	1.93×10^{-11}	0.23	0.971

indicates that the conversion from static to dynamic caption doesn't add much information. Figure 3.9(b) shows that the QoP scores of DC are remarkably higher than SC for the questions that are related to video text or visual content. The QoP scores of SC are even much worse than NC for the question related to video text. This indicates that the conventional captioning styles are rather distracting, and the hearing impaired participants need to focus on both the visual content and the caption at the bottom of frames. The proposed dynamic captioning scheme can be more easily glimpsed as the scripts are presented around the character faces.

Overall, it is clear that captioning is important for understanding the story in videos, but the conventional static captioning approach will degrade the information assimilation of hearing impaired audience from other sources, and the proposed dynamic captioning scheme can help them better perceive the content.

2. User Impression

For user impression, static captioning and dynamic captioning are compared using the following two criteria: *enjoyment* and *naturalness*.

- **Enjoyment**. It measures the extent to which users feel that the video is enjoyable.
- **Naturalness**. It measures whether the users feel the visual appearance of caption is natural.

In this test, there is no need to divide the users into groups, and thus each user was asked to assign a score of 1 to 10 (higher score indicates better experience) to the above two criteria. Figure 3.10 shows the results that are averaged over video clips and users. It can be seen that the dynamic caption remarkably outperforms static captioning in terms of enjoyment. However, the naturalness scores of the two captioning schemes are close.

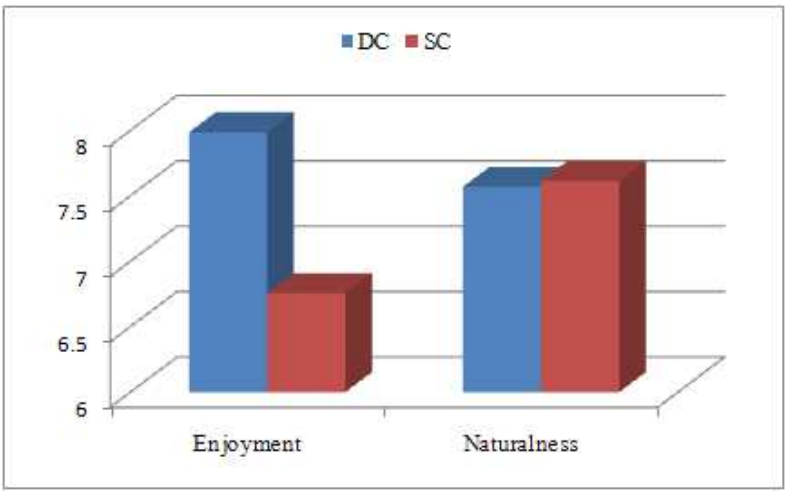


Figure 3.10: Study results of user impression. For the two criteria, namely *Enjoyment* and *Naturalness*, each user has been asked to assign a score between 1 and 10. Here we have demonstrated the scores averaged over users and video clips.

Via communicating with the audience, it is found that this is due to the fact that in several cases the regions of script presentation vary abruptly. One possible solution to address this problem is to smooth the variation of the regions for presenting the scripts.

3. Preference between Static and Dynamic Caption

Finally, each user is asked to choose between the static captioning and dynamic captioning that he/she prefers and wishes to use in the future considering all the above factors. The results show that 53 among the 60 users choose dynamic captioning. The remained 7 users choose static mainly because they have already been familiar with static captioning. This clearly demonstrates the usefulness of the proposed scheme.

Component Evaluation

This part further evaluates the components in the dynamic captioning scheme. We compare the following paradigms:

(1) Dynamic captioning (DC), *i.e.*, hearing impaired participants were shown videos with dynamic caption.

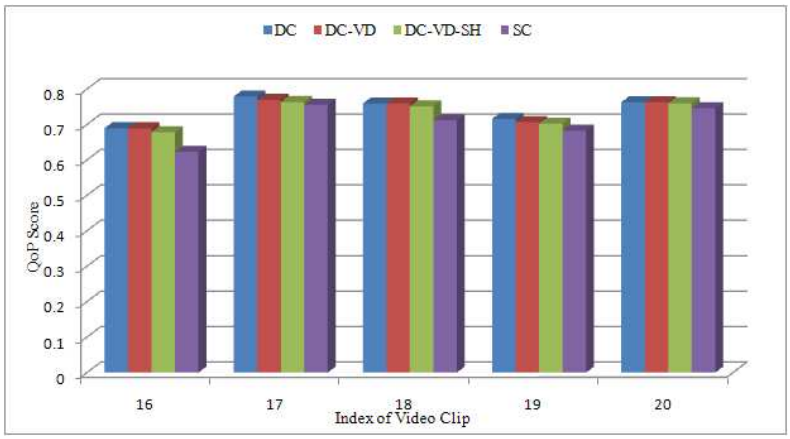


Figure 3.11: The comparison of QoP scores of: (1) dynamic captioning; (2) dynamic captioning without volume demonstration; (3) dynamic captioning without volume demonstration and script highligh; and (4) static captioning.

(2) DC without volume demonstration (DC-VD), *i.e.*, remove the voice volume demonstration from the dynamic captioning.

(3) DC without volume demonstration and synchronous highlight (DC-VD-SH), *i.e.*, remove both voice volume demonstration and script synchronous highlight from the dynamic captioning.

(4) Static captioning (SC), *i.e.*, hearing impaired participants were shown videos with static caption.

Analogous to the previous study, we test the content comprehension of the audience with the above paradigms. This study is conducted with the remaining 5 video clips (C-16 to C-20) and for each video there are also 50 questions. The 60 participants are randomly divided into four groups and then implement the question-answering test, and the process is the same with the previously introduced study. Figure 3.11 illustrates the average QoP scores for different clips. It can be seen that removing volume demonstration and synchronous highlight will reduce QoP scores, but DC-VD-SH is still able to outperform SC. This demonstrates the effectiveness of each component.

3.4.3 Discussion

In the experiments, it costs less than 4 minutes to process a video clip on average on a PC with Pentium 4 3.0G CPU and 2G memory. The average duration of the 20 video clips is 3.96 minutes, and this means that the processing time is roughly equivalent to the video duration (of course the processing time also depends on many factors such as the number of characters and the appearing frequency of dialogues). However, it is found that the cost can still be significantly reduced, such as by speeding up the solution process of Eqn. (3.2) and visual saliency analysis.

It is necessary to mention that this work mainly focus on the technical part of dynamic captioning and care less about user interface, such as the visualization of volume variation (currently we just use a very simple stripe with a highlighted part, see Figure 3.1(e)) and the style of script highlight. However, even with a simple interface, the proposed scheme has shown clear advantages through the study of user impression. User interface design is beyond the scope of this work although it is crucial for real-world application. We will leave it to the future work. Another problem worth mentioning is that inaccurate face-scripts mapping (though only few seen from the Table 3.1) will not confuse hearing impaired audience in this scheme since these scripts without accurate mapping will be displayed as static captions. This means that to some extent, static captioning can be viewed as the baseline of the proposed scheme.

Finally, we want to emphasize that although the focuses of the proposed scheme are on videos along with scripts, it can be extended to process general videos without scripts. Actually what we need is to employ speech recognition engine to convert speech to scripts, and use speaker clustering [103, 104] and identification [105, 106] to replace the face grouping and recognition techniques in the current scheme. Of course this task will be much more challenging, but it will be an important topic along this research direction.

3.5 Summary

This chapter describes a dynamic captioning scheme to enhance the accessibility of videos towards helping hearing impaired audience better enjoy videos. Different from

the existing static captioning methods, dynamic captioning put scripts at suitable positions to help hearing impaired audience better recognize the speakers. It also synchronously highlights the scripts by aligning them with the speech signal and illustrates the variation of voice volume to help hearing impaired audience better track and perceive scripts. Comprehensive user study with 60 real hearing impaired participants has demonstrated the effectiveness of the proposed scheme.

As this is the first work to our knowledge to help hearing impaired individuals better access videos, there is a lot of future work along this research direction. In future, it is possible to further improve the script-face mapping component to further boost the mapping accuracy and investigate the extension of the scheme to deal with videos without script will be investigated. A more comprehensive user study on a larger dataset could be conducted in future.

Chapter 4

Image Re-Emotionalizing

4.1 Introduction

Images may affect people into different emotions. For example, a photo taken in a rainy day looking at a dark street will usually give one a feeling of sadness; while a picture of a sunshine beach will mostly make people delighted.

Throughout the decade, the multimedia research community has shown great interest in affective retrieval and classification of visual signals (digital media). Bianchi-Berthouze [12] proposed an early Emotional Semantic Image Retrieval (*i.e.*, ESIR) system known as *K-DIME*. In *K-DIME*, individual models for different users are built using neural network. In [107], Itten's color contrast theory [108] is applied for feature extraction. Wang *et al.* [109] also developed emotion semantic based features for affective image retrieval, while other works, such as [110] and [111], used generic image processing features (*e.g.*, color histograms) for image emotion classification. In [112], Yanulevskaya *et al.* applied Gabor and Wiccest features, which are combined with machine learning techniques, to perform emotional valence categorization. Cho [113] developed a human-computer interface for interactive architecture, art, and music design. The studies [114] and [115] focused on affective content analysis for movies clips. More recently, an International Affective Picture System (IAPS) [13] was proposed for affective image classification. Inspired by the empirical concepts from psychology and

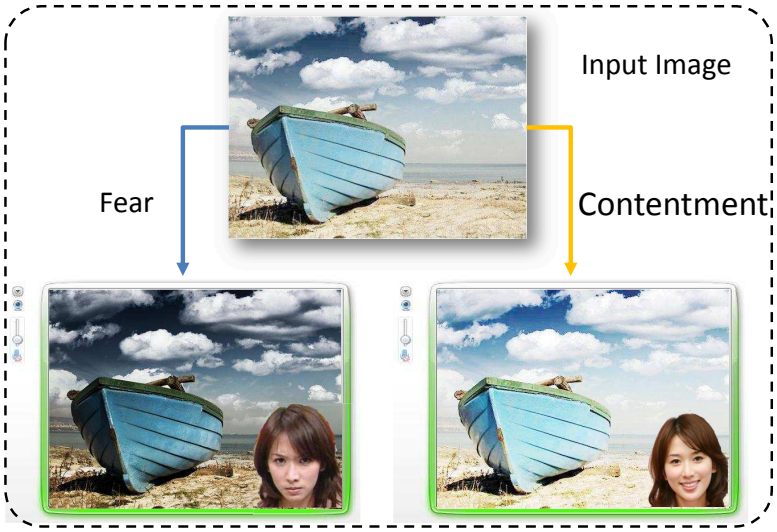


Figure 4.1: Objective of the proposed work: emotion synthesis. Given an input image, the proposed system can synthesize any user specific emotion on it automatically.

art theory, low-level image features, such as color, texture, and high-level features (composition and content), are extracted and combined to represent the emotional content of an image for classification tasks. The authors also constructed an IAPS image dataset which contains a set of artistic photography from a photo sharing site and a set of peer rated abstract paintings.

Beyond these emotion analysis efforts, one question naturally rises: could we endow a photo (image) with user specified emotions? An effective solution to this problem will lead to many potential interesting multimedia applications such as instant online messengers and photo editing software. This new function, illustrated in Figure 4.1, can help the inexperienced users create professional emotion-specific photos, even though they have little knowledge about photographic techniques. Nevertheless, this problem has rarely been studied. Not surprising, image emotion synthesis is a difficult problem, given that: 1) the mechanism of how image affect the human being's feeling evolves complex biological and psychological processes and the modern biology and psychology studies have very limited knowledge on it. Thus, mathematical modeling of this mechanism is intractable; and 2) human being's affection process is highly subjective, *i.e.*, the same image could affect different people into different emotions. Although to develop an expert system like computational model is intractable, it is quite possible that these problems could be alleviated by a learning-based approach. It is fortunate that it is not hard to obtain a large number of emotion-annotated images from photo sharing

websites such as *Flickr.com*. From a statistical point of view, images within each emotional group must convey some information and common structures which determine its affective property. Therefore, if the underlying cues that constitute an emotion-specific image can be mined by a learning framework, they can be further utilized for automatic image emotion synthesis.

The proposed solution is motivated by the recent advances in utilizing web data (images, videos, meta-data) for multimedia applications [61, 116]. First, an emotion-specific image dataset is constructed by collecting Internet images and annotating them with emotion tags by Amazon's Mechanic Turk [27]. Training images within each emotion group are clustered into different scene subgroups according to their color and texture features. Then these images are decomposed into over-segmented patch (super-pixel) representations and for each emotion + scene group, a generative model based (*e.g.*, Gaussian Mixture Models) on the color distribution of the image segments are constructed. To synthesize some specific emotion onto an input image, a piece-wise linear transformation is defined for aligning the feature distribution of the target image with the statistical model constructed from the source emotion + scene image subgroup. Finally, a Bayesian framework is developed by further incorporating a prior term enforcing the spatial smoothness and edge preservation of the transformation, and a *maximum a posteriori* (MAP) solution is inferred via a standard non-linear optimization method. Extensive user studies are performed to evaluate the validity and performance of the proposed system.

The organization of the rest of this work is as follows. Section 4.2 introduces the technical detail of image re-emotionalizing. Section 4.3 presents the experimental results and user study. Finally, the work is summarized in Section 4.4.

4.2 Learning to Emotionalize Images

This subsection first discusses the emotion-specific image dataset construction; then introduces the statistical modeling of the image emotion related features and proposes an emotion transfer model for synthesizing any user specified emotion onto the input images.

4.2.1 Dataset Construction

A training image dataset that contains emotion specific images is constructed. In this work, the images are mainly landscape images (for other categories of images, the same method applies). In [117], the International Affective Picture System (IAPS) was developed and widely used as emotional stimuli for emotion and attention investigations. Note that we do not use the dataset provided in [13] since most of the images in [13] are artistic photographs or abstract paintings, which are not appropriate for training emotion-specific models for real images such as landscape photos. A subset from the NUS-WIDE [118] image dataset, which is collected from web images, is selected as the training dataset. The interactive annotation scheme by Amazon *Mechanical Turk* is adopted to obtain emotion annotations. The web users are asked to annotate the images into 8 emotions, including *Awe, Anger, Amusement, Contentment, Sad, Fear, Disgust, Excitement* by Mechanical Turk. Only the annotations which are at least agreed by 3 (out of 5) users are accepted, resulting about 5000 emotion specific images. As mentioned, only landscape photos, *e.g.*, beach, autumn, mountain, etc. are used as the training set. Exemplar images are shown in Figure 4.2 and the statistics of the resulting image dataset are shown in Table 4.1. From Table 4.1, it can be observed that only a few landscape images are labeled as disgust or fear, thus we only consider 4 types of emotions, including two positive emotions (*i.e.*, *Awe* and *Contentment*) and two negative emotions (*e.g.*, *Fear* and *Sad*).

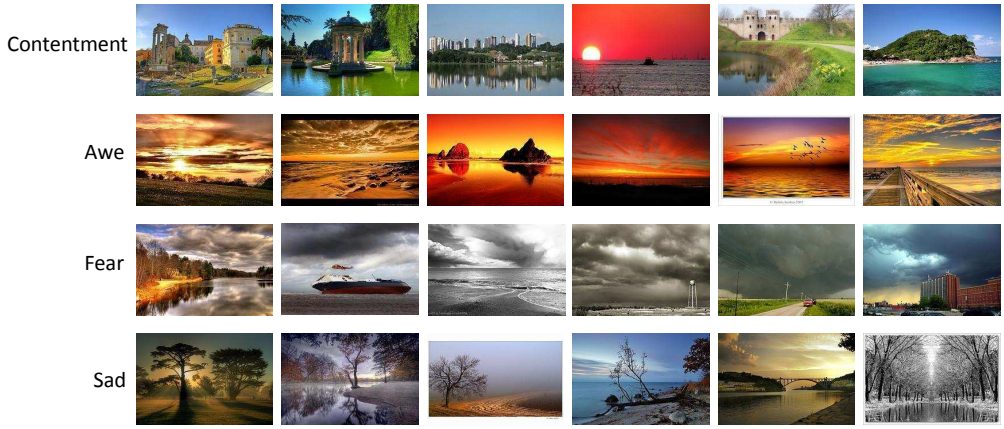


Figure 4.2: Exemplar emotion-specific images of the dataset. The exemplar images are from *Contentment, Awe, Fear and Sad*, respectively.

Table 4.1: Statistics of the constructed emotion annotated image dataset.

	Amuse.	Anger	Awe	Content.	Disgust	Excite.	Fear	Sad	Sum
Dataset	115	199	**1819**	**1643**	24	201	**238**	**627**	4866

4.2.2 Emotion-Specific Image Grouping

One can observe that even within each emotion group, image appearances may have large variations. Therefore, to develop a single model for each emotion image class is not reasonable. To cope with this problem, each emotion-specific image set is divided into several subsets first, such that the images within the same subgroup share similar appearances and structures. Then computational model is constructed for each of these image sub-groups. Similar with [116], each image is decomposed into a set of over-segmented image patches (*i.e.*, superpixels) by [119], then color (color moment) [120] and texture features (HOG) [121] are extracted and quantized by the bag-of-words model. Note that color and texture are complementary to each other in measuring image patch characteristics. Finally the images are grouped into several scene subgroups by K-means.

An illustration of the image grouping result is given in Figure 4.3. One can observe that within each scene subgroup, the images' appearances are quite similar. We can also note that different scene subgroups belong to different landscape types such as *beach, autumn, mountain,* etc.

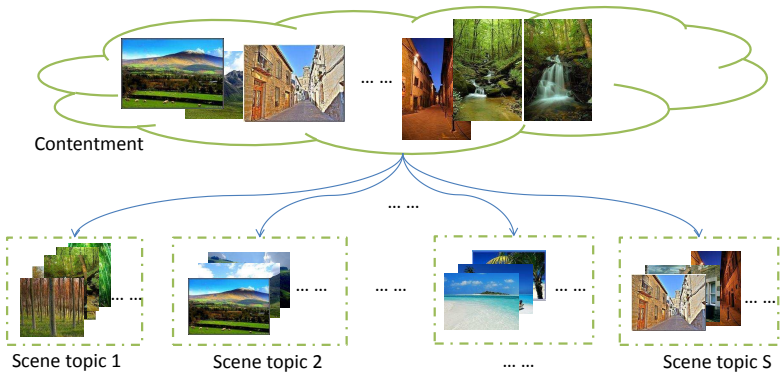


Figure 4.3: Example results of the image grouping process. The image set annotated with the emotion *contentment* is grouped into several scene subgroups.

4.2.3 Image Emotion Modeling

Emotion specific information is implicit within each emotion + scene subgroup. To uncover this information for the proposed emotion synthesis task, we use generative models, *i.e.*, Gaussian mixture models (GMM), for modeling the statistics of the image patch (segment) appearances within each emotion + scene image subgroup. We denote $\mathbf{x}$ as the appearance feature (*i.e.*, a $3D$ vector of R, G, B values) of an image patch segmented by [119]. Then each image is regarded as a bag of image segments. The reason for using this simple image features (*i.e.*, RGB color space) is that it is simple and direct for color transformation, which has been extensively demonstrated by previous works such as [36] [37]. We further denote that there are C emotion + scene image subgroups.

GMM is utilized to describe the patch feature distribution for each image subgroup $c \in \{1, 2, ...C\}$, which is given as follows:

$$p(\mathbf{x}|\Theta^c) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\mu_k^c, \Sigma_k^c), \tag{4.1}$$

where $\Theta^c = \{\mu_1^c, \Sigma_1^c, \omega_1^c, ..., \mu_K^c, \Sigma_K^c, \omega_K^c\}$. K denotes the number of Gaussian components. μ_k^c, Σ_k^c and ω_k^c are mean, covariance matrix and weight of the kth Gaussian component, respectively. The superscript c is dropped for notational simplicity for the rest of this subsection, while all the equations are presented for each emotion + scene subgroup. $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ denotes the uni-modal Gaussian density, namely,

$$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1}(\mathbf{x} - \mu_k)\}. \tag{4.2}$$

The parameters of GMM can be obtained by applying Expectation-Maximization (EM) approach.

The estimated GMM parameters $\{\Theta^1, \Theta^2, \cdots, \Theta^C\}$ can be obtained after EM, where each Θ^c characterizes the feature distribution of a specific emotion + scene subgroup.

4.2.4 Learning-based Emotion Synthesis

The input image is classified into the image subgroup within the target emotion group first. This can be achieved by first over-segmenting the input image and forming bag-of-words representation based on the color and texture features; then the nearest neighbor image in the target emotion group is found by computing the Euclidean distance of the histogram representations between the input image and the training images, and the scene label of the nearest database image is selected to be the scene label of the input image, denoted as c.

