

COMPUTATIONAL MEDIA AESTHETICS
FOR MEDIA SYNTHESIS

XIANG YANGYANG

(B.Sci., Fudan Univ.)

A THESIS SUBMITTED
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR
INTEGRATIVE

SCIENCES AND ENGINEERING

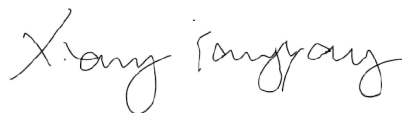
NATIONAL UNIVERSITY OF SINGAPORE

2013

DECLARATION

I hereby declare that this 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.



XIANG YANGYANG

January 2014

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor Professor Mohan Kankanhalli for his continuous support during my Ph.D study. His patience, enthusiasm, immense knowledge and guidance helped me throughout the research and writing of this thesis.

I would like to thank my Thesis Advisory Committee members: Prof. Chua Tat-Seng, and Dr. Tan Ping for their insightful comments and questions.

I also want to thank all the team members of the Multimedia Analysis and Synthesis Laboratory, without whom the thesis would not have been possible at all.

Last but not the least, I would like to express my appreciation to my family. They have spiritually supported and encouraged me through the whole process.

ABSTRACT

Aesthetics is a branch of philosophy and is closely related to the nature of art. It is common to think of aesthetics as a systematic study of beauty, and one of its major concerns is the evaluation of beauty and ugliness. Applied media aesthetics deals with basic media elements, and aims to constitute formative evaluations as well as help create media products. It studies the functions of basic media elements, provides a theoretical framework that makes artistic decisions less arbitrary, and facilitates precise analysis of the various aesthetic parameters.

Aesthetic assessment and aesthetic composition are two aspects of computational media aesthetics. The former one aims to evaluate the aesthetic level of a given media piece and the latter aims to produce media outputs based on computational aesthetic rules. In this dissertation, we focus on media synthesis, and exhibit how media aesthetics could help improve the efficiency and quality of media production.

First, we present an algorithm that can successfully improve the quality of hazy images and offer visually-pleasant haze-free results with vivid colors. The notion of “vivid colors” is related to the visual quality from an aesthetic point of view. We propose a full-

saturation assumption (FSA) based on the aesthetic photographic effect: photos of vivid colors are visually pleasant and first recover the degraded saturation layer. The depth image is also obtained as a by-product. Experimental results are compared with those of other dehazing approaches, and a synthesis-based test is also performed.

Second, we present a novel automatic image slideshow system that explores a new medium between images and music. It can be regarded as a new image selection and slideshow composition criterion. Based on the idea of "hearing colors, seeing sounds" from the art of music visualization, equal importance is assigned to image features and audio properties for better synchronization. We minimize the aesthetic energy distance between visual and audio features. Given a set of images, a subset is selected by correlating image features with the input audio properties. The selected images are then synchronized with the music subclips by their audio-visual distance. We perform a subjective user study to compare our results with those generated by other techniques. Slideshows based on audio pieces of different valence are also proposed for comparison.

Then we present an automated post-processing method for home

produced videos based on frame “interestingness”. The input single video clip is treated as a long take, and film editing operations for sequence shot are performed. The proposed system automatically adjusts the distribution of interestingness, both spatially and temporally, in the video clip. We use the idea of video retargeting to introduce fake camera work and manipulate spatial interestingness, then we perform video re-projection to introduce motion rhythm and modify the temporal distribution of interestingness. User study is carried out to evaluate the quality of the testing results.

We also present a web page advertisement selection strategy based on the force model. It refines the results of contextual advertisement selection by introducing aesthetic criteria. The web page is semantically segmented into blocks, and each block is an element in the two-dimensional screen. Aesthetic theories on the screen balancing are adopted in the proposed system. We compute the graphic weights of blocks and treat them as vertices in a graph. Weighted graph edges are the forces between the elements. The aesthetically optimal advertisement is the one that balances the force system. We invite users to compare our proposed scheme and the random advertisement selection strategy.

Contents

1	Introduction	3
1.1	Aesthetics and Applied Media Aesthetics	3
1.2	Methodology of Applied Media Aesthetics	5
1.3	Aesthetic Elements	7
1.4	Scope and Contributions	11
1.4.1	Aim	11
1.4.2	Approach	12
1.4.3	Contribution	12
1.5	Summary	13
1.6	Thesis Overview	14
2	Previous Work	17
2.1	Features that Represent Aesthetics	17
2.1.1	Object Position	20
2.1.2	Spatial Features	21
2.1.3	Motion	22
2.1.4	Composition and Object Detection	27
2.1.5	Audio	29
2.1.6	Fusion	30
2.2	The Applications of Multimedia Aesthetics	32
2.2.1	Aesthetic Evaluation	32
2.2.2	Aesthetic Enhancement	53
2.3	Discussions	58
3	Single Image Aesthetics: Hazy Image Enhancement based on the Full-Saturation Assumption	61
3.1	Introduction	62
3.2	Previous Work	64
3.3	The HSI Color Space and the Dehazing Problem	66
3.4	Full-Saturation Assumption	69
3.5	Relations with Dark Channel Prior	69
3.6	Our Example-based Approach	73
3.7	Experimental Results	75
3.8	Discussions	83