As studied in [36, 37], color (contrast, saturation, hue, etc.) can convey emotion related information. We therefore perform emotion synthesis via applying linear mapping on the RGB color space for the target image. Instead of performing global mapping for the entire image as in [37], we propose the following piece-wise linear mapping for each segment (superpixel or patch) of the target image as,

$$f_i(\mathbf{x}) = P_i\mathbf{x} + \Delta\mathbf{x}. \tag{4.3}$$

Equivalently, we can augment P, $\mathbf{x}$ with an additional constant values, *i.e.*, $\widetilde{\mathbf{x}} = [\mathbf{x}, 1]^T$, $\widetilde{P} = [P, \Delta\mathbf{x}]$ as,

$$f_i(\mathbf{x}) = \widetilde{P}_i\widetilde{\mathbf{x}}. \tag{4.4}$$

P, $\mathbf{x}$ is used to represent $\widetilde{P}$, $\widetilde{\mathbf{x}}$ for the rest of this subsection for notational simplicity. Here, $\mathbf{x}$ denotes the appearance feature of one superpixel (image segment). P_i denotes the mapping functions for operating the i-th image segment (superpixel). These image patches are superpixels which are obtained by using [119]. Note that every pixel within the same superpixel (image segment) shares the same mapping function f_i. The goal of the proposed synthesis process is to obtain the set of appropriate linear mapping functions for the entire target image (suppose there are M image segments), namely, $\mathcal{P} = \{P_1, ..., P_M\}$. The objective of emotion synthesis can be expressed by a *maximum a posteriori* (MAP) framework as,

$$\max_{\mathcal{P}}(\mathcal{F}_1 + \mathcal{F}_2), \tag{4.5}$$

The objective formulation contains two parts. The first part is a regularization term, which enforces the smoothness of the transformation and also maintains the edges of the

original image. $\mathcal{F}_1$ can be expressed as:

$$\mathcal{F}_1 = -\sum_{i,j\in N(a)} \omega_{ij}^a \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 + \lambda \sum_{i,j\in N(s)} \omega_{ij}^s \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 - \sum_{i,j\in N(c)} \|P_i - P_j\|_F^2, \tag{4.6}$$

where

$$N(a) = \{i,j|i,j \in N(c), \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \theta_1\}, N(s) = \{i,j|i,j \in N(c), \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \geq \theta_2\}. \tag{4.7}$$

Here, $N(c)$ denotes the spatial neighborhood, *i.e.*, two superpixels i and j are adjacent. θ_1 and θ_2 are the color difference thresholds. ω_{ij}^a and ω_{ij}^s are the weighting coefficients, which are defined as follows:

$$\omega_{ij}^a \propto \exp(-\|\mathbf{x_i} - \mathbf{x_j}\|_2^2/a), \omega_{ij}^s \propto 1 - \exp(-\|\mathbf{x_i} - \mathbf{x_j}\|_2^2/a). \tag{4.8}$$

Here, θ_1, θ_2, λ and a are set to be optimal empirically. We can note from this prior that: 1) The first term ensures that original contours in the target image will be preserved by enforcing that originally distinctive neighborhood segments present distinctive color values in the transformed image; 2) The second term encourages smooth transition from image segments to near-by segments. The second part of the framework is the emotion fitting term, which is expressed as:

$$\mathcal{F}_2 = \log(\prod_{i=1}^{M} p(\mathcal{I}|P_i)) = \log(\prod_{i=1}^{M} p(\mathbf{x}_i|\Theta^c)). \tag{4.9}$$

Here $p(\mathbf{x}|\Theta^c)$ is the trained GMM model for emotion+scene subgroup c, $\mathbf{x}_i$ is the color vector of the i-th image segment. We can note that this term encourages the distributions of the target image to move towards the statistical model of the training data. Finally the cost function is denoted as:

$$\begin{aligned}
&\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2 \\
&= -\sum_{i,j\in N(a)} \omega_{ij}^a \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 + \lambda \sum_{i,j\in N(s)} \omega_{ij}^s \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 \\
&- \sum_{i,j\in N(c)} \|P_i - P_j\|_F^2 + \sum_i \log(\sum_k \mathcal{N}_k).
\end{aligned} \tag{4.10}$$

Note that Eqn. (4.10) is nonlinear and complex. Therefore, to optimize the cost func-

tion, we adopt Newton's method with linear constraints, which can guarantee local optimum [122], as:

$$\max_{\mathcal{P}} \mathcal{F}, s.t. 0 \preceq P_i \mathbf{x} \preceq 255, \forall i = 1, 2, \cdots, M, Section5 \tag{4.11}$$

where $\preceq$ denotes component-wise inequality constraints. These constraints ensure that the resulting color value is within appropriate range. Our method is schematically illustrated in Figure 4.4.

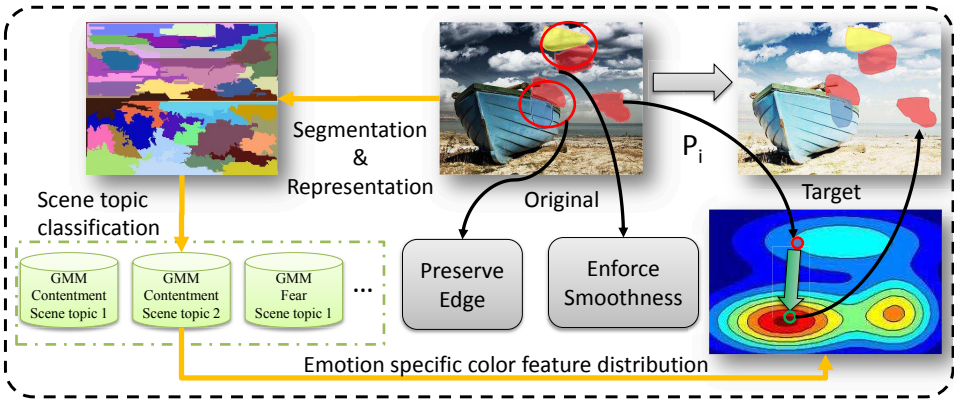


Figure 4.4: The learning-based emotion synthesis scheme.

4.3 Experiments

This subsection introduces the experimental settings, user studies along with discussions. As mentioned in the previous sections, during the pre-processing stage, training images within each emotion group are segregated into several scene subgroups (subclasses) based on the distributions of the image superpixels' HOG and color moment features. For each subclass, a GMM is trained to describe the distribution of superpixels' color information.

Given an arbitrary input image, in the emotion synthesis phase, we first assign it to the nearest scene subclass based on the HOG and color moment bag-of-words representation. Then the proposed task is to obtain a mapping function which can minimize Eqn. (4.11). Since the proposed probability function is non-convex, we can easily get trapped in a local minimum. Therefore, a good initial mapping matrix is crucial. To get a proper

initialization, we firstly assign patch (superpixel) j to the nearest Gaussian component center μ_i. After that, a pseudo inverse is performed as $P^i_{inv} = \mathbf{x}_j^{\dagger}\mathbf{u}_i$, here $\mathbf{x}_j$ denotes mean color feature value of image patch j. The linear multiplier transformation part of the initial mapping matrix becomes, $P^i_{ini} = \lambda I + (1-\lambda)P^i_{inv}$. Here I is the identity matrix. In this experiment, we set $\lambda = 0.8$ empirically. With a good initialization, we can mostly obtain a good mapping matrix using standard non-linear optimization algorithms such as Newton's method.

In this experiment, we choose 55 images from the NUS-WIDE dataset which serve as the testing images while the others construct the training image set. We compare the proposed method with the color transfer method proposed in [36], which directly aligns the mean and standard deviation of the color distribution between the source (reference) and the target image. The target image is mapped with the reference image chosen from the emotion + scene subclass by nearest neighbor assignment (in terms of the HOG and color moment based bag-of-words representation).

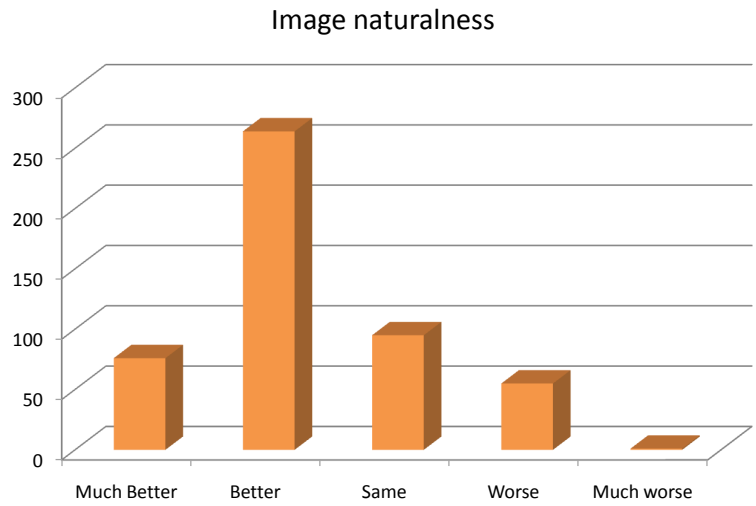


Figure 4.5: The statistics from the user studies. Nine participants are asked to compare the naturalness between 55 pairs of results from the proposed method and color transfer method. The yellow bar shows the summation of user's feedback based on naturalness, *i.e.* whether the result of the proposed method is Much Better, Better, Same, Worse, Much Worse than the result of color transfer.

The comparative user studies are conducted as follows. Firstly, the transformed images of both methods are presented to the participants in pairs (with the left-right order randomly shuffled). Participants are asked to decide whether these images can

express the specified target emotion. The naturalness of the synthesized images is also considered, since the naturalness will significantly affect the image quality. In this sense, the participants are also asked to compare which image of the same pair is more natural. In particular, participants are asked to give a judgment that whether the left image is Much Better, Better, Same, Worse, Much Worse than right one. In the user study, 9 participants are asked to judge the image's naturalness, and 20 participants with ages ranging from 20 to 35 years old are asked to judge whether these images can express the target emotion. The statistics of the results for the user study are illustrated in Figure 4.5 in terms of the naturalness. Several example results are shown in Figure 4.6 for both the proposed method and the color transfer method.

Table 4.2: Perceptibility comparison of each emotion set

	Awe	Contentment	Fear	Sad	Average
Baseline	0.4375	0.3600	0.4833	0.3788	0.4082
Our Method	0.6250	0.6550	0.6100	0.7904	0.7045

In Figure 4.5, yellow bars show the number of participants voting for each type of the ratings. It can be observed that the images resulting from the proposed method are more natural to the audience than the ones from the color transfer method. This could be explained by the fact that the statistic modeling using GMM is more generative and robust, while the exemplar image based color transfer might sometimes lead to over-fitting. Figure 4.5 and Table 4.2 show that in most cases images which are synthesized using the proposed method outperform the color transfer results in terms of the accuracies of emotion synthesis. Figure 4.6 further shows that the results of the proposed method are more natural than color transfer based results. As can be seen, color transfer based results rely on reference images. Therefore, if the color distribution of reference image is too far from the target image, the transformed result will be unnatural, *e.g.*, trees in the last example look red which are not realistic. However, the statistical learning based method does not have such a problem.

4.4 Summary

In this chapter, a learning based image emotion synthesis framework which can transfer the learnt emotion related statistical information onto arbitrary input images was devel-

oped. Extensive user studies well demonstrated that the proposed method is effective and the re-emotionalized images are natural and realistic.

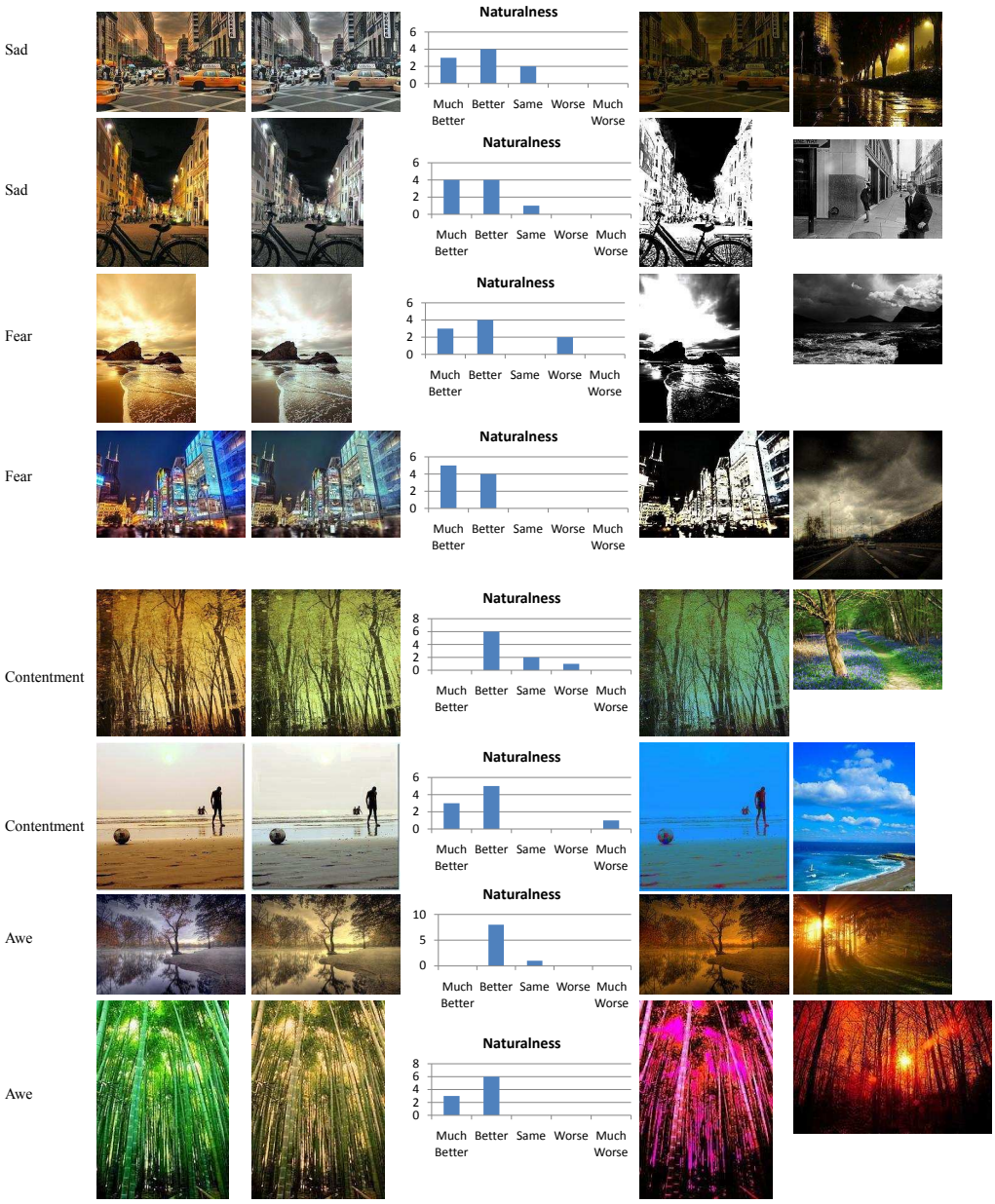


Figure 4.6: Example results of the image emotion synthesis. Each row, from left to right, show the original image, synthesized image using the proposed method, naturalness evaluation bar, color transfer result and reference image in color transfer. The middle blue bars show statistics of user's responses which indicate based on naturalness whether synthesizing result (left) is Much Better, Better, Same, Worse, Much Worse than the result from color transfer method (right). For better viewing, please see in x2 resolution and in color pdf file.

Chapter 5

Learning to Photograph

5.1 Introduction

Shooting a photograph has never been as easy as nowadays. Recent advances in computer vision algorithms along with the new generation embedded processors have introduced a new breed of built-in functionalities for digital cameras, such as automatic focus, face detection etc., all of which aim to assist the amateur photographers. These features usually adopt basic photographic rules. For instance, 1) focus length can be automatically adjusted based on the relative size and position of the detected subject/object; and 2) camera viewpoint should enclose the detected face region.

Beyond these features, there is a demand to develop more *intelligent* functionalities for digital cameras, among which automatically finding an aesthetic view rectangle from the wide camera input view has recently begun to attract research attention [14, 15]. Solution to this problem is promising, in the sense that it can bring dramatic convenience for millions of amateur photographers. An illustration of the aesthetically optimal view rectangle recommendation applications is shown in Figure 5.1, in the scenarios of a) wide angle photo input; and b) continuous photo sequence input.

Producing quality photograph requires mastery of various fundamental photographic skills. These skills are built on comprehensive considerations on photographic elements

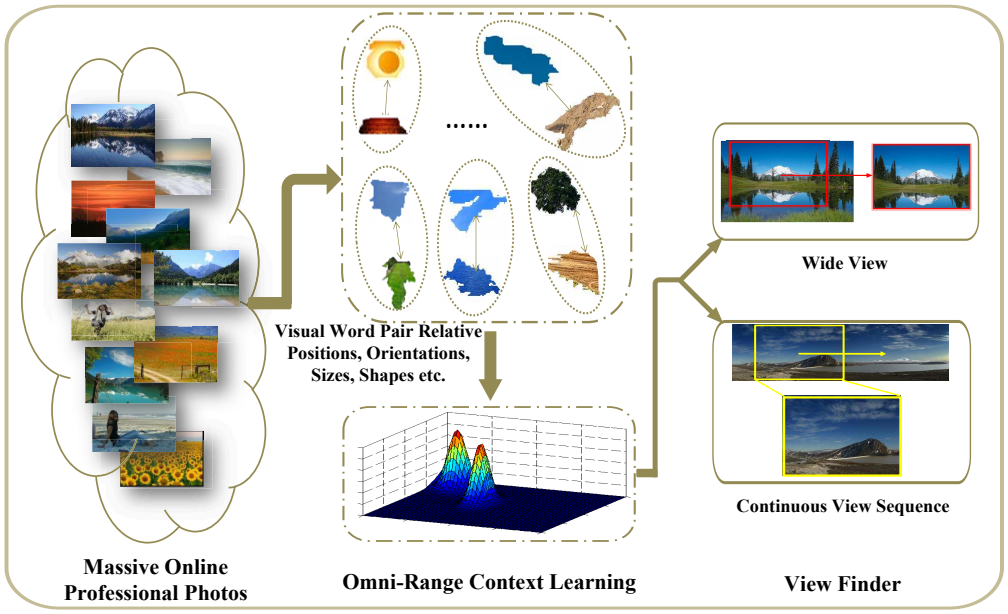


Figure 5.1: The objective of this work is to develop an automatic view finder by learning composition rules of photos with high aesthetic quality from massive online photo collections. Omni-range contexts are learned from the co-occurrent patch prototypes (visual word pairs) shown in the middle column. For better view, please see the colored pdf file.

such as: lines, shapes, colors, lighting, texture, contrast, focal length and photo compositions. Much of the previous work for the purpose of *optimizing the aesthetics of the original photograph* has concentrated on low level image processing, such as enhancing the *harmoniousness, attractiveness etc.* of the original images via 1) color transferring [123–125] or 2) changing the image illumination condition [126] (*e.g.*, image re-lighting). To take more compelling photographs, however, higher level considerations, *i.e.*, photo compositions, have to be taken into account. Photographic composition is the pleasing arrangement of subject matter elements within the photo area. Although it is true that *the only rule in photography is that there are no rules* [127], there are a number of well-established heuristic compositional guidelines in a general sense. These photographic compositional rules include *Rule of Thirds*, *Balancing Elements*, *Golden Section Rule*, *Diagonal Rule*, *Size of Region*, etc. [15], which lend the photo with a natural balance, draw the viewer's attention to the important parts of the scene, or lead the viewer's eye through the photo.

Several studies have been proposed for developing computational models of photo-

graphic compositional rules to increase the aesthetic appreciation of photographs. Traditional artistic compositions have been adopted in artistic rendering of $3D$ scenes [128]. *Rule of Thirds* has been applied to replace object boundaries of $3D$ models in view selection [14] and to position the region of interest (ROI) in robot application [129]. *Balance Elements* has been applied heuristically to arrange images and text objects [130]. Face [131] and eye [132] detection (tracking) results have been utilized for guiding the photo composition. [133] uses the web-based photo collections to construct a photo quality classifier that assesses the composition quality of images, which is then utilized for cropping the best view. More recently, Liu *et al.* [15] developed a novel computational means for evaluating the composition aesthetics of a given image based on measuring several well-grounded composition guidelines, including *Rule of Thirds*, *Visual Balance*, *Diagonal Dominance*, *Size of Region*, etc.

While the above mentioned rules are developed based on simplified scenarios, large variations of the photo content (*i.e.*, when the arrangements of the visual elements in the scene are very complex), however, complicate the problem. Besides, the judgment of photo aesthetics is subjective and involves sentiments and personal taste [16]. In both scenarios, the previously developed computational models for photo composition rules are not always applicable. Methods that embrace more generalization capability are yet to be explored.

The prevalence of the photo sharing websites such as *Flickr.com* has shed some lights. On the one hand, these websites provide millions of sharing photos, a large portion of which are taken by professional photographers. On the other hand, these images are intensively viewed and rated by users, and this type of rating information can be regarded as a very important indicator of aesthetics. Without loss of generality, the underlying visual structures shared by the highly rated photos can be significantly correlated with the photo aesthetics. Motivated by this observation, we develop a learning-based framework which discovers the underlying aesthetics photo composition structure from a large set of user-favored online sharing photos, for the purpose of viewpoint recommendation. Of particular interest is how to model photo composition in a general way. To cope with this problem, we develop an *Omni-Range Context* modeling method which lies in the synergy between the photo composition and the image spatial context [134–138]. In addition to local co-occurrence encoding [136, 138], we find that for the specific task, it is important to consider the co-occurrence contexts between image

visual elements (*i.e.*, regions or patches) which are situated within arbitrary distance to each other. Based on the omni-range context modeling, we then formulate aesthetically optimal view rectangle searching problem as *Maximum A Posterior* by imposing additional photographic constraints.

The main contributions of this work are summarized as follows:

1) A learning based framework for photo composition modeling based on omni-range context is developed.

2) A probabilistic inference framework is proposed for aesthetically optimal view rectangle recommendation, based on the combination of the mined omni-range context prior and other photographic constraints.

3) A large image dataset of professional landscape photos is constructed, which can be further made publicly available for encouraging this research direction.

This work is organized as follows. The proposed learning pipeline for omni-range context modeling is introduced in Section 5.2. The automatic aesthetically optimal view finding algorithm is presented in Section 5.3. Section 5.4 gives the experimental results in terms of qualitative evaluations as well as user studies. Section 5.5 concludes the work with future work discussion.

5.2 Learning Photographic Compositional Rules

5.2.1 System Overview

An image processing/machine learning pipeline is proposed for encoding aesthetic photographic compositional rules from a large set of *professional* images crawled from photo sharing websites, *e.g., Flickr.com*. Images are first classified into different sub-groups according to their color and texture features. Within each sub-group, each image is segmented into several coherent regions (patches) based on color and texture features. These patches are clustered into a set of visual elements (visual words). Then, for each visual

element pair, the generative mixture models is utilized to estimate the co-occurrence distribution of a set of geometric entries such as ratio of visual element areas and sizes, relative visual element positions and orientations etc, known as *omni-range context*. Besides, additional computational compositional rules are developed from training images in the presence of human subject. These mined rules are treated as prior knowledge and for a novel photo input, this prior information is further combined with other photographic constraints, yielding a posterior probability formulation. Finally, a sampling based mode seeking method is proposed to infer the Maximum A Posterior (MAP), which corresponds to finding the aesthetically optimal view rectangle in terms of users' favors. An illustration of the proposed system pipeline is given in Figure 5.2.

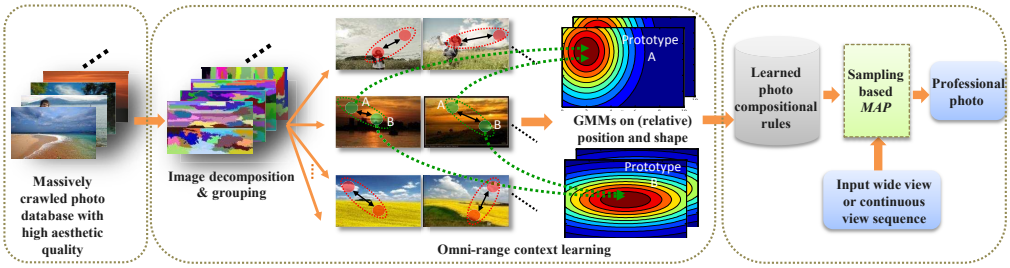


Figure 5.2: An illustration of the proposed pipeline for aesthetically optimal photo view recommendation. The database images are first clustered into several sub-topics according to color and texture features. Within each sub-topic, images are segmented into visually consistent patches (visual words). Then omni-range context is mined for each pair of patch prototype. In this example, the joint spatial (x, y) and shape (w, h) distributions of the visual word pair A and B is illustrated in terms of two marginal distributions for each of A and B, respectively, as the $4D$ joint distribution cannot be visualized. When a new view input is presented, the mined omni-range context is utilized to automatically search for an aesthetically optimal sub-view from the input view.

5.2.2 Professional Photo Database Construction

There are billions of images available through image sharing websites, however, photos with professional qualities are not trivial to obtain. For example, the popular photo sharing website *Flickr.com* has a collection of more than 10^{10} images uploaded by a large population of users, which include both professional photographers and amateurs. Therefore, the qualities of the uploaded images have large variations. To filter out the low

quality images, we leverage the interactive functions of *Flickr.com*, namely, the number of views and user ratings for each image. It can be observed that those highly rated images are likely to be more aesthetic than the lowly rated ones. This work focuses on several landscape-type images since taking a good landscape photo requires more dedicated photo composition knowledge to arrange various scene element such as sea, river, mountain, street etc. Nevertheless, the proposed framework can be applied to general scenarios. In particular, we use different text queries such as *mountain, beach, forest, river, street etc.* for crawling images from *Flickr.com*, and images which are favored by more than 10 users are retained as candidate images for the database. Note that ten users voting cannot strictly guarantee the quality of the images. To remove low quality photos, we first perform image clustering (which is introduced in the following section) and remove those images which are far from every cluster center. After some manual inspection for further noisy image removal, we finally obtain a large-scale professional photo pool consisting of more than $80,000$ highly rated images. Note that although in this work, we take natural landscape photos as example, the developed method is applicable to all type of photos, *i.e.*, family photos, traveling photos etc. We only need to change the relative parameters (such as the number of visual elements, the number of Gaussian mixture models etc.) of the proposed method to handle different types of photos.

5.2.3 Image Sub-topic Grouping

Images in the constructed database have large variations in content and it is intractable to derive a single composition rule for all type of images. To address this issue, database images are grouped into several classes (*i.e.*, topics) first, so that within each image subtopic, the image structures and appearances are similar and the learned photographic composition rules can be more informative.

The well-known bag of words model [139] is used to represent image. Histogram of oriented gradients (HOG) [140] (*i.e.*, a 72-D vector) features are extracted. More specifically, HOG feature is obtained by evaluating normalized local histograms of image gradient orientations in a dense grid. In practice, this is implemented by dividing each image block (patch) into several small spatial regions, and for each region accu-

mulating a local histogram of 8 gradient directions over the pixels, thus yielding a 72-D feature vector. In this work HOG features are extracted densely by a step size of 8 pixels.

About 1000 to 3000 HOG descriptors can be obtained from each image in average, given that the average image size is about 400×533 pixels. To form the bag of words representation, we then cluster the set of HOG descriptors into a set of visual words (*i.e.*, m) by K-means method. Then, each HOG descriptor is mapped to one visual word by nearest center assignment and each image therefore can be represented as a m-D histogram vector. In this work, the size of the visual vocabulary m is empirically set as 1000 according to [141], which is standard for image classification tasks. In addition to modeling texture, we also calculate the color histogram as in [142] for each image, namely, each image is then represented by a 256-D vector (*i.e.*, 256 colors). K-means is performed in this 1256-D space to group the images into K sub-topics (*i.e.*, $K = 100$ in this work, which is a trade-off between image grouping performance and representation complexity. We note that under the 1256-D image representation, 100 subgroups can well represent different types of scenes, *i.e.*, within each subgroup the images are visually consistent) according to the similarity measured by Euclidean distance. Figure 5.3 illustrates sample images of the collected photo database with examples of several subgroups. It can be observed that these database photos are of professional aesthetic quality. Also, the segregated images look visually similar within each sub-topic, *i.e.*, the scene type of each image sub-group is consistent.

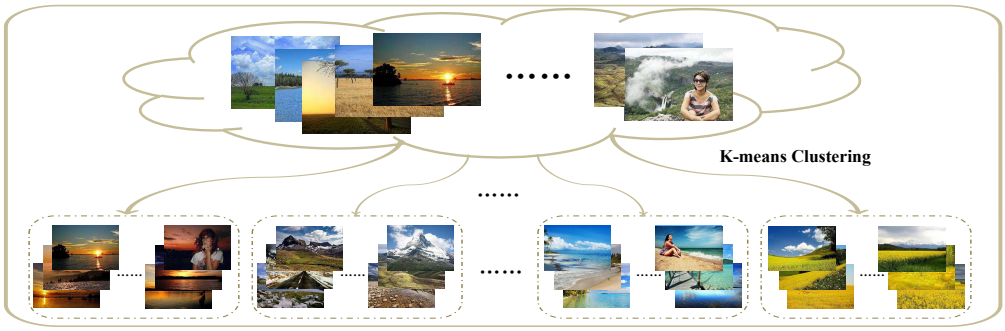


Figure 5.3: Illustration of sample images of the collected image database with examples of several photo sub-topics.

5.2.4 Visual Element Representation

Each image is composed of a set of image segments (patches) such as *sky*, *grass*, *sea*, etc., and any object of interest, with certain spatial arrangements. In photo composition, these image segments (patches) are denoted as visual elements and the way how they are spatially arranged corresponding to photo composition. The mean-shift [143] based segmentation algorithm is used to decompose the images into visual elements (*i.e.*, patches, segments), and the bandwidth parameter is set empirically. The segmentation process is as follows. First, all the images are resized and normalized and each pixel is initialized as one atomic patch. Then, the color features are used to describe the appearance of the initial image patches. We then apply the algorithm as in [144] to the segmentation results from the mean-shift algorithm to merge the smaller patches into larger ones iteratively. The merging process is stopped if the patch size is larger than a predefined threshold. An image is then segmented into $5 - 20$ regions in average. We use the same color plus texture feature descriptors as in subsection 5.2.3 to represent each patch (*i.e.*, a 1256-D vector), with pooling inside each image segment instead of the whole image. Without image annotations, to infer the semantic label for each segment is intractable. Therefore, the set of image segment descriptors are grouped into M visual words $V_1, V_2, \cdots, V_m$ (*i.e.*, image segment prototypes) by applying K-means again. In this work, the size of the visual vocabulary M is empirically set as 200 (this value is chosen based on the trade-off between visual element representation compactness and visual element grouping performance). Each image segment can be mapped to one visual word by nearest center assignment. It can be noted that the final image representation quality and the final inference algorithm will base on the quality of image segmentation. However, as the segmentation method which used in this work is deterministic (namely, the final segmentation results do not depend on initialization) and we do not require each segment convey explicit semantic meaning, the segmentation results will only depend on the image appearance itself and images which have similar appearance structures lead to similar segmentations so that the final inference algorithm will not be affected significantly. In some rare cases when the segmentation for the testing image deviates too much from the segmentations used for model training, the final inferred view rectangle will be poor. Such a failure example will be shown in the experimental section.

5.2.5 Special Visual Element: the Human Subject

Human subjects in a photo usually contain more semantic information and attract more attention from the observers.

A little change of the human subjects' poses, sizes, locations can affect human's aesthetic assessment process significantly. Therefore, special attention should be paid to the composition between human subject and other visual elements in the scene.

To address this problem, we explicitly detect human subject regions from the image by applying the LBP + HOG based human detector [145]. In the meantime, a multi-view face detector based on the Adaboost method [146] is used to extract face instances in three different poses, *i.e.*, left, right and front. To represent these three poses, we define three augmented visual words, namely subject faces to the left as V_{M+1}, right V_{M+2}, and front V_{M+3}, respectively (*e.g.*, assuming we have generated M visual words for ordinary image patches). Note that by using this augmented visual word representation, these special visual words can be treated in the same way as other visual words under a unified framework in the subsequent photographic compositional rule modeling. The scheme is also directly applicable when there are multiple human subjects in the scene, as each detected human subject in the scene is represented as a $V_i, (i = M+1, M+2, M+3)$ and the relationship between each $V_i, V_j, (i, j = M+1, M+2, M+3)$ pair can also be naturally encoded, and in fact the relative geometric property of each $V_i, V_j, (i, j = M+1, M+2, M+3)$ pair makes contribution to photo composition. Also note that we set 3 human visual words for the sake of trade-off between effectiveness and complexity. For most natural scene images (as concerned in this work) which contain human subject(s), three different facing directions are sufficient for compositional rule modeling (it can also directly deal with interactions between human subjects). For more complicated photos involving human subjects such as family photos, human subjects have more poses and facing directions. In those cases, we can increase the number of human visual element types straightforwardly, *i.e.*, by defining more and finer types of poses. For example, in the future work, part based model [147] can be used to simultaneously detect person and their fine poses, and assign different visual words to different fine poses.

5.2.6 Learning Photographic Compositional Rules: Omni-Range Context Modeling

From photographic compositional perspective, automatic finding an optimal view enclosure from the input view is equivalent to anchoring a view rectangle (*i.e.*, whose translation, rotation, and scaling parameters are to be decided) on the input view and let the objects/regions within the rectangle present in a way which is consistent with the spatial arrangement (or in other words, spatial context models) of the visual elements from the professional photos used for training.

There are two complementary aspects of spatial context modeling. First, each patch prototype (visual element) is associated with a spatial and geometric probabilistic distribution model, which illustrates the probability of the patch prototype over different image locations and shapes. Second, for pairs of patch prototypes (visual elements), the probabilistic distribution model of their spatial and geometric relationship well corresponds to the photo composition characteristic. Therefore the purpose in this work is to model these two aspects properly and the obtained model can be used to predict and recommend the best view rectangle.

An image I is represented as a bag of visual elements (or segments), denoted as $\{S^1, S^2, \cdots, S^{N(I)}\}$, where $N(I)$ denotes the number of visual elements in image I. Each segment S^i is encoded by a visual word $V^i \in \{V_1, V_2, \cdots, V_{M+3}\}$, as mentioned in subsections 5.2.4 and 5.2.5. The occurrence frequency, the spatial and geometric distribution of each type of visual element (visual word prototype) constitute one aspect of the photographic composition rules. In particular, the prior distribution $p(V(S) = V_i)$ for a visual element S assigned to visual word V_i for each image sub-topic $\mathcal{G}^k$, $k = 1, 2, \cdots, K$ (note that all discussions in this subsection consider composition rule modeling within one sub-topic) is modeled as a multi-nominal distribution as:

$$p(V(S) = V_i) = \theta_i, \forall i = 1, 2, \cdots, M+3, \tag{5.1}$$

$$\sum_i \theta_i = 1, \theta_i \geq 0, \forall i. \tag{5.2}$$

The parameters $\theta_1, \theta_2, \cdots, \theta_{M+3}$ can be estimated from all image visual elements extracted from sub-topic $\mathcal{G}^k$ by observing the number of its occurrences in this sub-topic,

namely,

$$\theta_i = \frac{|\{S|V(S) = V_i\}|}{|\{S\}|}, S \in \mathcal{G}^k, \forall i = 1, 2, \cdots, M+3, \tag{5.3}$$

where $|.|$ denotes the number of elements in the set.

In the meantime, the spatial and geometric distribution for each visual word V_i are assumed to be Gaussian mixture models (GMM). Namely, we denote $\mathbf{x}(S) = (x, y, w, h, a)^T$ as the normalized center of mass coordinate (x, y), and normalized width, height and area (w, h, a) with respect to the image rectangle, for image segment S (see Figure 5.4 for illustration). Given that the image segment S corresponds to visual word V_i, the conditional distribution of $\mathbf{x}(S)$ is then expressed as:

$$p(\mathbf{x}(S)|V_i) = \sum_{t=1}^{T_1} w_t^i \mathcal{N}(\mathbf{x}|\mu_t^i, \Sigma_t^i), \tag{5.4}$$

where $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ denotes a Gaussian component and T_1 denotes the number of mixtures. In this work, T_1 is empirically set as 100 for the trade-off between complexity and effectiveness, *i.e.*, when $T_1 = 100$ the log-likelihood value of Eqn. (5.5) is above a predefined threshold which guarantees that the models well fit the data. The Gaussian mixture model parameters $\Theta_{V_i} = (w_1^i, w_2^i, \cdots, w_{T_1}^i, \mu_1^i, \mu_2^i, \cdots, \mu_{T_1}^i, \Sigma_1^i, \Sigma_2^i, \cdots, \Sigma_{T_1}^i), i = 1, 2, \cdots, M+3$ can be estimated from the all the extracted visual elements (patches) within the sub-topic $\mathcal{G}^k$ via expectation-maximization (EM) algorithm [148]. The purpose of EM is to maximize the log-likelihood function:

$$\mathcal{L}(\Theta_{V_i}) = \ln p(\mathcal{D}^i|\Theta_{V_i}) = \sum_{i=1}^{N^i} \ln \sum_{t=1}^{T_1} w_t^i \mathcal{N}(\mathbf{x}|\mu_t^i, \Sigma_t^i). \tag{5.5}$$

Here N^i denotes the total number of patches corresponding to visual word V_i. $\mathcal{D}^i$ represents the total set of observation data associated with V_i.

In the E-step, we compute the class assignment probability:

$$Pr(t|\mathbf{x}_l) = \frac{w_t^i \mathcal{N}(\mathbf{x}_l|\mu_t^i, \Sigma_t^i)}{\sum_{t'=1}^{T_1} w_{t'}^i \mathcal{N}(\mathbf{x}_l|\mu_{t'}^i, \Sigma_{t'}^i)}, \tag{5.6}$$

$$N_t^i = \sum_{l=1}^{N^i} Pr(t|\mathbf{x}_l). \tag{5.7}$$

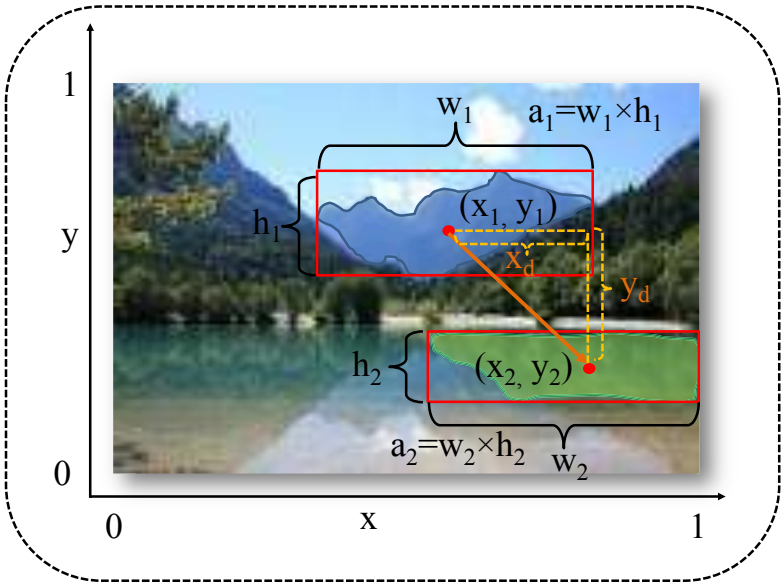


Figure 5.4: Definitions of various geometric entries for visual element and visual element pair.

Then the M-step updates the mean vectors and covariance matrices, namely:

$$\hat{\mu_t}^i = \frac{1}{N_t^i}\sum_{l=1}^{N^i} Pr(t|\mathbf{x}_l)\mathbf{x}_l, \tag{5.8}$$

$$\hat{\Sigma_t}^i = \frac{1}{N_t^i}\sum_{l=1}^{N^i} Pr(t|\mathbf{x}_l)(\mathbf{x}_l - \hat{\mu_t}^i)(\mathbf{x}_l - \hat{\mu_t}^i)^T, \tag{5.9}$$

$$\hat{w_t}^i = \frac{\sum_{l=1}^{N^i} Pr(t|\mathbf{x}_l)}{\sum_{t=1}^{T_1}\sum_{l=1}^{N^i} Pr(t|\mathbf{x}_l)}. \tag{5.10}$$

A more important aspect for photo composition is the relationship between visual elements in the same scene which including the co-occurrence pattern and spatial and geometric relations, etc. This type of relationship information, formally known as *spatial context* in computer vision, is widely adopted in image classification [136], object recognition [137] and face recognition [138] owing to its discriminative capability. The state-of-the-art methods for spatial context modeling [136, 138] only consider co-occurrence properties between neighboring image features; however, this scheme is insufficient for modeling photographic composition, as the visual element pairs which are not spatially adjacent even play an more important role in conveying aesthetic information. For ex-

ample, we can consider the case of *sunrise on the beach*, where the *sun* and the *beach* are concurrent yet always spatially separated by the sea in the photo. In this case, modeling the spatial joint layout for the visual element *sun* and *beach* is more aesthetically significant and this information cannot be captured by local spatial context. To address this limitation, an *omni-range* context framework which can model the spatial arrangement of two visual elements with arbitrary mutual distance is proposed.

On the one hand, we define a joint prior distribution $p(V(S) = V_i, V(S') = V_j)$ to measure the importance for each type of visual element (visual word) pair S, S' as:

$$\begin{aligned} & p(V(S) = V_j, V(S') = V_i) = \theta_{ij}, \\ \forall i, j \quad = \quad & 1, 2, \cdots, M+3, \qquad (5.11) \\ & \sum_i \sum_j \theta_{ij} = 1, \theta_{ij} \geq 0, \forall i, j. \qquad (5.12) \end{aligned}$$

Similarly, the parameters can be estimated from all image visual element pairs S, S' extracted from sub-topic $\mathcal{G}^k$ by observing the number of its occurrences in this sub-topic, namely,

$$\begin{aligned} & \theta_{ij} = \frac{|\{(S, S') | V(S) = V_i, V(S') = V_j\}|}{|\{(S, S')\}|}, S, S' \in \mathcal{G}^k, \\ & \forall i = 1, 2, \cdots, M+3. \qquad (5.13) \end{aligned}$$

Note that this joint prior distribution is the probabilistic form of the well-known co-occurrence model.