4	Aesthetics for Image Ensembles: A Synaesthetic Approach for Image Slideshow Generation	87
4.1	Introduction	88
4.2	Previous Work	91
4.3	Color and Sound Matching	93
4.3.1	Aesthetic Energy of Images	94
4.3.2	Aesthetic Energy of Audio	100
4.3.3	Color-Sound Matching	104
4.4	Our Photo SlideShow	107
4.4.1	Image Pre-Selection	107
4.4.2	Audio-Image Mapping	108
4.4.3	Image Saliency	110
4.4.4	Camera Work	111
4.4.5	Transition	116
4.5	Experimental Results	116
4.5.1	Scheme Comparison	117
4.5.2	Comparison between Different Input Audio	119
4.5.3	Comparison with the previous results	120
4.6	Discussions	121
5	Videos Aesthetics: Automatic Retargeting and Reprojection for Editing Home Videos	123
5.1	Introduction	123
5.2	Previous Work	127
5.3	Our Approach	131
5.3.1	Frame Saliency	132
5.3.2	Subclip Segmentation	136
5.3.3	Retargeting, Reprojection and The Fusion	138
5.3.4	Frame Re-Rendering	140
5.4	Experimental Results	142
5.5	Discussions	146
6	Aesthetics for Non-Traditional Medium: Force-Model Based Aesthetic Online Advertisement Selection	149
6.1	Introduction	150
6.2	Previous Work	153
6.3	Aesthetic Advertising	157
6.4	Our Approach	159
6.4.1	Visual Weights of Elements	160
6.4.2	Force-based System Formulation	164
6.4.3	An Optimization-based Solution	168

6.5	Experimental Results	172
6.6	Conclusion	176
7	Conclusion	179
7.1	Summary of The Dissertation	179
7.1.1	Aesthetics for Single Image	180
7.1.2	Aesthetics for Multiple Images	180
7.1.3	Aesthetics for Videos	181
7.1.4	Aesthetics for Online Advertising	181
7.2	Conclusions	182
7.2.1	Future Direction	185
	Bibliography	188
	Bibliography	189

List of Figures

1.1	Dominant colors. The left image(<i>The Twilight City</i> (2009)) has a cold dominant color and it delivers the feeling of grief. The right image (<i>Sherlock Holmes</i> (2009)) has a warmer dominant color. It implies the cheerfulness of the lucky survival.	8
1.2	Different horizons suggest different natures of the whole scene. The horizontal camera view gives a stable scene while the right images has an unstable horizon, and it exaggerates the feeling of speed.	8
1.3	Different shot points. The left image uses a horizontal angle, and it shows the sense of sacred. The middle image is taken from the side face. It emphasizes the continuity between buildings. The right image is taken from below, and it highlights the height and impact of the skyscraper.	9
2.1	The statistic scoring results of ACQUINE [DW10].	40
2.2	The extracted features of Chen et al. [LC09]	43
2.3	A summary of the extracted aesthetic features in the media assessing systems.	54
3.1	The left shows an image free of haze. The right one is taken on a foggy day and degraded by haze.	62
3.2	A sample natural image of vivid color. (a). The natural image. (b). The saturation layer.	70
3.3	Distribution of local maximum saturation. (a). The natural outdoor scene. (b). Indoor objects with post-processed color effects.	70
3.4	Color saturation under different Intensity.	72
3.5	Haze removal result. (a) Input hazy image. (b) The saturation layer of the original image in the HSI color space. (c) The initial downsampled transmission map. (d) The corresponding pixel index of downsampled transmission map in the up-sampled map. The joint bilateral filter is performed on (d), and the estimated transmission map is shown in (e). (f) The saturation layer of the dehazed image. (g) The output haze-free image.	77
3.6	Haze removal results. First column: input hazy images. Second column: the transmission map. Third column: Output haze-free images.	78