On the other hand, to describe the spatial and geometric dependence of two visual elements S and S', we define a set of normalized relative spatial and geometric entries. These entries include relative offset vector $(x_d = x - x', y_d = y - y')$ which denotes the 2-D displacement vector from the center of mass of the visual element S to S', and the ratios of the heights, widths and areas of both visual elements as $(r_h = h/h', r_w = w/w', r_a = a/a')$. Figure 5.4 illustrates the definitions of these entries. The conditional distributions of $\mathbf{y}(S, S') = (x_d, y_d, r_h, r_w, r_a)^T$ given the visual element prototype pair V_i, V_j is represented by Gaussian mixture models (GMM) as:

$$p(\mathbf{y}(S, S') | V_i, V_j) = \sum_{t=1}^{T_2} w_t^{ij} \mathcal{N}(\mathbf{y} | \mu_t^{ij}, \Sigma_t^{ij}). \qquad (5.14)$$

In this work, T_2 is also empirically set as 100 for the trade-off between complexity and effectiveness. Similarly, the Gaussian mixture model parameters can be represented as:

$$\Theta_{V_i,V_j} = (w_1^{ij}, w_2^{ij}, \cdots, w_{T_2}^{ij}, \mu_1^{ij}, \mu_2^{ij}, \cdots, \mu_{T_2}^{ij}, \Sigma_1^{ij}, \Sigma_2^{ij}, \cdots, \Sigma_{T_2}^{ij}), \tag{5.15}$$

where $i, j = 1, 2, \cdots, M + 3$. It can be estimated from the training patch pairs within the sub-topic via expectation-maximization (EM) algorithm.

The advantages of the proposed omni-range context modeling method are that: 1) both single patch spatial and geometric distribution information as well as the spatial and geometric co-occurrence relationship (*i.e.*, spatial context in other words) for pairs of prototype patches are modeled, which constitute two important and complementary aspects for image spatial context modeling; 2) using Gaussian mixture models, the co-occurrence relationships between all pairs of image patches with arbitrary spatial distance can be well described within a unified probabilistic distribution framework, which apparently overcomes the theoretical limitations of the previous local spatial context modeling methods [136, 138], *e.g.*, far context cannot be modeled by those methods; and 3) the learning framework is a unified framework because it makes no difference between the ordinary visual words (extracted from image patches) and the 3 face visual words.

Relation to Heuristic Photographic Composition Rules: the proposed omni-context framework is a general probabilistic model which can well represent the heuristic photographic composition rules, *e.g.*, *Rule of Thirds*, *Balancing Elements*, *Golden Section Rule*, *Diagonal Rule*, *Size of Region*, etc. For instance, *Rule of Thirds* and *Size of Region* can be well expressed in terms of the conditional distribution $p(\mathbf{x}(S)|V_i)$ as in Eqn. (5.4); and *Diagonal Rule* can be modeled by the conditional distribution of $p(\mathbf{y}(S, S')|V_i, V_j)$ as in Eqn. (5.14). The merit of the proposed method lies in the fact that by using generative models, any types of composition learned from the professional photo collection could be captured by probabilistic distributions.

5.3 Apply Omni-Range Contexts: Aesthetically Optimal View Finding

This subsection introduces a probabilistic framework for aesthetically optimal view finding, by utilizing the learned omni-range context model(see Section 5.2.6). Given an input image I, we first calculate the HOG + color histogram feature representation as in Section 5.2. Then, based on this feature representation, the input image is classified into one of the image sub-topics by searching for its nearest neighbor image in the training database and inherit its sub-topic label $\mathcal{G}_k$. In the rest of this subsection, we assume the omni-context models and the algorithmic formulations to be mentioned are associated with the sub-topic $\mathcal{G}_k$. The sub-topic subscript is ignored for notational simplicity.

The input image I is further decomposed into a set of visual elements $\{S^1, S^2, \cdots, S^{N(I)}\}$ as in Section 5.2. Given the learned omni-context model $\mathbf{\Omega}$ (which includes $\{\Theta_{V_i}\}$, $\{\Theta_{V_i,V_j}\}$, $\{p(V(S) = V_i)\}$ and $\{p(V(S_1) = V_i, V(S_2) = V_j)\}$) of image sub-topic $\mathcal{G}_k$, the objective is to find a maximum value of the posterior probability $p(\boldsymbol{\eta}|I, \mathbf{\Omega})$ for the parameter set $\boldsymbol{\eta}$ of the view rectangle to be pursued, where $\boldsymbol{\eta} = (s, r, \theta, \mathbf{t})$ and s, r, θ and $\mathbf{t} = (t_x, t_y)^T$ refer to the scaling, width/height ratio, rotation and the translation parameters of the predicted view rectangle. Note that we assume that the input image is already horizontally aligned (which is quite reasonable in realistic), therefore the parameter θ is ignored in the rest of this subsection, *i.e.*, $\boldsymbol{\eta} = (s, r, \mathbf{t})$.

Using the Bayesian rule, this *Maximum A Posterior* (MAP) problem can be transformed into the joint probability $p(\boldsymbol{\eta}, I|\mathbf{\Omega})$ maximizing problem, by recognizing that the evidence term $p(I|\mathbf{\Omega}) = p(I)$ is a constant and could be ignored in the maximizing operation, denoted as:

$$
\begin{aligned}
Q(\boldsymbol{\eta}) &= p(\boldsymbol{\eta}, I|\mathbf{\Omega}), && (5.16)\\
&= p(\boldsymbol{\eta}|I, \mathbf{\Omega})p(I|\mathbf{\Omega}), && (5.17)\\
&\propto p(\boldsymbol{\eta}|I, \mathbf{\Omega}). && (5.18)
\end{aligned}
$$

This term can be further expanded into several factors by recursively applying the Bayesian

rule as:

$$p(\boldsymbol{\eta}, I|\boldsymbol{\Omega}) = p(\boldsymbol{\eta}|\boldsymbol{\Omega})p(I|\boldsymbol{\eta}, \boldsymbol{\Omega}), \tag{5.19}$$

$$= p(\boldsymbol{\eta})p(I|\boldsymbol{\eta}, \boldsymbol{\Omega}), \tag{5.20}$$

where we use the fact $p(\boldsymbol{\eta}|\boldsymbol{\Omega}) = p(\boldsymbol{\eta})$.

On the one hand, the prior probability $p(\boldsymbol{\eta})$ can be further decomposed into three factors as:

$$P(\boldsymbol{\eta}) = p(s)p(r)p(\mathbf{t}). \tag{5.21}$$

Here we assume the independence of $p(s)$, $p(r)$ and $p(\mathbf{t})$, which represent the prior distribution for the scaling, rotation and translation in the following form, respectively:

$$p(s) = \mathcal{N}(s|s_0, \sigma_s), \tag{5.22}$$

$$p(r) = \sum_{i=1}^{i=K_r} \omega_i \mathcal{N}(r|\mu_r^i, \sigma_r^i), \tag{5.23}$$

where $\mathcal{N}(x|\mu, \sigma)$ denotes a Gaussian distribution. The parameters s_0, σ_s, μ_r^i and σ_r^i are estimated from training images. Note that we set $K_r = 2$ for the mixture of Gaussian distribution $p(r)$ and the two Gaussian models corresponds to portrait and landscape images, respectively. $p(\mathbf{t})$ is modeled as a uniform distribution.

On the other hand, the likelihood probability term $p(I|\boldsymbol{\eta}, \boldsymbol{\Omega})$ can incorporate the learned omni-range context model, which essentially measures the compatibility between the concerned sub-image and the photo composition rules mined from professional photos. Specifically, if we denote visual element set enclosed by the concerned view rectangle $I(\eta)$ are $\{S_1, S_2, \cdots, S_{N(I(\eta))}\}$, where are in turn quantified as $\{V(S_1), V(S_2), \cdots, V(S_{N(I(\eta))})\}$, the likelihood $P(I|\boldsymbol{\eta}, \boldsymbol{\Omega})$ could be expressed as:

$$\begin{aligned} p(I|\boldsymbol{\eta}, \boldsymbol{\Omega}) = & \prod_{i=1}^{N(I(\eta))} p(V(S_i))p(\mathbf{x}(S_i)|V(S_i)) \\ & \times \prod_{i=1,j=1}^{N(I(\eta)),N(I(\eta))} p(V(S_i), V(S_j)) \\ & p(\mathbf{y}(S_i, S_j)|V(S_i), V(S_j)). \end{aligned} \tag{5.24}$$

This posterior probability is complicated and has no explicit analytical form, therefore, in order to derive the optimal parameter, *i.e.*, $\boldsymbol{\eta}_{opt}$, we adopt the commonly used Markov Chain Monte Carlo (MCMC) [149] based sampling method to seek the mode (maximum) of the target posterior $Q(\boldsymbol{\eta})$. Namely, in each sampling step, we fix other parameters to their old values and then randomly sample a new value for the current concerned parameter according to certain proposal probability. For instance, in the step of updating s, we fix θ and $\mathbf{t}$ for the current time τ and sample the new value of s according to proposal distribution $q(s|s^\tau, \theta^\tau, \mathbf{t}^\tau)$. In this work, let $q(s|s^\tau, \theta^\tau, \mathbf{t}^\tau) \sim \exp\{-\frac{(s-s_\tau)^2}{\sigma^2}\}$, where σ is a bandwidth parameter which trades off sampling stableness and convergence speed. The sampling procedure iterates until convergence, *i.e.*, the estimated posterior distribution becomes stable. Finally, the parameter set, which yields the highest value is retained as the optimal solution of view rectangle. Although the solution distribution is typically multi-modal, MCMC based sampling methods can always yield globally optimal solution. A multiple random initialization scheme is adopted to accelerate the convergence speed. Given that the parameter space is only $4 - D$, the convergence of the sampling procedure is quite fast [149]. Figure 5.5 shows a sampling procedure in terms of distribution convergence and sample trajectory for a four-visual elements example. Stable distribution is attained after about 10000 sampling iterations, refereing to the distribution and trajectory convergence illustrations. Note also that in the current work, there are about 200×200 visual word pairs, however, as mentioned in the previous section, the number of visual segments (elements) obtained in an image is typically less than 20, and thus the number of pairs is much less than 200×200, *i.e.*, computation (*e.g.*, for evaluating Eqn. (5.24)) is quite feasible during testing. In the current computational platform, (*i.e.*, 2.8GHZ CPU, 4GB memory, un-optimized Matlab code), the whole testing process which includes both feature extraction, sub-topic classification, image segmentation and the optimal view inference takes less than 5 second. More specifically, feature extraction and image segmentation takes around 1 second and the MCMC based sampling method takes around 4 seconds for inferring the best view rectangle. To make the algorithm flexible for embedded systems such as digital cameras and mobile devices, we plan to improve the computational efficiency by 1) replacing the current image segmentation method with recently proposed fast segmentation method such as [150]; and 2) using GPU-based parallel computing architecture for implementing the sampling algorithm. We believe by adopting these approaches, the system can be efficiently implemented into these portable devices with embedded cameras.

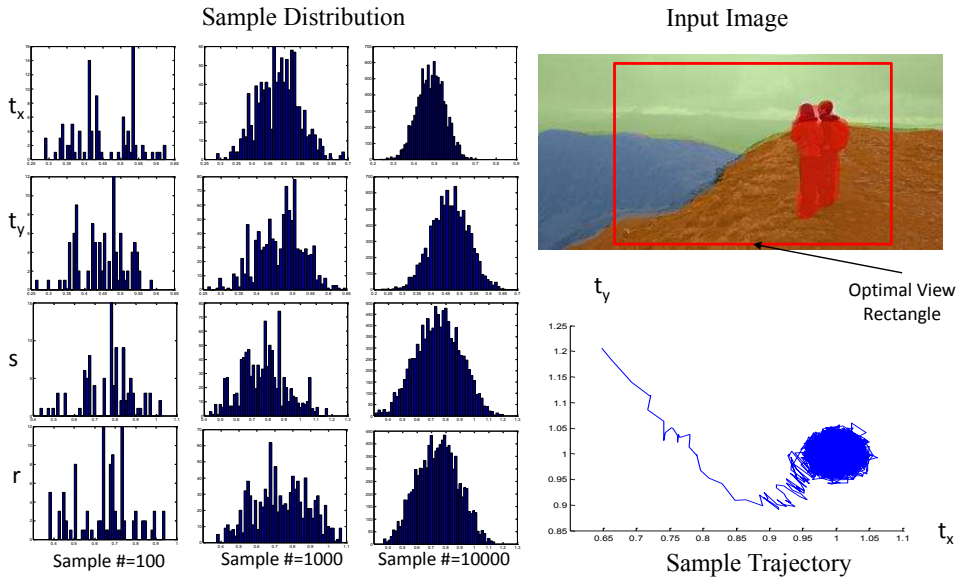


Figure 5.5: Example of sampling convergence procedure. The left three distribution columns show the distributions of samples after 100, 1000 and 10000 sample iterations for t_x, t_y, s and r, respectively. The right upper figure shows the input photo, the four segmented visual elements and the solution rectangle. The right lower figure shows the sampling trajectory of (t_x, t_y). It can be seen that the sampling process converges after about 10000 iterations.

5.4 Experiments

This subsection presents both qualitative and quantitative experimental results on automatic aesthetically optimal view finding. We will first visualize the omni-context models learned from the constructed professional photo database and then present user studies that compare the proposed method with other state-of-the-art methods for view recommendation.

5.4.1 Qualitative Evaluation: Omni-Range Context Visualization

Photographic compositional rules are represented by modeling the statistical distributions of each visual word and word pair's geometric (or relative geometric) configurations such as (relative) $2D$ positions, (relative) height, width and area. These distributions are described by Gaussian mixture models (GMMs). In Figure 5.6, several exemplar distributions $p((x, y)|V_i)$ ($2D$ position), $p((w, h)|V_i)$ (weight and height),

$p((dx, dy)|V_i, V_j)$ (displacement between visual word pairs) and $p((w_i/w_j, h_i/h_j)|V_i, V_j)$ for different visual words and visual word pairs with high priors ($p(V_i)$, or $p(V_i, V_j)$) are illustrated. We also relate the visual elements (pairs) to their corresponding positions in Gaussian mixture models. Figure 5.6 shows that relative displacements and relative shape parameters (*e.g.*, height and width) of visual element pairs such as *sky over flowers*, *mountain over grass/water/boat* constitute the far contexts. These are typical examples of photo compositional rules which can be commonly found in most user favored natural scene photos.

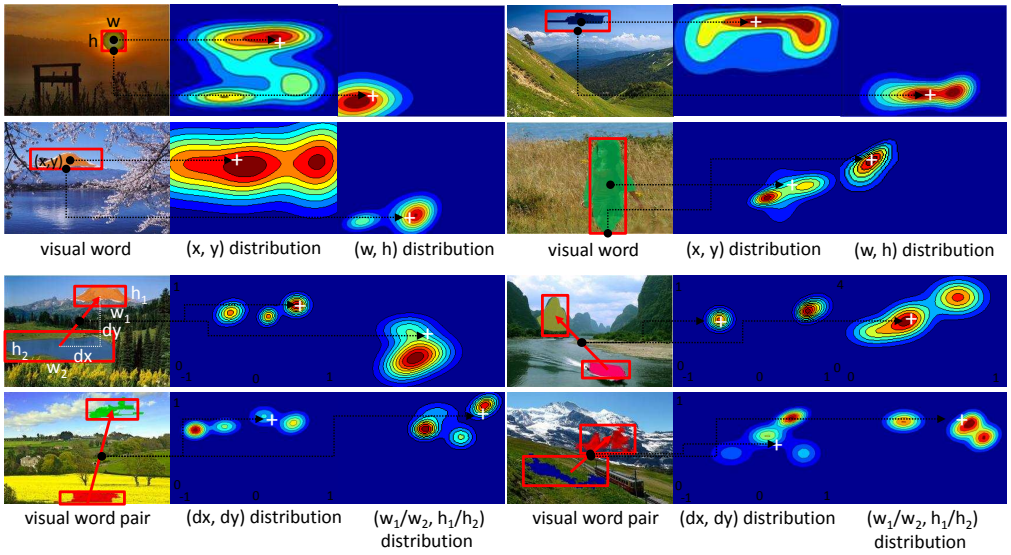


Figure 5.6: Visualization of the spatial (x, y), (dx, dy) and shape (w, h), $(w_1/w_2, h_1/h_2)$ distributions for individual visual words (first row) and the omni-range contexts between patch prototype-pairs (second row), respectively. Except for those values shown on the distribution figures, the default range of x-axis and y-axis is $[0, 1]$.

5.4.2 Quantitative Evaluation: User Studies

For the purpose of algorithmic evaluation, 76 wide-field (*e.g.*, panoramic) photos are collected from the Internet as the input testing photos. Then the proposed automatic view finding algorithm is applied on each input image. For algorithm comparison, we also compare the state-of-the-art view finding algorithms including: 1) the visual attention model proposed in [41], denoted as VA; 2) the photo composition optimization method

Table 5.1: The first two columns illustrate the mean and standard deviation values of the rating scores from the user studies on the comparison of the proposed method ORC-2 and other methods including VA, OPC and ORC-1. The rest columns illustrate the ANOVA test results. The p-values show that the difference of two comparing methods is significant and the different of users is insignificant.

ORC-2 vs. OPC		The factor of algorithms		The factor of users	
ORC-2	OPC	F-statistic	p-value	F-statistic	p-value
1.4 ± 0.808	0.3 ± 0.647	31.37	9.58×10^{-7}	0.11	1.0
ORC-2 vs. VA		The factor of algorithms		The factor of users	
ORC-2	VA	F-statistic	p-value	F-statistic	p-value
1.0 ± 0.904	0.3 ± 0.647	13.27	7.0×10^{-4}	0.34	0.9999
ORC-2 vs. ORC-1		The factor of algorithms		The factor of users	
ORC-2	ORC-1	F-statistic	p-value	F-statistic	p-value
0.7 ± 0.789	0.3 ± 0.646	5.44	0.02038	0.42	0.9987

proposed in [15], denoted as OPC; 3) early version of the omni-range context learning based method (denoted as ORC-1) as in [63] (*i.e.*, without modeling the height, width or area for visual words). The proposed method in this work is denoted as ORC-2. For all comparing algorithms, their related parameters are empirically set to be optimal. As there lacks a computational metric for measuring the aesthetic quality of photos, we conduct user studies in terms of user preference for all comparing results. For each input testing image, the recommended view rectangles by all methods are presented to human subjects in the randomly shuffled order. Each human subject is required to give a judgment score to each resulting sub-photo. A five-point scale is used for rating the user's preference, where the score value of 5.0 corresponds to the most favorable. Note that for each testing image, different methods might receive same scores. In this experiment, 50 human subjects with uniform gender distribution and with ages ranging from 23 to 35 years old conduct user studies. The statistics of the user study result are illustrated in Figure 5.7. The score results are converted into ratings to further demonstrate the statistical significance of the proposed method. For each user, we assign a rating of 0 to the worst method (if his average rating of this method on all testing images is lower than that of the other methods), and the other methods are assigned a rating of 1, 2, or 0 if it is better than (average rating is higher), much better than (average rating is much higher), or comparable to this method, respectively. Since there will be disagreements among the evaluators, we perform a two-way analysis of variance (ANOVA) test [151] to statistically analyze the comparison. It partitions the observed ratings into components

corresponding to different explanatory factors, and it is able to test the significance levels of the rating differences with respect to the factors of comparing algorithms and user. The comparison results are illustrated in Table 5.1. The results demonstrate the superiority of the proposed approach over the other methods. The ANOVA test shows that the difference of algorithms is statistically significant and the difference of the evaluators is not significant. This further confirms the effectiveness of the proposed approach. It can be observed from Figure 5.7 that the view rectangle found by the proposed method is more preferred than other state-of-the-art methods (including the earlier version, ORC-1), from a statistical view point. Figure 5.8 shows several examples of the resulting view recommendations by all comparing algorithms along with the corresponding users' scores [1]. It can be observed from Figure 5.8 that 1) the proposed method consistently outperforms other state-of-the-art, and it also improves the earlier version of this work; 2) the resulting sub-photos from the visual attention model (VA) are quite similar to that of the proposed method in cases that the regions with rich texture coincide with the aesthetically optimal view rectangle that are favored by users. However, when the photo has some spurious highly textured regions (*e.g.*, see example (a)), the visual attention model based method fails; 3) in some cases, using pre-defined compositional rules such as *Rule of third*, *Visual Balance* or *Diagonal Dominance* do not result in aesthetically optimal view (*i.e.*, see examples (c), (d) and (e) respectively), due to the complexity of human's judgement model. However, the proposed model directly learns compositional rules from a large number of user favored photos, and it can avoid such over-fitting problem and mostly obtain better view recommendations for the users; 4) the proposed method is more general than the early version as it includes not only the position information but also the size and shape information for the visual elements (patches), which is important for some scenes (*e.g.*, in example (i) the scale of the *human* visual word is important in photo aesthetical quality assessment); 5) When human subjects are present in the photos, the proposed approach properly captures the composition rule for this case, *i.e.*, when a person faces to the right (left), then there should be more margin in the right (left) side of the human face (see examples (i) and (j)). However, neither attention model based method or the OPC method are able to model such important rules; and 6) in some cases when segmentation quality is low such as the highly textured sky region in example (n), *i.e.*, the sky region is segmented into many noisy small visual elements, these visual elements cannot fit the trained generative model well and the proposed method could fail

[1] All testing results are uploaded to: https://sites.google.com/site/ltpprojectpage/

to capture the optimal view rectangle.