3.7	Comparison with He et al's work [HST09]. (a) The input hazy image. (b) Dark channel prior. (c) Our result.	79
3.8	Comparison with others' work. (a) The input hazy image. (b) Fattal's result [Fat08]. (c) Dark channel prior [HST09]. (d) Our result.	80
3.9	More comparisons with other work. (a) The input hazy image. (b) Our results. (c) Fattal's results [Fat08]. (d) Dark channel prior [HST09], (e) Zhang's results [ZLY ⁺ 10]	81
3.10	A synthetic experimental result. (a) the synthetic hazy image. (b) the ground truth image. (c) output haze-free image. (d) the estimated transmission map. (e) the ground truth map. . .	82
3.11	A failure case of the proposed algorithm. (a) Input hazy image. (b) Output image.	82
4.1	Aesthetic Energy of Colors	94
4.2	. (a) The color wheel under Red-Yellow-Blue(RYB) model. (b) The color wheel under RGB model (in the HSV color space). .	95
4.3	The assigned energy coefficients for different colors.	97
4.4	Color quantization for categorization.	97
4.5	. The gray scale images in different color spaces.	99
4.6	Color aesthetic energy for two test images.	100
4.7	Sound elements and their effects on perception.	102
4.8	Structural transition from images to music for audio matching. .	104
4.9	A brief description of our audio-visual mapping scheme.	106
4.10	. The flowchart of our proposed music-photo SlideShow scheme. .	107
4.11	Music Structure and Camera Motion.	111
4.12	An example of the camera path.	116
4.13	. Sample images of the experimental image dataset. Each group contains 200 images and 36 random images of each group are displayed in the figure.	117
4.14	. User Evaluation of Group 1.	118
4.15	. User Evaluation of Group 2.	119
4.16	. User Evaluation of Group 3.	120
5.1	The four frames (a)-(d) from a stage performance video clip. This segment lasts more than 4 seconds.	125
5.2	Saliency and detected foreground. Column(a) Original frames; Column (b) motion saliency; Column (c) spatial saliency; Column (d) fused foreground.	135
5.3	Frame Interest.	137

5.4	The synthesis example for the accelerated frame generation. Frame (1)-(6) are 6 continuous frames. The object motion velocity seems to increase by reducing the projection time. Within the same exposure time, the trace of moving object is longer and results in more noticeable motion blur. Figure (a) shows the ideal continuous combination of the 6 frames. In our implementation, we use the weighting combination in Equation 5.19 to accumulate temporal information (b).	141
5.5	The flowchart of the whole system.	143
5.6	Subjective User Evaluation. SD: segment detection. CW: camera work. PS: projection speed, FR: fusion result.	145
6.1	The flowchart of the proposed system.	160
6.2	The procedures of the proposed system. I. The input web page; II. The input web page is semantically segmented into blocks; III. Blocks are abstracted into vertices in a graph system by feature vectors containing the style and saliency information; IV. The graph system is built up by integrating nodes and forces.	161
6.3	The color wheels and color harmony. Left: RGB wheel. Right: RYB wheel. Take red ($c_i = 0$) as an example on the RYB color wheel, the 3 sets of harmonized color patches are: red/red purple & red/orange(the dashed-blue-line), red/green (the dashed-green-line), red/blue,red/yellow(the dashed-red-line).	164
6.4	The segmentation of cold and warm colors. Warm colors that are further away from the segmentation line have higher graphic weights, and it is the same as the cold colors.	167
6.5	Graphic mass and screen position. I. Screen-centered position provides the maximum stability; II. Object-counterweighting can also be balanced if the objects have similar graphic weights; III. The larger and heavier graphic mass on the right surpasses the one on the left, and the system becomes unstable.	169
6.6	Left: Experimental Result 1. A snap shot of CNN news with inserted advertisement. Some of advertisement candidates are listed on the right. Right: The estimated graphic weights of Experimental Result 1	170

List of Tables

2.1	Weights for different factors in Equation . Unsta:unstable, Infid:infident, orient:orientation. [MZZH05]	31
2.2	Media Representation Models	33
2.3	Comparison of the properties of current databases containing aesthetic annotations. PN: Photo.net [DJLW06], DP: Dpchallenge.com [KTJ06a], CUHKPQ [LWT11], Aesthetic Visual Analysis (AVA) [MMP12], CLEF: Visual Concept Detection and Annotation Task 2011	37
2.4	Features of Ke et al. [KTJ06b]	38
2.5	Features of Datta et al. [DJLW06]	39
2.6	Features of Li et al. [LGLC10]	41
2.7	Features of Khan et al. [KV12]	42
2.8	Bag-of-aesthetics- preserving (BoAP) features [SCK ⁺ 11].	44
2.9	Features of Luo et al. [LT08]	45
2.10	Features of Luo Wei et al. [LWT11]	46
2.11	Features of Niu et al. [NL12]	48
2.12	Features of Yang et al. [YYC11]	48
2.13	Features of VisQ [WCLH10].	51
2.14	Statistically significant correlations between features and patterns [ZCLR09].	52
3.1	Related Parameters	76
5.1	Details of User Study.	143
5.2	Output Rendering Parameters of Clip 02.	146
6.1	Comparison between graph drawing and the proposed advertisement selection framework.	153
6.2	Factors influencing graphic weight [Zet99].	165
6.3	Evaluation criteria for subjective user study.	173
6.4	User Evaluation. E.C: Eye Catching. In.: Intrusiveness. V.P: Visual pleasure. Cnt: Contribution. P.M: proposed method; R.D: random results.	176
7.1	A summary of the proposed media aesthetic applications.	183

Introduction

1.1 Aesthetics and Applied Media Aesthetics

Aesthetics, derived from the Greek word *aisthese-aisthanomai* (to perceive-feel-sense), is a branch of philosophy and closely related to the nature of art. Linked to culture, personal emotion and many other subjective judgments, it is common to think of aesthetics as the systematic study of beauty [Sax10].

“ Aesthetics is a term commonly used to refer to such diverse matters as theories of beauty and the elegance of a logician’s axiomatic system. Philosophically, the term has a far more precise designation. Today, those philosophers called aestheticians are concerned with two general enterprises - the theory of art and the theory of the aesthetic that emerged in the eighteenth and nineteenth centuries from the theory of beauty. ”[DSR89]

Since aesthetics refers to the study of aesthetic phenomena and judgement, one of its major concerns is the evaluation of beauty and ugliness. Actually, we make aesthetic decisions in our daily life consciously or unconsciously. When we choose a picture to decorate the bedroom, select flowers for the garden, or stand in front of the wardrobe, we are making aesthetic judgements. We need certain guidance or principles for such decision making, and this leads to the

study of aesthetics. However, different from the traditional interpretations, there have been controversies over aesthetics, art and beauty in the domain of philosophy. In modern art, beauty is no longer a necessary feature. For example, Goya's *Disasters of Wars* can not be predicated as "pleasant", but it is still regarded as a great work. Meaning and significance overcome the visual pleasure in aesthetic evaluation. More precisely, there are three important aesthetic concepts: beauty, art and the aesthetic experience – and they have slightly different meanings. The tragedy form of art is included in the concept of aesthetic experience, but not in that of beauty.

In spite of the confusions between aesthetic experience and the experience of beauty, it is still true that the focus of aesthetics today is on art and quite a good amount of art is beautiful and pleasing. To specifically describe the concerns of philosophical aesthetics is difficult, but in the domain of applied media aesthetics, it is much clearer and more direct. [Zet99] put forward the notion of *applied media aesthetics*, which concerns basic media elements, and aims to constitute formative evaluations as well as help create media products.

“Media aesthetics is a process of examining media elements such as lighting, picture composition, and sound – by themselves or jointly – and a study of their roles in manipulating our perceptual reactions, communicating messages artistically, and synthesizing effective media productions.” [DV01]

The intent is to “provide a theoretical framework that makes artistic decisions in video and film less arbitrary, and facilitate precise analysis of the various aesthetic parameters ([DV02])”. Compared to the traditional abstract philosophical definition, applied media aesthetics is different in several aspects.

- Applied media aesthetics does not try to answer the eternal question for aesthetics - the truth of beauty. It is not a question of the truth. Instead, it examines a series of aesthetic-related media elements, such as color and motion.
- Media platforms are no longer considered as neutral means of message distribution, but important elements of the aesthetic system. For example, in traditional art, artists exhibit their thoughts and emotions through their works, no matter whether by sculpture or oil painting. But in applied media aesthetics, medium itself acts as an important structural agent. The video shown on a film screen is quite different from that on a home television. Both the impact and the way of information delivery are different (details will be discussed in the later chapters.)
- Traditional aesthetics is restricted to analysis, while applied aesthetics can also serve to the case of synthesis. Under the guidance of applied aesthetics, we can both evaluate and compose aesthetic products.

1.2 Methodology of Applied Media Aesthetics

According to Zettl ([Zet99]), applied media aesthetics is an inductive process which works by combining aesthetic-related elements in a certain way. The five fundamental media elements are:

1. light and color,
2. two-dimensional space,
3. three-dimensional space,
4. time and motion,

5. sound.

These basic elements have their own characteristics, potentials and perspective aesthetic fields. They constitute the aesthetic “vocabulary”. Applied media aesthetics begins with the analysis of these elements, extends to the understanding of their contextual functions, and then helps examine how they can effectively classify and intensify the impact of media products. The five elements serve as the essential prerequisite in applied media aesthetics. It corresponds to the definition of media aesthetics given by Chitra Dorai ([DV01]), i.e. media aesthetics examines the media elements and studies their roles in media production. The analysis of the underlying principles starts from the interpretation of media elements.

These fundamental aesthetic elements are contextual. An image of bright colors and high contrast does not really show happiness (Van Gogh’s *Starry Sky*). In practice, people first setup a theme, and then use various mediums to communicate with others. It is the content that plays the most important role in aesthetics. But we still need to realize that the molding process of these ideas influences the effective delivery of authors’ intent. These production tools, taking our media production as an example, include the manipulation of cameras, the specification of colors, the control of light, the selection of focus and so on. From this point of view, the understanding of the fundamental aesthetic elements helps us to effectively clarify, interpret and produce mass communication. Therefore, this thesis commences with the analysis of basic elements, and then followed by the discussion on algorithms are based on the analysis and interpretation of aesthetic elements.