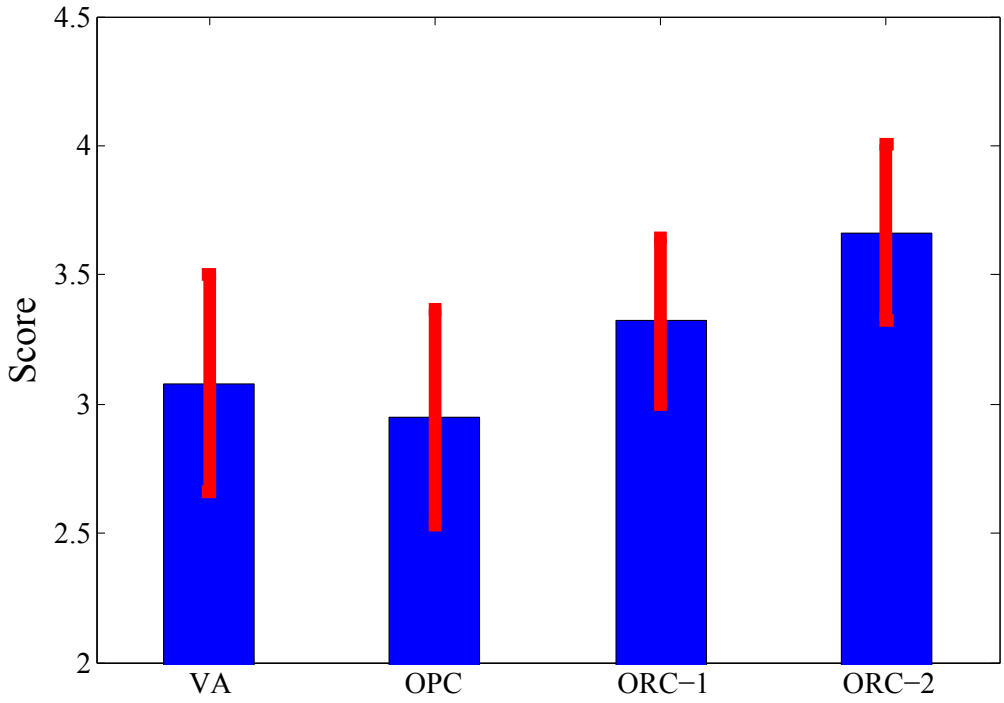


Figure 5.7: Statistics of the user study results. The results are illustrated in term of average ratings from 50 subjects for 76 testing images with standard deviations.

An additional experiment is performed to further validate the proposed photo evaluation metric for aesthetically optimal view finding, as shown in Figure 5.9. We slide a fixed-sized view rectangle along the horizontal direction on the input panoramic image (*i.e.*, which simulates the case of camera movement for seeking best view) and crop the sub-photo stepped by equal distance (10 pixels). 50 human subjects are then asked to rank these cropped sub-photos in terms of their preference level. In particular, the five levels are Very Good (5), Good (4), Moderate (3), Bad (2), Very Bad (1). In the meantime, we evaluate the objective function Q defined in Section 5.2 (*i.e.*, (5.16)) for each cropped sub-photo and take its logarithm. The user rankings as well as the corresponding objective function values are shown in the aligned mode for better comparison. It can be observed that the developed photo quality assessment metric is quite consistent with users evaluations, which further validates the effectiveness of the proposed method.

5.5 Summary

In this chapter, we have proposed an omni-range spatial context model for encoding professional photographers' experiences of photo composition rules, which are mined from massively crawled photos of good aesthetics quality from photo sharing website. A photo quality evaluation metric for automatic aesthetically optimal view finding is developed based on the learned omni-range contexts. Extensive experiments as well as comprehensive user studies demonstrated the effectiveness of the proposed method.

This work is not meant to provide a full solution to the automatic camera view finding problem, but rather it aims to inspire more interests in this new and practically important and challenging research direction. In the future work we plan to develop a more general and comprehensive professional photo quality assessment model which considers not only photo compositional rules but also other important photography elements such as exposure, contrast, etc.

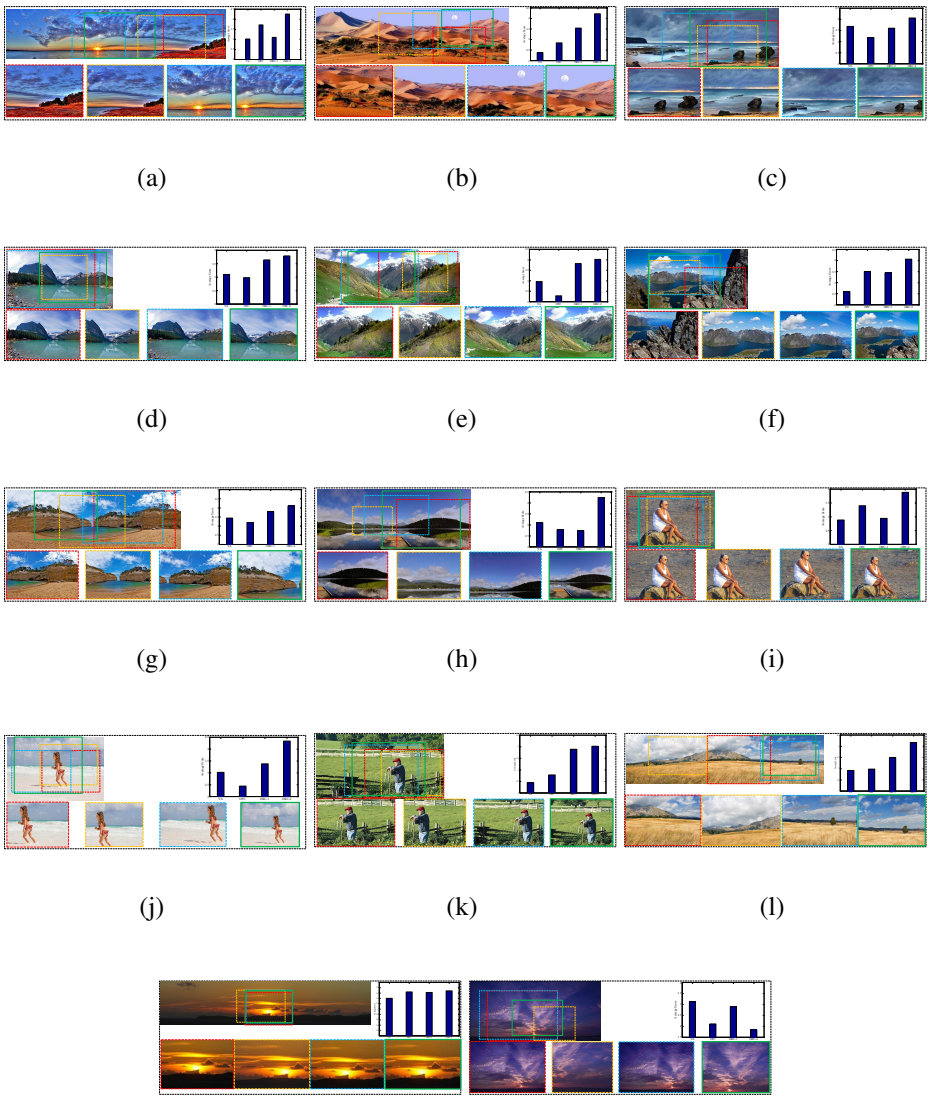


Figure 5.8: Examples of the comparison results. The upper rows show the original input photo and the average user's ratings for VA (red), OPC (yellow), ORC-1 (blue) and ORC-2 (green) methods, respectively. The bottom rows show the corresponding recommended view rectangles by different methods. For better view, please see the original color pdf.

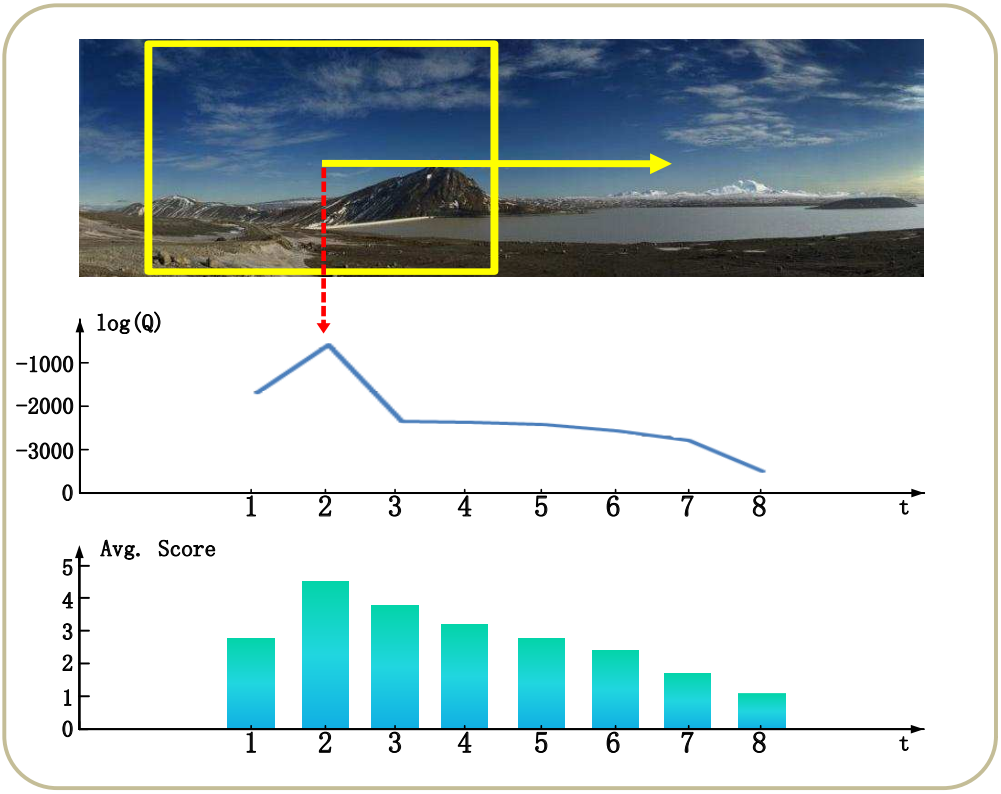


Figure 5.9: The photo aesthetic quality assessment function values (*i.e.*, (5.16) in the logarithmic form) and the corresponding average user's rankings for the sub-photos cropped by sliding a fixed size view rectangle along the horizontal direction. It can be observed that trends of both evaluations are consistent.

Chapter 6

Touch Saliency

6.1 Introduction

Visual saliency, which refers to the preferential fixation on the meaningful region in a scene, has been extensively studied in vision and psychology literature (*e.g.*, [17–19]), due to its various applications in image segmentation [20], multimedia [21, 22] and image retargeting [23, 24], etc. Fixation datasets are required for visual attention study. Typically, human subject is required to look at the image stimuli and his eye fixation data is recorded. Many fixation datasets [20, 25, 26, 64–67] use eye tracking devices to collect human fixation points to study human visual attention.

However, eye tracking devices are not as popular as mobile devices and are complicated for operation (which requires non-trivial calibration efforts). Therefore, it is impractical to construct a large-scale (*e.g.* million scale) eye-fixation database by using eye tracking devices.

The recently emerging usage of smart phones and pocket PCs with touch screen, such as iPhone and iPad, inspires us to develop an alternative to the conventional eye tracking based fixation data acquisition method. People spend several hours a day on online surfing using smart phones or pocket PCs with touch screen. With the limited screen size, when browsing images, most users are used to sliding fingers to move or

zoom in the images, and fix the region of interest from time to time. These finger movements and fixation operations indicate human interest and attention on certain regions of the image, which is in essence similar to the eye-fixation data in visual attention study. We believe that such screen touch information, which is referred as *Touch Saliency*, can be used the same way as the eye-fixation map for visual attention study. Compared with the conventional eye-fixation map, touch saliency possesses its unique advantages: 1) it is generated more easily and efficiently due to easy operation nature of touch devices, even for a little kid since no complicated calibration phase like that of the eye tracking devices is required at all; and 2) it is much cheaper owing to the wide popularity of touch screen devices as well as the prevalence of online photo and image sharing web sites. Therefore, the mission to construct a web-scale touch saliency database is possible.

One question naturally arises: whether the touch saliency can reasonably capture the true underlying visual attention of human and is its performance compare with that of eye-fixation data in visual attention study? To address this question, in this work, we conduct a series of studies on the proposed touch saliency with the conventional eye-fixation based saliency served as touch ground truth. In particular, this work aims to find the relationship between touch saliency and visual saliency, and study the possibility of using touch information instead of visual information for attention study. To facilitate this study, we 1) build an interface based on the touch screen smart phone and collect a database of users' touch saliency data; 2) conduct comprehensive comparative studies between touch saliency and conventional visual saliency.

Since saliency detection has been extensively studied for many years, a lot of prediction models (*e.g.*, [17, 18, 68, 69]) and features have been studied. Existing saliency detection methods can be divided into two categories: bottom-up and top-down. Typically, some low-level features, such as color, intensity, texture, etc., are extracted. The saliency maps are detected for the features and finally combined as the output saliency map. Some researchers combine the efficient low, mid and high-level features to train the model and finally obtain a high performance [26]. The simple face and object detection results are always used as high level feature. Thus it is possible to further improve the saliency detection performance by adding other middle or high level features. In this work we train models for image segments and use the prediction result as the middle-level category features to study the performance. The contributions of this work can be summarized as follows:

- An iPhone interface to is built collect human touch saliency and generate a saliency map, which is firstly studied to our knowledge besides the preliminary study [152].
- By comparing the performances of some state-of-the-art methods, we analyze the differences between visual saliency and touch saliency. Further discussions show that it is efficient to get human attention information with the help of mobile devices.
- We propose middle-level category features which represent the segment information. The middle-level category features are utilized into the Multi-task Sparse Pursuit (MTSP) saliency detection model. Evaluation results show that this method outperforms other state-of-the-art unsupervised models.

The rest of the work is organized as follows. Section 6.2 introduces touch saliency dataset collection and experiment results. In Section 6.3 we will define middle-level category features and apply it in Multi-task Sparse Pursuit (MTSP) saliency prediction model. The experimental results will also be shown in this part. Finally, the conclusion and future work are summarized in Section 6.4.

6.2 Touch Saliency VS. Visual Saliency: Fixation Maps Observation

Different from the preliminary touch saliency study [152], we choose another widely used visual saliency dataset MIT dataset [26] to collect touch saliency. The image size within this saliency dataset is around 1024×700. Instead of using the entire dataset, we choose 500 representative images from this released dataset to relieve users' burden. These 500 images are chosen by computing BOW feature and k-means clustering to ensure that they are representative. The same as the preliminary touch saliency work [152], an interface on iPhone is designed to collect touch saliency information. During data collection process, users are asked to freely view the photos. The movement of the image will be recorded. To get the fixation point, we assume that the center point of the image shown on the screen is the fixation point. The zooming-in scale for each point is different, and the image part which can be seen on the screen is the visual window.

This subsection shows the detailed data collection, saliency map generation process and observations.

6.2.1 Dataset Collection

Stimuli

500 representative images in MIT dataset [26] are used as stimuli. This dataset includes 1003 images, which are landscape images, street view images, human face images, etc. It is constructed to study the saliency based on low, middle and high-level features. These 500 representative images are selected by using BOW feature and k-means clustering.

Participants

There are totally 69 participants joined this experiment. They are students and non-students from China, aged from 5 to 70 years old (with mean age μ=29.59, and age standard deviation σ =15.89). The same as the preliminary study [152], all participants are naive users. They are only asked to freely view the images. Noted that here "freely view" means participants can view the images freely; they are not advised to focus on the specific object or part of the image, *e.g.* building, people. The only requirement is that they have to zoom in the images when necessary. Each participant viewed as least 100 images and most participants viewed at least 400 images.

Data Collection Procedure

Similarly to the preliminary study, a data collection interface is built on iPhone. Figure 6.1 shows the interface of the data collection app. Here, (a) shows the login interface. Once the app is opened, the users are asked to put in name, age and gender. There are two choices, "continue" and "new", to ensure that users can continue the labeling after pauses. (b) shows the interface after login. The images are presented in a random order to avoid bias. The users are asked to freely view the images after login. (c) shows

Figure 6.1: Interface designed for data collection. (a) login interface. (b) the interface after login. (c) image after zoom in (d) black screen.

the images after zooming in. Users are asked to zoom in the images when possible. (d) shows the black screen. The images will be shown for ten seconds, after which a black screen will be shown. When the black screen is shown, users have to release their fingers to avoid the influence of the next image.

Users are asked to freely view the images. This program will record the center pixel, the scale and also the current time of the displayed image. The center point is treated as the fixation point. Staying time can describe the interest level of each part. For interesting part, the staying time would be longer since people would like to find more details. For less interesting part, the staying time would be shorter and they just skip this part. So the image with long staying time can be treated as saliency region and saliency level can be rated using Gaussian function. The center part on the screen is the most interesting part, then less to the edges.

Touch Image Selection

Since there are more user response in this work than the preliminary study, we decide to choose the ”good” data and abandon the ”bad” data. Images without zoom-in should be abandoned because they cannot provide useful information. Besides, this kind of images will lead to strong central bias. So ”good” data is defined as response with enough fixation points. The ”good” collected data is chosen using the following steps.

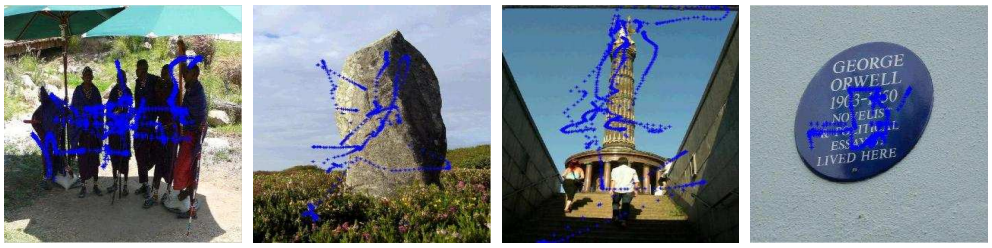


Figure 6.2: Examples of touch fixation points.

- If the duration of the first point is too long, it usually means that the participant do not focus on the experiment, and the image may be removed. Here, the duration threshold is set as 2 seconds
- When move the image, there may be some noise caused by the movement, so the points with very short duration are removed. The duration threshold is set as 0.02 seconds empirically.
- After the previous steps, if the summation of the duration is too short or only several points are left, it usually means that there is too much noise in this image. So the image with short duration summation and small amount of points is removed. The summation threshold is set as 8 seconds and point number threshold is set as 10 empirically.

The parameters are set empirically. Experimental results show that images chosen using these steps can ensure the quality of the generated saliency map. Since there are more participants in this data collection process, after the image selection process, there are still enough users for each image.

Figure 6.2 shows the collected fixation points. Here the images are rescaled for easily display. We also rescale the other result images in the rest part of this work for the same reason.

Touch Fixation Map

The visualization method mentioned in [152] is used to generate the saliency map. As mentioned in the previous part, center point shown on screen is treated as touch fixation

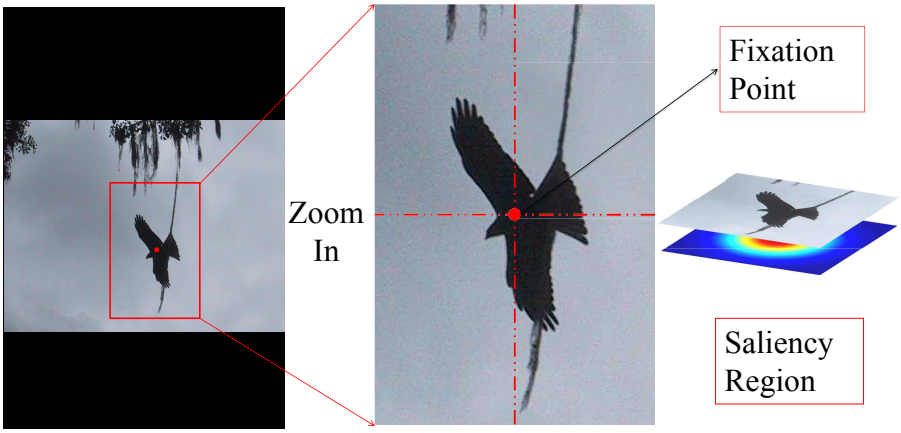


Figure 6.3: Saliency map visualization process. The center point of the image shown in screen is the fixation point. To visualize the saliency map, we use different Gaussian bandwidth parameters for different touch fixation points

point. For each point, there is a zoom-in scale. We filter out the noisy fixation points by removing the points with small zoom-in level (less than 1.2) and short duration (less than 0.02 seconds). Figure 6.2 shows fixation points. In order to build a smooth fixation map of an image, we convolve 2-D Gaussian filters for the touch fixation point. As shown in Figure 6.3, after zoom-in and other sliding operations, the center point of the image shown on screen is the fixation point. Note that the scales are different for different fixation points. Therefore different Gaussian bandwidth parameters are used for different touch fixation points thus generating a touch fixation map. The bandwidth parameters of Gaussian filters are calculated based on the zoom-in scales. For each point i with scale s_i, the Gaussian filter size is $[320/s_i, 320/s_i]$ with $\sigma = a/s_i$ (a is between 32 to 40). Here $s_i = \frac{zx}{ox}$, where zx is the width of the image after zoom-in and ox is the width of the original image. In this experiment, s_i can be larger than 1.

6.2.2 Touch and Visual Fixation Maps Comparison

This subsection systematically compares the generated touch fixation map and ground truth visual fixation map. The comparison consists of two aspects: 1) general comparison by investigating heat and focus maps, most salient regions and central bias; 2) cross estimation of these two maps to evaluate their similarity.

General Comparison

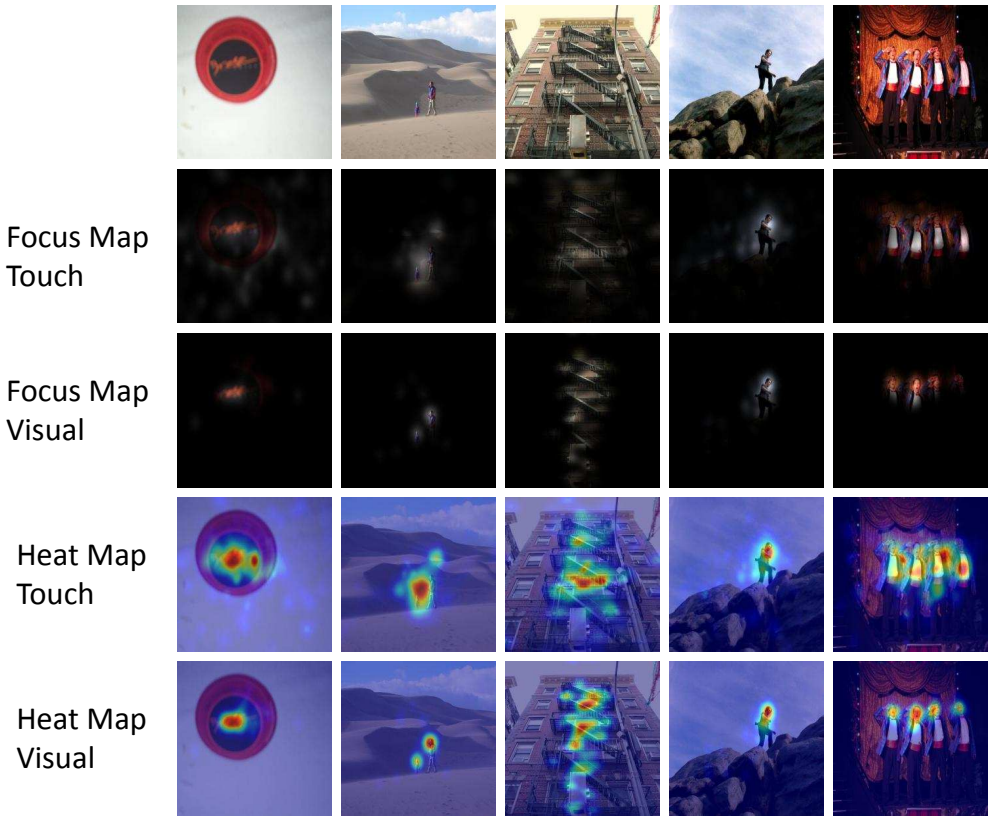


Figure 6.4: Focus and heat map of touch and visual fixation maps. From top to down are original image, heat map of touch fixation map, heat map of visual fixation map, focus map of touch fixation map and focus map of visual fixation map, respectively.

Figure 6.4 shows focus map and heat map of touch and visual fixation map. Focus map is computed by dot multiplication of the saliency map and the original image. Heat map is a summation of saliency heat map and the original image; the hotter region is more salient. It can be seen that, although not the same, the results of visual saliency and touch saliency are quite similar.

The most salient regions of touch and visual fixation maps are shown in Figure 6.5 to better investigate the touch and visual fixation maps. Figure 6.5 shows the top 5%, 10%, 15% and 20% most saliency region of visual and touch fixation maps. As can be seen in the figure, the 20% salient regions are very similar for touch and visual fixation maps. The difference on the top 5% salient regions indicates the AUC (area under curve) and CC (Correlation Coefficient) value between these two maps maybe not high.

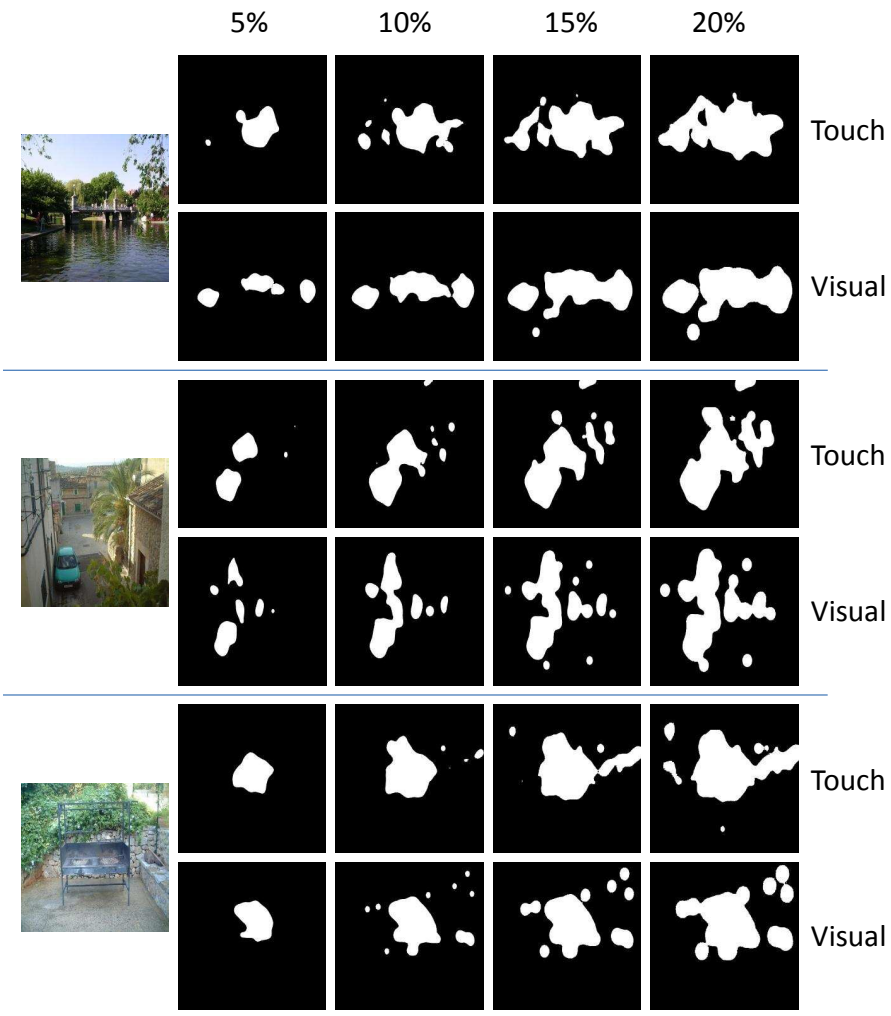


Figure 6.5: Top 5%, 10%,15% and 20% most salient region of touch and visual fixation map.

For eye tracking datasets, there is always a central bias [153] caused by making early fixations near the center of the image. To compare the central bias between touch and visual fixation maps, Figure 6.6 shows the average of all fixation maps for touch and visual saliency. It can be seen that both of these datasets have central bias, which is consistent with other eye fixation datasets [26]. The average visual fixation map is different from the map shown in [26] because only 500 images instead of the entire MIT dataset are used in this work. Human touch habit is different from viewing habit. When watch the image using touch devices, usually we will put the region of interest near the middle of the screen. However, most of the times, the interesting region is not exactly in the middle of the screen. Although we have already removed some data to reduce central

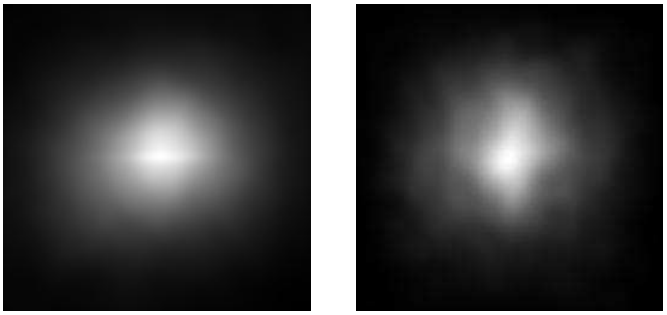


Figure 6.6: Average of all fixation maps for touch and visual saliency. Both of them have central bias.

bias, it is still strong, as shown in Figure 6.6.

Cross Estimation

This subsection compares the touch and visual saliency by computing the similarity between touch and visual fixation maps. The widely used AUC (Area Under ROC Curve) and CC (Correlation Coefficients) are utilized to evaluate the cross estimation results.

AUC is the area under the ROC curve. ROC curve can be obtained by trying all threshold values of the saliency map, plotting the false-positive-rate value on the x-axis against the true-positive-rate on the y-axis for each value. Correlation Coefficients (CC) [154] is also computed to evaluate the similarity between two saliency maps. Here, CC is defined as:

$$CC = \frac{\sum_x (M_g(x) - \mu_g)(M_p(x) - \mu_p)}{\sqrt{\sum_x (M_g(x) - \mu_g)^2 \sum_x (M_p(x) - \mu_p)^2}}, \tag{6.1}$$

where x indexes the pixels in the maps, $M_g(x)$ and $M_p(x)$ are the ground truth and prediction saliency maps, and μ_g and μ_p are the mean values of these two maps.

In this work, when compute the AUC and ROC, we use the thresholded fixation maps as ground truth. That is to say, thresholded the fixation map such that the 20 % of the pixels are salient. We use the original fixation maps as ground truth to compute CC.

Table 6.1 shows the AUC and CC values between these two maps. It can be seen that both AUC and CC values are very high, which indicates that these two maps are highly correlated with each other.

Table 6.1: The AUC and CC (correlation coefficient) comparison on the dataset.

Criteria	Ground Truth	Touch	Visual
AUC	Visual	0.8444	
	Touch		0.7908
CC	Visual	0.5187	
	Touch		0.5187

Conclusion

We qualitatively and quantitatively compare the touch and visual saliency. As shown in Figure 6.4 and Figure 6.5, touch fixation map is highly correlated with visual fixation map. The high AUC and CC values in Table 6.1 further support this conclusion. In Figure 6.5, although the most 5% salient regions are different for touch and visual fixation map, the 10%, 15% and 20% most salient regions for touch and visual fixation map are quite consistent. Thus we draw conclusion that touch and visual fixation maps are highly correlated to each other.

As mentioned above, a new touch dataset is collected and the touch saliency map is generated using the new image selection rule. Extensive studies show that the touch and visual fixation maps are similar. A deeper investigation will be provided in the following section by comparing the predictability of touch and visual saliency.

6.3 Touch Saliency VS. Visual Saliency: Prediction Model

This subsection compares the predictability of the visual and touch saliency. We compute saliency maps using some state-of-the-art saliency prediction models and compare the performance on touch and visual ground truth. Also a hybrid saliency detection method is proposed, in which segment information is added as the middle-level category features in saliency detection process to improve the detection performance. Here

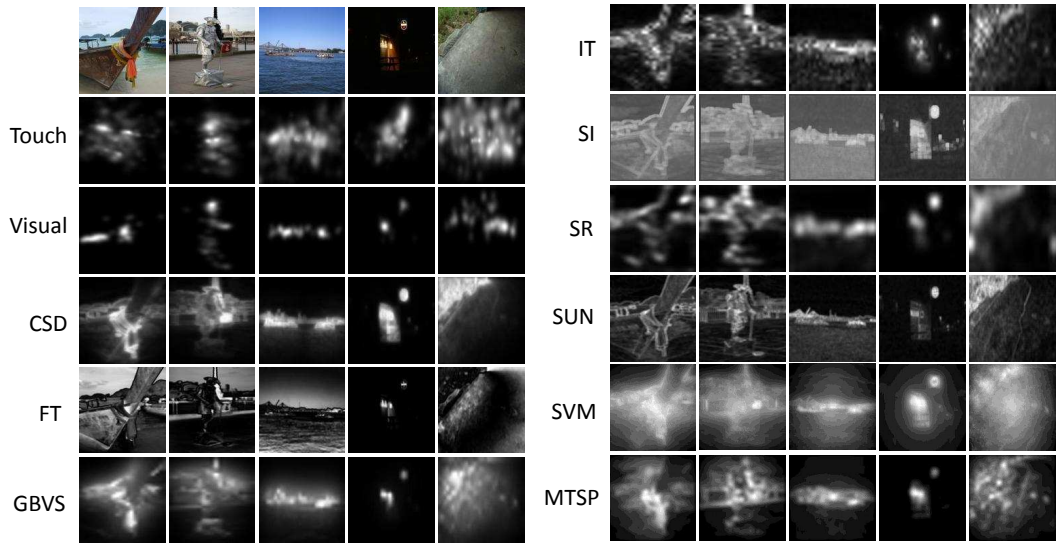


Figure 6.7: Saliency maps predicted using different methods. For the left images, from top to bottom are: original image, touch fixation map, visual fixation map, saliency map predicted using CSD, FT and GBVS. For the right images, from top to bottom are: saliency map predicted using IT, SI, SR, SUN, SVM and MTSP. Here images are rescaled to the same size for easily display.

the state-of-the-art multi-task sparsity pursuit (MTSP) model [81] is chosen as the basic model and investigate the saliency detection performance with and without the middle-level category features respectively. We choose this model because it outperforms other methods and can combine different types of features efficiently.

6.3.1 Comparison of the State-of-the-art Prediction Models

This subsection compares the predictability of touch and visual saliency. To achieve this, we compute the saliency map using state-of-the-art models and evaluate the performance on visual and touch ground truth.

The following state-of-the-art models are implemented for comparison: context aware-based saliency detection (CSD) [77], frequency-tuned method (FT) [69], graph-based-visual saliency (GBVS) [68], Itti Model (IT) [17], self-information (SI) [18], SR [155], SUN [156], SVM [26] and multi-task sparsity pursuit (MTSP) [81]. For the baseline methods, we use the released code and the default settings given by the authors.

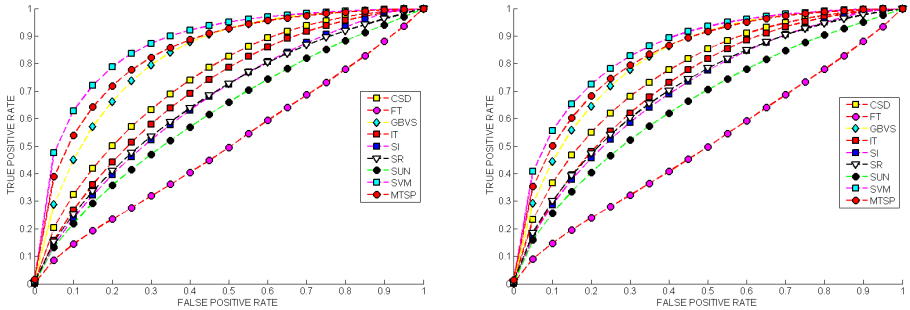


Figure 6.8: Left: ROC curve for the state-of-the-art saliency prediction results and thresholded touch ground truth. Right: ROC curve for the state-of-the-art saliency prediction result and thresholded visual ground truth.

Each image is resized to 256×256. Figure 6.7 shows some exemplar saliency maps predicted by different models.

ROC, AUC and CC are utilized to compare these methods. In the comparison, all the saliency maps are rescaled to the original size. We use the thresholded fixation maps as ground truth to compute ROC and AUC and the original fixation maps as ground truth to compute CC. Table 6.1 shows the AUC and CC values between these two maps. Figure 6.8 shows ROC curves of the comparison results. Left figure shows the ROC curve for the state-of-the-art saliency prediction results and thresholded touch ground truth. Right figure shows the ROC curve for the state-of-the-art saliency prediction results and thresholded visual ground truth. It can be seen that these prediction methods have similar values on these two ground truths. Table 6.2 shows the AUC and CC results. For AUC, "Ground Truth" means using thresholded touch or visual fixation map as ground truth; for CC, "Ground Truth" means using original touch or visual fixation map as ground truth. The AUC and CC value for touch and visual ground truth are close to each other. One may have noticed that for GBVS, SVM, and MTSP, the AUC value is higher on touch ground truth than visual ground truth. As can be seen from Figure 6.7, saliency maps of these three methods have larger salient region. Considering that touch saliency is less dense than visual saliency, it is not surprising that for these three methods (GBVS, SVM, and MTSP), AUC value on touch ground truth is higher. But all these differences are very small, both for AUC and CC value. Thus visual and touch have the similar predictability. It can be observed that MTSP outperforms other methods consistently except for SVM, which is reasonable because MTSP an unsupervised method while SVM inte-

Table 6.2: The AUC and CC (correlation coefficient) values between ground truth and the saliency maps predicted by state-of-the-art models. For AUC, Ground Truth means using thresholded touch or visual fixation map as ground truth; for CC, Ground Truth means using original touch or visual fixation map as ground truth.

Criteria	Ground Truth	CSD	FT	GBVS	IT	SI
AUC	Visual	0.7609	0.5062	0.8142	0.7258	0.7022
	Touch	0.7360	0.5046	0.8206	0.7019	0.6641
CC	Visual	0.3676	0.1534	0.4928	0.3095	0.2750
	Touch	0.3295	0.1291	0.3860	0.2663	0.2588
Criteria	Ground Truth	SR	SUN	SVM	MTSP	
AUC	Visual	0.7080	0.6524	**0.8486**	0.8269	
	Touch	0.6649	0.6185	**0.8745**	0.8422	
CC	Visual	0.2760	0.2311	**0.5988**	0.4234	
	Touch	0.2776	0.2265	0.4146	**0.5675**	

grates bottom-up (unsupervised) and top-down priors (supervised). As an unsupervised method, MTSP integrates multiple types of features collaboratively and treats saliency prediction as a sparsity pursuit problem [81]. Since MTSP considers cross-feature information, we will use it as the basic method in the following part to study middle-level category features. Details of MTSP and middle-level feature extraction will be discussed in the following part.

6.3.2 Enhancement of MTSP with Middle-level Category Features

This subsection introduces middle-level category features which measure the segment information. The middle-level category features is integrated with Multi-Task Sparsity Pursuit (MSTP) [81] model to boost the saliency detection performance. MSTP model is used as the basic model because it outperforms other unsupervised state-of-the-art models, as shown in Table 6.2, and it can combine different types of features efficiently. In the rest of this part, we will give a brief introduction of MTSP and a detailed introduction of the middle-level category features extraction.

Multi-Task Sparsity Pursuit (MTSP): A Review [81]

Image superpixels are used as basic image elements. Given an image, we first partition it into N superpixels. If the superpixel is judged to be salient, the pixels within this superpixel are salient. Since only multiple features case is considered, the problem can be formulated as follows.

Let $X_1, X_2, \cdots, X_K$ be K feature matrices corresponding to K types of features, where $X_k = [x_1, x_2, \cdots, x_N]$. The dimension of X_k is $d_k \times N$, where d_k is the dimension of k-th feature and N is the number of superpixels. x_i is the k-th feature vector correspond to an image superpixel P_i. The target here is to find an assignment function $S(P_i) \in [0, 1]$ by combining all the feature matrices $X_1, \cdots, X_K$. Here, the function $S(P_i)$ is used as saliency map.

Considering that the salient region should be different from the background region, the matrix X_k can be decomposed as:

$$X_k = X_k Z_k + E_k \tag{6.2}$$

where $X_k Z_k$ denotes the non-salient region which can be reconstructed by itself and Z_k is reconstruction coefficient. E_k denotes the salient region. We assume that only a small region of the image is salient, and E_k should be a sparse matrix.