1.3 Aesthetic Elements

Artists manipulate audiences' perceptions, emotions and feelings via the manipulation of aesthetic grammars. Applied media aesthetics looks into and analyzes the language of media aesthetics and provides guidelines with which we can evaluate the effectiveness of media aesthetic products and optimally decide the structure of basic aesthetic elements, which include ([Zet99] [DV02]):

- *Light and Color.* Light is the most important factor to show shapes, space and time. The proper combination of light and shadows gives information of object shapes. The intensity of light can be the clue for time. For example, it is believed that light representing winter should be more bluish than for summer because the sun is weaker during winter days. Also the orientation of light can manipulate the emotion of the whole scene. The below-eye-level lighting, for example, shows instability, exaggerates tense and evinces horrible feelings. Colors, on the other hand, offer a new dimension of information by influencing global atmosphere of an event and constructing the primary mood of the scene (Figure 1.1). For example, the *Twilight City* (2009) uses a blue and unsaturated dominant color, which gives the audiences a feeling of quietness and grief, the emotional tone of the whole story.
- *Two-Dimensional Space.* The area within the two-dimensional screen places constraints on the arrangement of different objects. It is especially important for paintings, photography and screen composition. Just like painters and photographers, video producers need to consider the size and the aspect ratio of the screen. They carefully plan the composition of shots with some universal aesthetic rules. For example, the magnetism

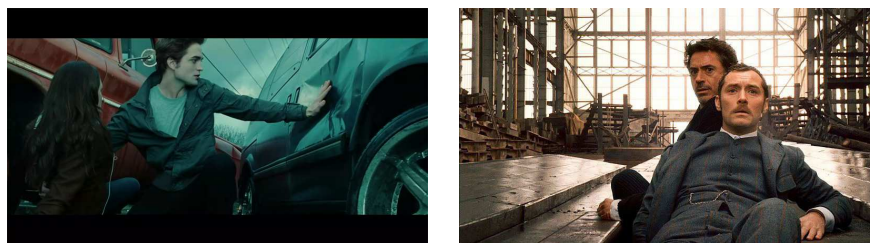


Figure 1.1: Dominant colors. The left image(*The Twilight City* (2009)) has a cold dominant color and it delivers the feeling of grief. The right image (*Sherlock Holmes* (2009)) has a warmer dominant color. It implies the cheerfulness of the lucky survival.



Figure 1.2: Different horizons suggest different natures of the whole scene. The horizontal camera view gives a stable scene while the right images has an unstable horizon, and it exaggerates the feeling of speed.

of frames requires reasonable space between screen boundaries and the region of interest, and different horizons suggest different natures of the whole scene, either stability or dynamism (Figure 1.2). There are some special composition rules for video production, like the safe area and display media. The former requires directors to place important objects towards the center of the frame, while the latter influences the kind of shots producers would like to choose. Meanwhile, video production enjoys its own features in the two-dimensional space. For those videos displayed on large movie screens, wider shots are able to show details quite well, while for family television, long-shots might lead to the loss of details.



Figure 1.3: Different shot points. The left image uses a horizontal angle, and it shows the sense of sacred. The middle image is taken from the side face. It emphasizes the continuity between buildings. The right image is taken from below, and it highlights the height and impact of the skyscraper.

- *Three-Dimensional Space.* Media products - photos and videos - are the projection of the 3D world onto a two-dimensional plane. They try to create the illusion of a 3-dimensional space on the 2D plane. Perspective plays an important part in constructing the illusion of depth. Camera focus effect creates the depth of the scene and emphasizes certain objects. Additionally, different shot points could create different levels of impact (Figure 1.3), and it serves as an important way to deliver producer's subjective views.
- *Time Motion.* The fourth dimension, time line, makes video unique from images and single photos. Motion is the most obvious and direct sign of time. But motion offered by videos is also an illusion because videos are nothing more than a series of still images. A sequence of images with slight shifts give the viewers the feeling of motion in their brain. Neatly controlled motion velocity can offer special aesthetic effects. For example, a slow motion during a race can intensify speed while accelerated motion is able to trigger certain moods because of the unpredictable jerks.
- *Sound.* Sound is an indispensable part for modern media production. Proper combination of video and audio tracks can produce higher impact

than using any one of them only. Not only speech provide additional information to the video track, but also non-literal sounds, like background music can quickly build up certain moods. Moreover, spatial sound enables video sound tracks to offer additional information beyond 2D video frames. This technique helps to build up a 3D world for audiences.