To get the salient region $E_1, E_2, \cdots, E_K$ by integrating all the K features, the problem here is to seek the jointly sparse matrix E, $E = [E_1; E_2; \cdots; E_K]$, by solving the following convex optimization problem.

$$\begin{aligned} \min_{\substack{Z_1,\cdots,Z_K \\ E_1,\cdots,E_K}} \quad & \sum_{i=1}^{K} \|Z_i\|_* + \lambda \|E\|_{2,1} \\ \text{s.t.} \quad & X_i = X_i Z_i + E_i, i = 1, \cdots, K. \end{aligned} \tag{6.3}$$

The problem 6.3 is convex and can be solved efficiently [81]. We finally get the optimal solution $\{E_1^*, \cdots, E_K^*\}$ (with respect to E_i's). The saliency score for the i-th

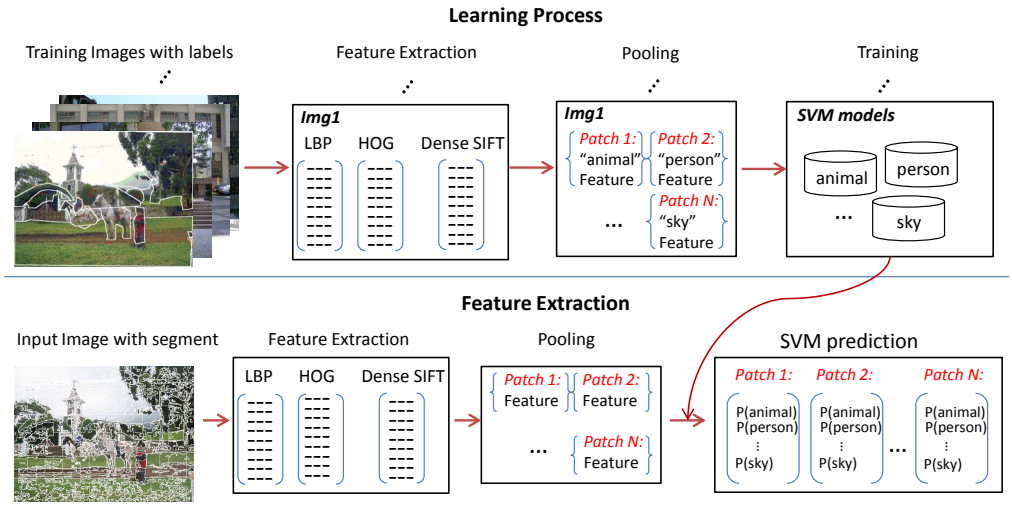


Figure 6.9: Middle-level category features extraction process. There are two processes: learning process and feature extraction. For learning process, given the learning images and labels, first extract the LBP, HOG and dense SIFT features for the images. For each labeled patch, the feature is extracted using second-order pooling. We train SVM model for each tag and get tag-SVM models. For feature extraction, given the image and the corresponded segments, we also extract the LBP, HOG and dense SIFT features, and get the segment features using second-order pooling. Then for each segment, concatenate the tag-SVM estimations for each tag as the middle-level category features.

superpixel P_i can be obtained by adding the ℓ_1-norm of the optimal solutions:

$$\mathrm{S}(P_i) = \sum_{j=1}^{K} \|E_j^*(:,i)\|_1 \tag{6.4}$$

where $\{E_1^*, \cdots, E_K^*\}$ denotes the optimal solution (with respect to E_i's). $\|E_j^*(:,i)\|_1$ is the ℓ_1-norm of the i-th column of E_j^*.

Middle-Level Category Features Extraction

This part introduces the middle-level category features extraction method. Here user-labeled segments are used as ground truth and a model for each label is learned. The superpixels are used as the basic elements, thus the features are extracted within the superpixels and the models are trained for superpixels which are labeled as the corresponded tags.

This feature extraction method is learning based. Firstly a SVM model is trained for each tag. Then, partition the given image into superpixels. For each superpixel, we concatenate the estimation results of each tag-SVM model as the middle-level category features.

- **Learning Process**: Given the image I, we label the image into K segments and represent the j-th segment as $(S_j^i, L_j^i), (j = 1, 2, ..., K)$. K is the number of segments. S_j^i and L_j^i are the pixels within the j-th segment and the label of the j-th segment respectively. LBP, HOG and dense SIFT feature are extracted for each image. The segments are represented using the second-order pooling [157]. A SVM model is trained for each tag and finally N (N is the number of labels) tag-SVM models could be obtained.
- **Middle-level Category Features Generalization**: Given an image, LBP, HOG and dense SIFT features are extracted first, and then segment the image into superpixels using SLIC [158]. A second-order pooling is used to get the final feature of each superpixel. Each superpixel could be represented by the possibility that it belongs to the n-th tag. Thus each superpixel can be represented as an N-D vector, N being the number of tags. And the n-th element of the N-D feature is the SVM estimation result of the n-th tag-SVM model. This N-D feature is the middle feature of the segment.

Experimental Settings

The 500 images within MIT dataset is used to do this experiment. For segment labeling, labels of the images are provided and undergraduate students are asked to label the segment regions manually. The labels include face, TV, sky, Sun, etc. To simplify the training and testing process, the labels are summarized into 17 tags based on WordNet [159], which is a large lexical database of English. These 17 tags are *abstraction*, *animal*, *artifact*, *body part*, *food*, *geological formation*, *ground*, *land*, *material*, *natural object*, *person*, *piece*, *plant*, *process*, *region*, *sky* and *water*.

In the learning process, for each tag, we use the segments labeled by participants to train tag-SVM model. While in the prediction part, we use the superpixels generated

from SLIC [158] as basic elements to predict the salient region. Instead of using that 500 images, we randomly select 400 images as training images to train tag-SVM models, and predict the middle-level category features for the rest 100 images. We only conduct saliency detection on these 100 testing images. We run five trials and report the average result (including AUC, ROC and CC), thus, the AUC and CC values for MTSP may be different from Table 6.2. The number of segments in the training set for each tag is between 10 to 900.

Besides the middle level feature, we choose three effective features: color, local energy [160] and local contrast [17] as the basic features.[1] Specially, we construct color feature by concatenating the six values of Red, Green and Blue channels as well as the probabilities of these three color channels. Besides, we compute RGB 3D color histograms of the image and filter them using a median filter at six different scales. For local energy, we use steerable pyramid filters as the feature. We use intensity, orientation and color contrast as local contrast feature.

In this experiment, the images are sized to 256×256 and the tradeoff parameter λ is set as 0.001.

Results and Analysis

Figure 6.10 shows some examples of the generated saliency maps. As can be seen here, saliency map obtained from middle-level category features is comparable with other maps. MTSP-MID integrates the F, LC, C and M maps efficiently and produces the best saliency map.

Figure 6.11 shows the original image, saliency ground truth and saliency maps predicted from MTSP and MTSP with middle-level feature (noted as MTSP-MID) and other baselines. Although the saliency detection results for MTSP and MTSP-MID are very similar to each other, we can still observe improvement after adding middle-level category features. For the first image (see the first sub-column), the salient region for MTSP and MTSP-MID are the cat's face, while for the other methods, the salient region is the

[1] Source codes of these features are available at http://people.csail.mit.edu //tjudd//WherePeopleLook//index.html.

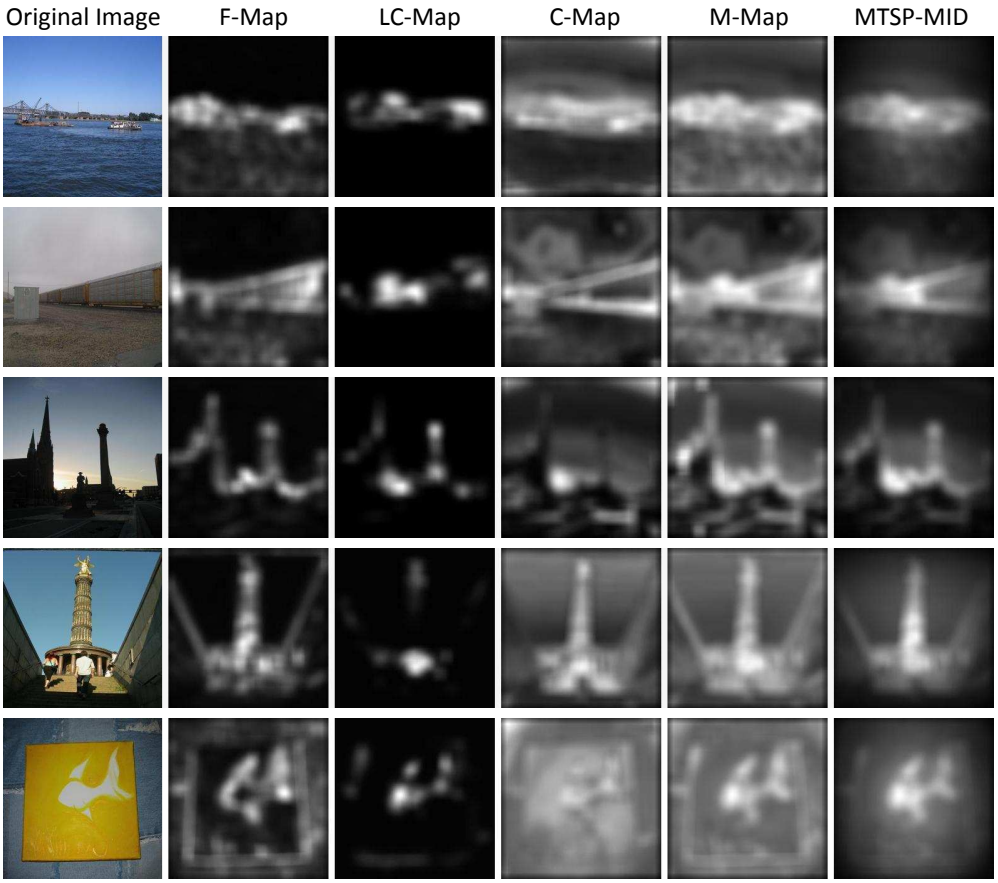


Figure 6.10: Some examples of integrating different types of features. From left to right: original image, the saliency map obtained from the local energy feature; the saliency map obtained from local contrast features; the saliency map obtained from color features; the saliency map obtained from middle-level category features; final saliency map.

forehead or even ear. Furthermore, MTSP-MID has more positive points than MTSP. We can also see similar results on other saliency maps. Note that only one feature is added to the existing three low-level features to boost the saliency detection process, so this improvement is reasonable.

The same as previous touch saliency study, we quantitatively evaluate the results by computing ROC, AUC (area under curve) and CC (Correlation Coefficient). Here the saliency maps are resized to the original image size. For ROC and AUC, we use threshold touch and visual fixation maps as ground truth; for CC, we use the original touch and visual fixation maps as ground truth.

Figure 6.12 shows the ROC curve of the saliency map produced from local con-

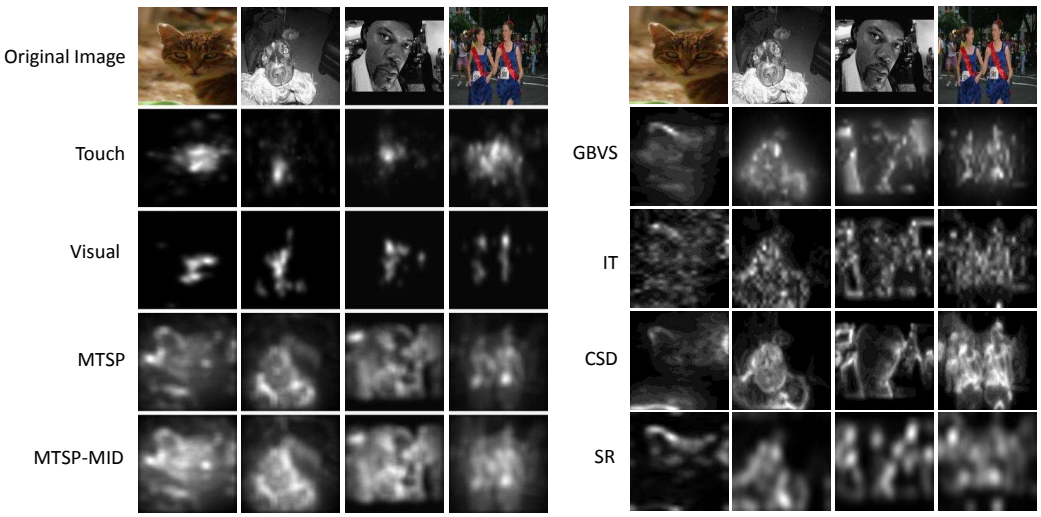


Figure 6.11: Some saliency prediction results. For the left column, from top to bottom are original image, touch fixation map, visual fixation map, saliency map produced by MTSP and MTSP-MID, respectively. For the right column, from top to bottom are original image, saliency map produced by GBVS, IT, CSD and SR, respectively. Note that we only compare GBVS, IT, CSD and SR because the other methods are outperformed by MTSP distinctly.

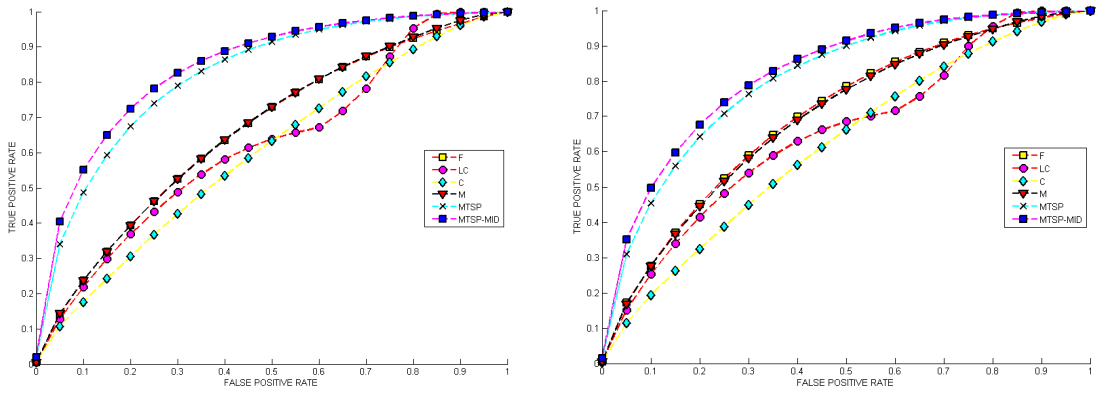


Figure 6.12: ROC curve of local contrast saliency map, color saliency map, local energy saliency map, middle-level saliency map, MTSP saliency map and MTSP-MID saliency map. Left figure shows ROC curve with thresholded touch ground truth, right figure shows ROC curve with thresholded visual ground truth.

Table 6.3: AUC and CC values for feature maps. For AUC, Ground Truth means using thresholded touch or visual fixation map as ground truth; for CC, Ground Truth means using original touch or visual fixation map as ground truth. Note that the AUC and CC value for MTSP is not the same as table 6.2 because we only predict saliency map on 100 testing images for five random trails.

Criteria	Ground Truth	F-Map	LC-Map	C-Map	M-Map	MTSP	MTSP-MID
AUC	Visual	0.7032	0.6551	0.6196	0.6992	0.8081	**0.8246**
	Touch	0.6618	0.6238	0.5998	0.6641	0.8230	**0.8457**
CC	Visual	0.2607	0.1675	0.1382	0.2437	0.3866	**0.4063**
	Touch	0.2526	0.1575	0.1484	0.2654	0.5214	**0.5697**

trast feature, color feature, local energy feature, middle-level category features, MTSP and MTSP-MID. Left figures shows the results with thresholded touch ground truth and right figure shows the results with thresholded visual ground truth. It can be seen that MTSP-MID and MTSP significantly outperform other saliency maps. It is consistent with the conclusion provided in [81] that MTSP can integrate different types of features efficiently. It can also be seen that the saliency map produced from middle-level category features has the similar performance as local energy saliency map, although the saliency maps are quite different, as shown in Figure 6.10. Through comparison of the AUC map between GT-Touch and GT-Visual, it can be seen that the saliency maps have the similar performance under these two ground truths.

The widely used AUC and CC results are shown in Table 6.3 to further investigate the difference. It can be seen that MTSP-MID outperforms MTSP under both touch and visual ground truths. Both MTSP-MID and MTSP performs better on touch ground truth.

This subsection compares the evaluation ability between touch saliency and visual saliency. Extensive studies show that touch saliency and visual saliency have comparable evaluation ability. We also propose middle-level category features to boost the MTSP saliency detection model, which called MTSP-MID. Saliency maps and quantitative study results show that MTSP-MID outperforms MTSP, and can boost the saliency detection model.

6.4 Summary

In this chapter, a larger touch saliency dataset with more users and more images is established. An iPhone interface which is similar to the previous study is built. And a criterion to choose good touch data is developed, which is quite useful in touch behavior study. Since not all images are enlarged during the data collection process, adding all the images without choosing will lead to very strong bias. We compare the generated touch fixations and some state-of-the-art saliency prediction results to evaluate the performance of touch saliency. The experimental results show that touch saliency is highly correlated visual saliency and can be used to study human attention, which is consistent with the preliminary study. Besides, we propose middle-level category features which measure the segment information. The state-of-the-art method MTSP is used as the basic model because it outperforms other state-of-the-art methods and can combine different types of features efficiently. The saliency prediction results of "use" and "do not use" middle-level category features are compared. ROC, AUC and CC results show that middle-level category features can improve the performance of saliency prediction. Again, the experimental results demonstrate that visual saliency and touch saliency have similar evaluation ability.

In future, we would like to build a touch profile for each user to study their touch habits. The touch habits can be used in many applications, such as smart advertisement, automatically enlarging image, etc. Better data chosen criteria can be proposed to generate a more reliable touch saliency map.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

7.1.1 Dynamic Captioning

In this work, a dynamic captioning system which helps hearing impaired audience understand video better was described. Specifically, this system was built to help hearing impaired audience recognize the speakers more easily and perceive and track the video scripts better. To help hearing impaired audience better recognize the speakers, dynamic captioning puts scripts at suitable position. Meanwhile, the scripts are synchronously highlighted and the variation of voice volume is illustrated. These functions could help hearing impaired individuals better perceive and track the scripts. A user study on 60 real hearing impaired participants was conducted to evaluate this system. User study results show that this system is effective. The audience can better understand the video with the proposed dynamic captioning process.

7.1.2 Image Re-emotionalizing

In this work, a brand new emotion synthesis framework, namely image re-emotionalizing, was developed. It transfers the image emotion from the original emotion to the user in-

put emotion. Figure 4.6 shows some transformed results. It can be seen that these transformed images are more natural than color transferring baseline results. A user study on 20 participants was conducted to quantitatively evaluate these results. User study results show that the proposed framework is effective and the transformed images are realistic and natural. Compared with the baseline method, the proposed method transforms the image automatically without choosing any reference image.

7.1.3 Learning to Photograph

In this work, an intelligent photography system which recommends the most user-favored view rectangle for arbitrary camera input was proposed. A user study to study user preference was conducted to quantitatively evaluate the system. Figure 5.8 shows the experimental results as well as the user ratings. It can be seen that the proposed method outperforms existing state-of-the-art methods. User ratings show that users prefer the images generated from the proposed method to other state-of-the-art methods.

7.1.4 Touch Saliency

In this work, how to collect user touch data on touch devices when browsing images was studied. The touch data is used to generate an alternative visual saliency ground truth, touch saliency. To evaluate the performance, qualitative and quantitative study of touch saliency map were investigated. Also, a user study on users who participated in both touch and visual saliency collection process was conducted to study user experience. The study of touch saliency map shows that touch saliency is highly correlated with human visual saliency ground truth. And the user study results show that touch data collection process is much more interesting, much more comfortable and less tiring than visual saliency collection process. These results show that touch saliency can be naturally used as visual saliency in the same way. Compared with visual saliency ground truth, touch saliency ground truth is much easier to obtain. Thus it is possible to create a large touch saliency database. Moreover, we proposed middle-level category features to enhance the saliency detection process. Extensive experiments demonstrate the effectiveness of the proposed middle-level category feature. This study opens up a new research direction

of saliency study by analyzing human touch information on touch smart mobile devices which are increasingly popular.

7.2 Future Work

7.2.1 Dynamic Captioning

To our knowledge, it is the first work to help hearing impaired audience better access videos, lots works along this direction could be done in future. One possible future work is to design a better interface. User interface design which is important for real-world application is not mentioned in this work. That is because it is not a crucial problem here since even with a simple user interface, this system is already able to help hearing impaired individuals access videos. But a better interface could further improve the proposed system. Another possible future work is to extend this system to process video without script by employing speech recognition engine and user speaker clustering [103, 104] and identification [105, 106] methods. Moreover, the script-face mapping component could be improved to increase the accuracy of face identification in future. Finally, a larger dataset should be built to further investigate the effectiveness of this system.

7.2.2 Image Re-emotionalizing

As the first emotional image synthesizing work, this study shows a new direction of emotional image study, and could be utilized to provide better service to users in social networking multimedia. It should be noted that only 4 emotions instead of 8 emotions were studied, because user labeling result shows that these 4 emotions (including two negative and two positive emotions) are able to describe most images. In this work, color features were explored to transform the image because color features are straight forward and easy to transfer. The results are already better than baseline result. Although the transformed images are natural, it is necessary to consider other efficient features to further improve the performance. It might not be easy; however, with current emotional

image study, it is possible to find suitable features which could produce natural image with better emotion expression. Note that in this experiment, only landscape images were studied. This disadvantage is small because other kinds of images can be studied using the same approach. In future, it is necessary to implement the proposed method on a larger dataset which includes variety types of images. This synthesizing system could be used in online chatting applications to enhance user experience during online chatting.

7.2.3 Learning to Photograph

Note that in this work, the images were classified into different categories and a model was generated for each category. Thus in future, a more general model could be developed by considering other important photography factors such as contrast, exposure, etc. Some photographic compositional rules can also be considered.

7.2.4 Touch Saliency

It would be possible to build a large dataset and study human touch habit in future. Other saliency generation methods should also be studied. By collecting persons touch fixations, it is possible to study human touch behavior and build touch profile for individual users. This touch profile could be used in social networking sites to utilize personalized searching or recommendation. Considering that touch mobile devices are increasingly popular, it should be useful to explore more applications which combine both touch and multimedia study.