The above five elements of applied media aesthetics are dependent and contextual. Reliable analysis and evaluation must be based on the content of media themselves. Instead of understanding the content and trying to discover how it successfully creates higher meanings from series of shots, applied media aesthetics deals with properties of basic elements that make up the grammar and their structural composition. It aims at providing theories to make once unpredictable media production grammars less arbitrary. [DV01] defined *Computational Media Aesthetics* as “the algorithmic study of a number of image and aural elements in media and the computational analysis of the principles that have emerged underlying their use and manipulation, individually or jointly, in the creative art of clarifying, intensifying, and interpreting some event for the audiences.” It originally aims to interpret media data in order to automatically understand and make up the semantic gap. In other words, the gap between the richness of interpretation users want and the limitations of content descriptions that computer can generate today.

Computational media aesthetics also offers a new point of view towards media enhancement. Media grammars can be categorized into five classes as presented by [Zet99]. Professional producers are able to compose the fundamental elements in a way such that the media impact could be maximized.

While for home media production, constraints on equipment functions and producers' aesthetic sense limit media clips' interestingness as well as capacity of intent delivery. Based on established computational media aesthetic theories and frameworks, we want to find out if it is also possible to enhance the efficiency and effectiveness of home media productions from an applied media aesthetical point of view.

1.4 Scope and Contributions

1.4.1 Aim

There are two research areas related to multimedia aesthetics:

- *Aesthetic evaluation* studies the automatic rating of media products: the quality of images/video, the layout of websites etc. They extract corresponding aesthetic features and study to what level the features could influence the aesthetic appeal of media pieces. The aesthetic features are computationally interpreted for the integration assessment. The ACQUINE system [DW10] allows users to upload photos and rates the files automatically for their aesthetic quality.
- *Aesthetic processing* looks into the aesthetic enhancement of media products. Under the guidance of existing theories, aesthetic processing computes the features and improves the quality of media products from an artistic perspective. For example, Mubarak et al. makes use of aesthetic rules including the rule of thirds and golden ratio to rearrange the composition of internet photos, which are taken by amateurs using common consumer digital cameras [BSS10].

The two topics consider and deal with the aesthetic features, which have been discussed in the previous parts. But the different goals make each of them a special, and equally important, problem. In this dissertation, we will focus on the application of aesthetic grammars on multimedia processing problems, especially on aesthetic-interpretation of visual features, and their correlation with audio features. Aesthetic evaluation is out of the scope of this dissertation, instead we adopt the method of subjective user study to evaluate the success of results.

1.4.2 Approach

The semantic gap between the rich meaning that users want when they query and browse media, and the low-level nature of content descriptions that can actually be computed at present is still large. Computational aesthetics, therefore, aims to bridge the analytic and synthetic gap between computer science and arts. It investigates the creation of tools that can enhance the expressive power of applied arts, seeks to facilitate both the analysis and the generation of media and furthers our understanding of aesthetic evaluation.

1.4.3 Contribution

The computational media aesthetics framework proposed by Dorai et al. [DV01] begins at the study of a variety of media elements with insights into media production. In this dissertation, we propose four applications of computational media aesthetics on media enhancement and media authoring, including images, videos and webpages. We start at the extraction and interpretation of basic media elements, build computational models for aesthetic theories, and utilize the models to automatically or semi-automatically improve media

aesthetics. Based on our proposed media processing frameworks, we demonstrate the competence and advantages of media aesthetics from the following aspects:

- Aesthetic-related rules ensure the visual quality of outputs. Media aesthetics aims at understanding compositional and aesthetic media principles to guide content analysis. And its very initial target is to improve the aesthetic level of output media.
- Aesthetic-related criteria can simplify the classical media processing problems by placing subjective constraints on these problems, which are often ill-posed.
- Computational media aesthetics can optimize the results of traditional algorithms, such as image ranking, retrieval and online advertising.

1.5 Summary

Aesthetics studies beauty in art, and computational media aesthetics is different from the traditional content in the following ways:

- Traditional aesthetics considers the abstract philosophy of art, while applied media aesthetics studies the basic elements that are related to aesthetics, including light, color, space, motion and sound.
- Traditional aesthetics is mainly applied in art analysis while media aesthetics can analyze and process media products.
- Computational media aesthetics is more important in the production process [Zet99].

The study of media aesthetics adopts the inductive approach, i.e. the fundamental features related to aesthetics are first examined. This artistic information is computationally modeled, quantified and extracted from media pieces. We first examine their aesthetic characteristics, and then extend to the structures in the potentially aesthetic fields. The process of identification, interpretation and application is based on the selection of elements for a specific application. Professional producers manipulate the elements to influence recipients' perception. From the point of computational study, we want to utilize these formal elements to facilitate effective automatic, or semi-automatic, aesthetic manipulation.

1.6 Thesis Overview

The dissertation is organized as follows: Chapter 2 categorizes and reviews the literature of computational multimedia aesthetics. Chapter 3 applies the aesthetic criterion to solve the problem of single image dehazing. The proposed algorithm shows how the application of computational aesthetics can dramatically improve the efficiency and quality of traditional image processing. Chapter 4 proposes an image slideshow framework by equalizing the weights of visual and audio features. Aesthetic energy overcomes the gap between the two. Chapter 5 proposes an aesthetic-based home video post-processing framework, and it shows how aesthetic film grammars can be applied to home video processing. The method integrates traditional video retargeting and reprojection, and improves the performance of these independent techniques. Chapter 6 describes a force-based computational advertising scheme. The optimal advertisement candidate is defined to be the one that equalizes the

aesthetic force system of the visual features within a given webpage. And the summary and some conclusive discussions about our current work is given in Chapter 7.

Previous Work

Aesthetic assessment and aesthetic composition are two aspects of computational media aesthetics. The former aims to evaluate the aesthetic level of a given media piece and the latter aims to produce media outputs based on computational aesthetic rules. In spite of the different objectives, they adopt similar aesthetic grammars and models. As discussed in the previous chapter, media aesthetics begins at the analysis of fundamental elements, studies their contextual functionality, and utilizes the knowledge to guide media production. In this section, we will first go through the extraction, analysis and interpretation of media features, then look into their functionality and the ways to utilize the existing models by looking into applications in different areas.

2.1 Features that Represent Aesthetics

Aesthetic feature models are closely related to aesthetic analysis and applications providing concise and informative descriptions of media clips. Human perception system is too complex to be modeled by the current techniques, and hence automatic semantic understanding of media contents is still a tough problem. Most existing aesthetic descriptive models make use of low level features to interpret high-level semantic content by adopting widely accepted

evaluation criteria. For example, according to Rule of Thirds, an image should be imagined as divided into nine equal parts by two equally-spaced horizontal lines and two equally-spaced vertical lines. The important compositional elements should be placed along these lines or their intersections [Pet03].

On referring to the low level information, different description models make use of essentially similar features. Paintings, photography and videos share similar spatial visual criteria. So in the following discussions, we will mainly consider videos. Compared to the other media, videos have their unique features in the temporal domain. For example, [YLSL07] builds a visual perception model based on low-level features: motion, contrast, and scene rhythm. To interpret these low level features, they present some criteria:

1. Moving objects will attract more attention;
2. Objects those appear more frequently will attract more attention;
3. The position of the objects will also influence perceptual analysis;
4. Human beings pay more attention to the objects at the center of the frames.

The first criterion considers the importance of motion. The second considers object recognition. The third considers the frame composition. This is a typical process of building video feature models: extract features, propose widely accepted rules related to these features, and formalize constraints. It also follows the standard procedure of media aesthetics. Among the 3 criteria, the first two are unique temporal features for videos, and the third one is common for both videos and still art pieces.

Generally speaking, common basic aesthetic elements include:

- *Luminance and chroma.* Color is one of the most important features for visual analysis. It has direct influence on viewers' perception. Some color-related properties, such as saturation and harmonic color pairs, also play important roles in aesthetic evaluation. In professional film production, the dominant color is manipulated in post-processing to control the emotional tone of the movie.
- *Motion.* Motion is a unique attribute of videos which makes them different from still images. It is also an important attention-grabbing attribute in human perception and can be categorized into object motion and camera motion. The application of motion models ranges from low-level camera motion detection to high-level aesthetic video understanding.
- *Composition.* Frame composition is an aesthetic notion. Common criteria include Rule of Thirds and the magnetism between object placement and boundaries. In photographic theories, for example, the salient objects shall never be too close to the frame boundaries.
- *Object detection.* Certain objects are believed to be more competitive in attracting human attention. For example, human faces, animals, captions etc. In video content analysis, special importance is attached to such objects.
- *Audio.* The audio track, another unique feature for videos, is often made up of two parts: the dialogue and the music track. The informative content is more important for the dialogue track while for the music track, we mostly make use of beat, tempo, genre to analyze their emotional functions.

Current feature models depict the media stream from different aspects. The seemingly independent features should be semantically assembled for descriptive models. The most straightforward scheme is to linearly combine them with proper weighing factors. Some more sophisticated models have been presented to distinguish the different importance of those features based on experimental results of human perception [Mic06].

2.1.1 Object Position

It is widely accepted that the position of objects will influence human perception [MLZL02]. Objects in the center of the frame will attract higher attention than those off at the boundaries. So empirical weighing factors are often assigned to different regions of the frame.

In a standard visual weight model, the 2-dimensional frame is evenly divided into a 3×3 block matrix. The weighing factor matrix is in the form of a Gaussian matrix, with the highest value in the center and the lowest on the boundaries. A typical matrix given in [YLSL07] is

$$\begin{pmatrix} 1/6 & 1/3 & 1/6 \\ 1/2 & 1 & 1/2 \\ 1/3 & 1/2 & 1/3 \end{pmatrix} \quad (2.1)$$

The entry in the matrix denotes the weighing factor w_{ij} of the corresponding region. In practice, the sum of the matrix is often set to 1 in order to ensure stability of the whole system [DW10].