List of Publications

1) Tam Nguyen*, **Mengdi Xu***, Guangyu Gao, Mohan Kankanhalli, Qi Tian, Shuicheng Yan, "Static Saliency vs. Dynamic Saliency: A Comparative Study", ACM Multimedia(ACM MM), 2013.

2) Bingbing Ni, **Mengdi Xu**, Bin Cheng, Meng Wang, Shuicheng Yan, Qi Tian, "Learning to Photograph: a Compositional Perspective", IEEE Transactions on Multimedia (TMM), 2013.

3) **Mengdi Xu**, Bingbing Ni, Jian Dong, Zhongyang Huang, Meng Wang, Shuicheng Yan, "Touch Saliency", ACM Multimedia(ACM MM), 2012. (Short paper)

4) Bingbing Ni*, **Mengdi Xu***, Jinhui Tang, Shuicheng Yan, Pierre Moulin, "Omni-Range Spatial Context for Visual Recognition", Computer Vision and Pattern Recognition (CVPR), 2012.

5) **Mengdi Xu**, Bingbing Ni, Jinhui Tang, Shuicheng Yan, "Image Re-emotionalizing", Pacific-Rim Conference on Multimedia(PCM), 2011. [Best Paper Award]

6) Richang Hong, Meng Wang, Xiao-Tong Yuan, **Mengdi Xu**, Jianguo Jiang, Shuicheng Yan, Tat-Seng Chua, "Video accessibility enhancement for hearing-impaired users", TOMCCAP, 2011.

7) Richang Hong, Meng Wang, **Mengdi Xu**, Shuicheng Yan, Tat-Seng Chua, "Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment", ACM Multimedia, Firenze, Italy, 2010. [Best Paper Award]

*indicates equal contribution.

8) **Mengdi Xu**, Xiao-Tong Yuan, Jialie Shen, Shuicheng Yan, ”Cast2Face: Character Identication in Movie with Actor-Character Correspondence”, ACM Multimedia, Firenze, Italy, 2010. (Short paper)

9) Richang Hong, Xiao-Tong Yuan, **Mengdi Xu**, Meng Wang, Shuicheng Yan, Tat-Seng Chua, ”Movie2Comics: A Feast of Multimedia Artwork”, ACM Multimedia, Firenze, Italy, 2010. (Short paper)

List of Awards

1) 2010 ACM Multimedia 2010 Best Paper Award

2) 2011 Pacific-Rim Conference on Multimedia 2011 Best Paper Award

Bibliography

[1] Grudin, J.: Three faces of human-computer interaction. Annals of the History of Computing **27** (2005) 46–62

[2] Sears, A., Jacko, J.A.: The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications. (2007)

[3] Barrett, R., Maglio, P.P., Kellem, D.C.: How to personalize the web. In: Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. (1997) 75–82

[4] Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. Internet Computing **7** (2003) 76–80

[5] Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the 2008 ACM conference on Recommender systems. (2008) 259–266

[6] Anderson, C.R., Domingos, P., Weld, D.S.: Personalizing web sites for mobile users. In: Proceedings of the 10th international conference on World Wide Web. (2001) 565–575

[7] Merialdo, B., Lee, K.T., Luparello, D., Roudaire, J.: Automatic construction of personalized tv news programs. In: Proceedings of the seventh ACM international conference on Multimedia (Part 1). (1999) 323–331

[8] Yu, Z., Zhou, X.: Tv3p: an adaptive assistant for personalized tv. IEEE Transactions on Consumer Electronics **50** (2004) 393–399

[9] Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized tv. Signal Processing Magazine **23** (2006) 90–100

[10] Nielsen(ed), J.: Advances in human-computer interaction. Volume 5. Intellect Publishers (1995)

[11] Gulliver, S., Ghinea, G.: How level and type of deafness affect user perception of multimedia video clips. Universal Access in the Information Society **2** (2003) 374–386

[12] Bianchi-Berthouze, N.: K-dime: an affective image filtering system. IEEE Transactions on Multimedia **10** (2003) 103–106

[13] Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: ACM International Conference on Multimedia. (2010) 83–92

[14] Gooch, B., Reinhard, E., Moulding, C., Shirley, P.: Artistic composition for image creation. In: Eurographics Workshop on Rendering. (2001) 83–88

[15] Liu, L., Chen, R., Wolf, L., Cohen-Or, D.: Optimizing photo composition. Computer Graphics Forum **29** (2010) 469–478

[16] MARTINEZ, B., BLOCK, J.: Visual Forces, an Introduction to Design. Prentice-Hall, New York (1998)

[17] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259

[18] Bruce, N.D., Tsotsos, J.K.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems. (2006)

[19] Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): Meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010)

[20] Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., Chua, T.S.: An eye fixation database for saliency detection in images. In: European Conference on Computer Vision, Crete, Greece (2010)

[21] Hong, R., Wang, M., Xu, M., Yan, S., Chua, T.S.: Dynamic captioning: Video accessibility enhancement for hearing impairment. In: ACM International Conference on Multimedia. (2010)

[22] Hong, R., Wang, M., Yuan, X.T., Xu, M., Jiang, J., Yan, S., Chua, T.S.: Video accessibility enhancement for hearing impaired users. ACM Transactions on Multimedia Computing, Communications, and Applications **7S** (2011) 24–42

[23] Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B.: Automatic image retargeting. In: In Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia(MUM). (2005)

[24] Wang, M., Sheng, Y., Liu, B., Hua, X.S.: In-image accessibility indication. IEEE Transactions on Multimedia **12** (2010) 330–336

[25] Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision **9** (2009) 1–24

[26] Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: International Conference on Computer Vision. (2009)

[27] Amazon: Amazon mechanical turk. (https://www.mturk.com/mturk/welcome)

[28] TV: A brief history of captioned television. (http://www.ncicap.org/caphist.asp)

[29] Boyd, J., Vader, E.: Captioned television for the deaf. American Annals of the Deaf **117** (1972) 34–7

[30] Braverman, B.B., Hertzog, M.: The effects of caption rate and language level on comprehension of a captioned video presentation. American Annals of the Deaf **125** (1980) 943–48

[31] Lewis, M.S.J., Jackson, D.W.: Television literacy: Comprehension of program content using closed captions for the deaf. Journal of Deaf Studies and Deaf Education **6** (2001) 43–53

[32] Garrison, W., Long, G., Dowaliby, F.: Working memory capacity and comprehension processes in hearing impaired reader. Journal of Deaf Studies and Deaf Education **2** (1997) 78–94

[33] Gulliver, S., Ghinea, G.: Impact of captions on deaf and hearing perception of multimedia video clips. In: International Conference on Multimedia and Expo. Volume 1. (2002) 753–756

[34] Chang, Y., Saito, S., Nakajima, M.: Example-based color transformation of image and video using basic color categories. IEEE Transactions on Image Processing **16** (2007) 329–336

[35] Gupta, M.R., Upton, S., Bowen, J.: Simulating the effect of illumination using color transformation. SPIE CCI **111** (2005) 248–258

[36] Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. CGA **21** (2001)

[37] Thompson, W.B., Shirley, P., Ferwerda, J.A.: A spatial post-processing algorithm for images of night scenes. Journal of Graphics Tools **7** (2002) 1–12

[38] Guo, Y., Yu, J., Xu, X., Wang, J., Peng, Q.: Example based painting generation. CGI **7** (2006) 1152–1159

[39] Zhang, X., Constable, M., He, Y.: On the transfer of painting style to photographic images through attention to colour contrast. In: Pacific-Rim Symposium on Image and Video Technology. (2010) 414–421

[40] Hong, R., Yuan, X., Xu, M., Wang, M., Yan, S., Chua, T.S.: Movie2comics: A feast of multimedia artwork. In: ACM International Conference on Multimedia. (2010) 611–614

[41] Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)

[42] Sun, X., Yao, H., Ji, R., Liu, S.: Photo assessment based on computational visual attention model. In: ACM International Conference on Multimedia. (2009) 541–544

[43] Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B.: Automatic image retargeting. In: ACM International Conference on Mobile and Ubiquitous Multimedia. (2005) 59–68

[44] Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. ACM Transactions on Graphics **26** (2007) 1–9

[45] Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. ACM Transactions on Graphics (SIGGRAPH) **27** (2008) 1–9

[46] Rubinstein, M., Shamir, A., Avidan, S.: Multi-operator media retargeting. ACM Transactions on Graphics (SIGGRAPH) **28** (2009) 1–11

[47] Wolf, L., Guttmann, M., Cohen-or, D.: Non-homogeneous content-driven video-retargeting. In: International Conference on Computer Vision. (2007)

[48] Wang, Y.S., Tai, C.L., Sorkine, O., Lee, T.Y.: Optimized scale-and-stretch for image resizing. ACM Transactions on Graphics (SIGGRAPH ASIA) **27** (2008)

[49] Cho, T.S., Butman, M., Avidan, S., Freeman, W.T.: The patch transform and its applications to image editing. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)

[50] Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)

[51] Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics (SIGGRAPH) **28** (2009)

[52] Guo, Y., Liu, F., Shi, J., Zhou, Z.H., Gleicher, M.: Image retargeting using mesh parametrization. IEEE Transactions on Multimedia **11** (2009) 856–867

[53] Luo, Y., Tang, X.: Photo and video quality evaluation: Focusing on the subject. In: European Conference on Computer Vision. (2008) 386–399

[54] Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: IEEE Conference on Computer Vision and Pattern Recognition. (2006) 419–426

[55] Sheikh, H., Bovik, A., Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Transactions on Image Processing **14** (2005) 2117–2128

[56] Luo, W., Wang, X., Tang, X.: Content-based photo quality assessment. In: International Conference on Computer Vision. (2011)

[57] Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: International Conference on Computer Vision. (2011)

[58] Zheng, Y.T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009)

[59] Ji, R., Xie, X., Yao, H., Ma, W.Y.: Mining city landmarks from blogs by graph modeling. In: ACM International Conference on Multimedia. (2009) 105–114

[60] Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.M., Pang, Y., Zhang, L.: Equip tourists with knowledge mined from travelogues. In: International World Wide Web Conference. (2010)

[61] Ni, B., Song, Z., Yan, S.: Web image mining towards universal age estimator. In: ACM International Conference on Multimedia. (2009) 85–94

[62] Ni, B., Song, Z., Yan, S.: Web image and video mining towards universal and robust age estimator. IEEE Transactions on Multimedia (2011)

[63] Cheng, B., Ni, B., Yan, S., Tian, Q.: Learning to photograph. In: ACM International Conference on Multimedia. (2009)

[64] Liu, T., Sun, J., Zheng, N.N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)

[65] Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., Oliva, A.: Modeling search for people in 900 scenes. Visual Cognition **17** (2009) 945–978

[66] van der Linde, I., Rajashekar, U., Bovik, A.C., Cormack, L.K.: Doves: A database of visual eye movements. Spatial Vision **22(2)** (2009) 161–177

[67] Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model the bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006)

[68] Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 545–552

[69] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009) 1597–1604

[70] Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: European Conference on Computer Vision. (2012) 414–429

[71] Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. MIT tech report (2012)

[72] Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 802–817

[73] Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12** (1980) 97–136

[74] Gopalakrishnan, V., Hu, Y., Rajan, D.: Random walks on graphs for salient object detection in images. IEEE Transactions on Image Processing **19** (2010) 3232–3242

[75] Hu, Y., Rajan, D., Chia, L.T.: Detection of visual attention regions in images using robust subspace analysis. Journal of Visual Communication and Image Representation **19** (2008) 199–216

[76] Li, J., Levine, M.D., An, X., Xu, X., He, H.: Visual saliency based on scale-space analysis in the frequency domain. IEEE Transactions on Pattern Analysis and Machine Intelligence (In press) (2013)

[77] Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1915–1926

[78] Itti, L., Koch, C.: A comparison of feature combination strategies for saliency-based visual attention systems. SPIE human vision and electronic imaging IV **3644** (1999) 373–382

[79] Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging **10** (2001) 161–169

[80] Navalpakkam, V., Itti, L.: Search goal tunes visual features optimally. Neuron **53** (2007) 605–617

[81] Lang, C., Liu, G., Yu, J., Yan, S.: Saliency detection by multitask sparsity pursuit. IEEE Transactions on Image Processing **21** (2012) 1327–1338

[82] Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. Advances in Neural Information Processing Systems **20** (2008)

[83] Deaf: (http://en.wikipedia.org/wiki/deaf)

[84] Xu, M., Jin, J.S., Luo, S., Duan, L.: Hierarchical movie affective content analysis based on arousal and valence features. In: ACM International Conference on Multimedia. (2008) 677–680

[85] Emotion. (http://www.scholarpedia.org/article/Speech emotion analysis)

[86] Moreno, P.J., Joerg, C., Van Thong, J.M., Glickman, O.: A recursive algorithm for the forced alignment of very long audio segments. In: International Conference on Spoken Language Processing. Volume 8. (1998)

[87] Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy– automatic naming of characters in tv video. In: British Machine Vision Conference. (2006)

[88] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. (2001)

[89] Yang, T., Pan, Q., Li, J., Li, S.: Real-time multiple objects tracking with occlusion handling in dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005)

[90] Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J., Darrell, T.: Visual speech recognition with loosely synchronized feature streams. In: International Conference on Computer Vision. Volume 2. (2005) 1424–1431

[91] Wang, M., Hua, X.S., Tang, J., Hong, R.: Beyond distance measurement: constructing neighborhood similarity for video annotation. IEEE Transactions on Multimedia **11** (2009) 465–476

[92] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009) 210–227

[93] Wang, M., Hua, X.S., Hong, R., Tang, J., Qi, G.J., Song, Y.: Unified video annotation via multigraph learning. IEEE Transactions on Circuits and Systems for Video Technology **19** (2009) 733–746

[94] Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing **20** (2010) 231–252

[95] Arandjelovic, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 1. (2005) 860–867

[96] Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. SIAM (2008)

[97] Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. In: ACM International Conference on Multimedia. (2003) 374–381

[98] Wang, M., Zhang, H.J.: Video content structuring. Scholarpedia **4** (2009) 9431

[99] Daelemans, W., van den Bosch, A.: Tabtalk: Reusability in data-oriented grapheme-to-phoneme conversion.'. In: Proceedings of Eurospeech. Volume 93. (1993) 1459–1466

[100] Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., Lee, K.F., Rosenfeld, R.: The SPHINX-II speech recognition system: an overview. (1992)

[101] Xu, M., Chia, L.T., Yi, H., Rajan, D.: Affective content detection in sitcom using subtitle and audio. In: International Conference on MultiMedia Modeling. (2006)

[102] Fisher, R.A., Genetiker, S., Fisher, R.A., Genetician, S., Britain, G., Fisher, R.A., Généticien, S.: Statistical methods for research workers. Volume 14. (1970)

[103] Ajmera, J., Wooters, C.: A robust speaker clustering algorithm. In: IEEE Workshop on Automatic Speech Recognition and Understanding. (2003) 411–416

[104] Stadelmann, T., Freisleben, B.: Unfolding speaker clustering potential: a biomimetic approach. In: ACM International Conference on Multimedia. (2009) 185–194

[105] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital signal processing **10** (2000) 19–41

[106] Wan, V., Campbell, W.M.: Support vector machines for speaker verification and identification. In: Proceedings of the 2000 IEEE Signal Processing Society Workshop. Volume 2. (2000) 775–784

[107] Colombo, C., Bimbo, A.D., Pala, P.: Semantics in visual information retrieval. IEEE Transactions on Multimedia **6** (1999) 38–53

[108] Itten, J.: The art of color: the subjective experience and objective rationale of color. John Wiley, New York (1973)

[109] Wang, W.N., Yu, Y.L., Jiang, S.M.: Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: IEEE International Conference on Systems, Man and Cybernetics. (2006) 3534–3539

[110] Hayashi, T., Hagiwara, M.: Image query by impression words-the iqi system. TCE **44** (1998) 347–352

[111] Wu, Q., Zhou, C., Wang, C.: Content-based affective image classification and retrieval using support vector machines. Affective Computing and Intelligent Interaction (2005)

[112] Yanulevskaya, V., van Gemert, J.C., Roth, K., Herbold, A.K., Sebe, N., Geusebroek, J.M.: Emotional valence categorization using holistic image features. In: IEEE International Conference on Image Processing. (2008)

[113] Cho, S.B.: Emotional image and musical information retrieval with interactive genetic algorithm. In: Proceedings of the IEEE. (2004) 702–711

[114] Hanjalic, A.: Extracting moods from pictures and sounds: towards truly personalized tv. IEEE Signal Processing Magazine **23** (2006) 90–100

[115] Wang, H.L., Cheong, L.F.: Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology **16** (2006) 689–704

[116] Cheng, B., Ni, B., Yan, S., Tian, Q.: Learning to photograph. In: ACM International Conference on Multimedia. (2010) 291–300

[117] Lang, P., Bradley, M.M., Cuthbert, B.N.: International affective picture system (iaps): Affective ratings of pictures and instruction manual. In: Technical Report A-8, University of Florida , Gainesville, FL. (2008)

[118] Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: Nus-wide: A real-world web image database from national university of singapore. In: CIVR. (2009)

[119] Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal of Computer Vision **59** (2004) 167–181

[120] Stricker, M., Orengo, M.: Similarity of color images. In: SPIE. (1995) 381–392

[121] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005) 886–893

[122] Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)

[123] Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21** (2001)

[124] Zhang, X., Constable, M., He, Y.: On the transfer of painting style to photographic images through attention to colour contrast. In: Pacific-Rim Symposium on Image and Video Technology. (2010) 414–421

[125] Chang, Y., Saito, S., Nakajima, M.: Example-based color transformation of image and video using basic color categories. IEEE Transactions on Image Processing **16** (2007) 329–336

[126] Wang, Y., Liu, Z., Hua, G., Wen, Z., Zhang, Z., Samaras, D.: Face re-lighting from a single image under harsh lighting conditions. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)

[127] Rules: 10 top photography composition rules. (http://www.photographymad.com/pages/view/10-top-photography-composition-rules)

[128] Kowalski, M.A., Hughes, J.F., Rubin, C.B., Ohya, J.: User-guided composition effects for art-based rendering. In: Symposium on Interactive 3D graphics. (2001) 99–102

[129] Byers, Z., Dixon, M., Smart, W.D., Grimm, C.M.: cheese!: Experiences with a robot photographer. In: Innovative Applications of Artificial Intelligence Conference. (2003) 65–70

[130] Lok, S., Feiner, S., Ngai, G.: Evaluation of visual balance for automated layout. In: International Conference on Intelligent User Interfaces. (2004) 101–108

[131] Zhang, M.: Auto cropping for digital photographs. In: International Conference on Multimedia and Expo. (2005)

[132] Santella, A., Agrawala, M., Decarlo, D., Salesin, D., Cohen, M.F.: Gaze-based interaction for semi-automatic photo cropping. In: International Conference on Computer Human Interaction. (2006)

[133] Nishiyama, M., Okabe, T., Sato, Y., Sato, I.: Sensation-based photo cropping. In: ACM International Conference on Multimedia. (2009) 669–672

[134] Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **3** (1973) 610–621

[135] Birchfield, S., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2005) 1158–1163

[136] Li, J., Wu, W., Wang, T., Zhang, Y.: One step beyond histogram: Image representation using markov stationary features. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)

[137] Zheng, Y., Zhao, M., Neo, S.Y., Chua, T.S., Tian, Q.: Visual synset: towards a higher-level visual representation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)

[138] Ni, B., Yan, S., Kassim, A.: Contextualizing histogram. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009)

[139] Fei-fei, L.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005) 524–531

[140] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005) 886–893

[141] Feng, J., Ni, B., Tian, Q., Yan, S.: Geometric p-norm feature pooling for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011)

[142] Shapiro, L.G., Stockman, G.C.: Computer Vision. Prentice Hall (2003)

[143] Comaniciu, D., Meer, P., Member, S.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 603–619

[144] Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal of Computer Vision **2** (2004) 167–181

[145] Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: International Conference on Computer Vision. (2009)

[146] Viola, P., Jones, M.: Robust real-time object detection. International Journal on Computer Vision **75** (2004) 137–154

[147] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010)

[148] Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B **39** (1977) 1–38

[149] Gilks, W., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics. Chapman and Hall/CRC (1996)

[150] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012)

[151] King, B.M., Minium, E.W.: Statistical Reasoning in Psychology and Education. New York: Wiley (2003)

[152] Xu, M., Ni, B., Dong, J., Huang, Z., Wang, M., Yan, S.: Touch saliency. In: ACM International Conference on Multimedia. (2012)

[153] Tatler, B.W., Baddeley, R.J., Gilchrist, I.D., et al.: Visual correlates of fixation selection: Effects of scale and time. Vision research **45** (2005) 643–659

[154] Ouerhani, N., Von Wartburg, R., Hugli, H., Muri, R.: Empirical validation of the saliency-based model of visual attention. Electronic letters on computer vision and image analysis **3** (2004) 13–24

[155] Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8

[156] Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision **8** (2008)

[157] Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: European Conference on Computer Vision. (2012)

[158] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. École Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep **149300** (2010)

[159] Miller, G.A., et al.: Wordnet: a lexical database for english. Communications of the ACM **38** (1995) 39–41

[160] Simoncelli, E.P., Freeman, W.T.: The steerable pyramid: A flexible architecture for multi-scale derivative computation. In: IEEE International Conference on Image Processing. Volume 3. (1995) 444–447