2.1.2 Spatial Features

Spatial features are important for still image analysis. Common features include color, brightness, shape etc. Human recognize color in terms of hue, contrast, saturation. In film theories [Zet99], color information is thought to reveal the emotional tone of the media products. It seems to be straight-forward to adjust emotional tone by altering color properties. [AYK06] manipulates color characteristics in their video editing framework under the assumption that darker and colder dominant colors signal negative feelings, while brighter and warmer colors imply happy, positive emotions. In their work, hue is attached with the highest importance on referring to emotion, and the rest of the chromatic features are not considered.

[YLSL07] proposes a contrast model which contains two aspects: luminance contrast and clearness contrast. Human vision system is sensitive to luminance changes, so the authors extract luminance information from the histogram statistics of DC coefficients. The *area contrast* is defined by the macroblock proportion of those in the foreground and the rest belonging to the background. *Clearness contrast* is defined by the subtraction of the AC coefficients between foreground and background. More specifically speaking, let MC_l represent the luminance contrast, DL_1 and DL_2 denote the dominant luminance value of foreground and background respectively, then the luminance contrast is

$$MC_l = \frac{|DL_1 - DL_2|}{64} \quad (2.2)$$

For each macroblock within the frame, the area contrast is defined by the absolute sum of 3 AC coefficients: $AC = |AC_{1,0}| + |AC_{0,1}| + |AC_{1,1}|$. Let C_1, C_2 denote the mean value of the area contrast of foreground and background

respective. The clearness contrast is given by

$$MC_c = \frac{|C_1 - C_2|}{\max(|C_1 - C_2|)} \quad (2.3)$$

where $\max(|C_1 - C_2|)$ is the maximal $|C_1 - C_2|$ of all frames in the video clip.

Chroma and brightness are common low level features that many aesthetics-related studies utilize. And the statistical models are widely used when we analyze the corresponding features. However, even though these spatial features have direct influence on human beings, information interpretation is essentially an aesthetic issue. Ideally speaking, reasonable color models for media spatial analysis shall be based on both psychological and aesthetic interpretations. Therefore, models of higher levels are needed for the seemingly low-level features.

2.1.3 Motion

Motion is one of the most important attributes of video data and makes it different from still images. The relative positional shifts between frames can give clues for region-of-interest detection and video saliency detection. On referring to the video analysis, the motion detection results need not necessarily be that accurate, hence many models simply utilize macroblock-based motion vectors because they are directly available in compressed video files. Some other models, which emphasize the importance of motion detection accuracy, choose optical flow for their wider applications.

Motion is classified as *local motion* (foreground motion, real motion, object motion etc.) and *global motion* (camera motion, background motion). Mostly, background is assumed to be a still scene, and its motion is introduced by the

camera work. Moving objects are usually classified as the foreground. Their motion is called local motion, because it is often independent from camera motion. In motion attribute models, local motion is attached with higher importance than camera motion, because it reveals the region of interest (ROI) and is more important for human perception. Global motion is related to camera work and plays an important role in video aesthetic analysis for deliberate camera work often reveals the intent of directors. And video producers also utilize proper camera motion to guide viewers' attention.

Like the spatial features in the previous discussions, motion is also modeled by their statistical properties. Let $\{(u_k, v_k), k = 1, 2, \dots, M\}$ denote the background macroblock motion vectors. [KCKK00] and [YLSL07] use an affine camera motion model

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} \quad (2.4)$$

This is a standard affine transformation model, where (x_1, y_1) denotes the source pixel while (x_2, y_2) is the destination pixel, and $a_i (i = 1, 2, \dots, 6)$ are the affine parameters. In the case of motion estimation, we have $x_2 = x_1 + u, y_2 = y_1 + v$, i.e. x_1, x_2, y_1, y_2 are the position indexes of macroblocks and the shift is the corresponding motion vector. The least-square scheme is applied to solve the 6 affine parameters. Once the parameters have been determined, the global motion vectors GMV can be computed for each macroblock based on their coordinates. Then the foreground object motion vectors are given by

$$FMV = MV - GMV \quad (2.5)$$

where GMV represents the estimated global motion vectors based on the affine model, MV is the real macroblock motion vectors and FMV is the foreground motion vector. In practice, it is not a trivial problem to segment background and foreground macroblocks. So when the affine model is estimated, all the motion vectors are taken into consideration. The foreground motion will inevitably influence the accuracy of the results. [YLSL07] uses an iterative scheme to reduce the influence of foreground motion. They iteratively use the affine parameters to update GMV . Based on the definition of background motion, if a macroblock belongs to the background, the estimated FMV will approximate to zero. Thus the small residue values of Equation 2.5 are thought to be brought in by foreground motion. The affine parameters are modified to make up for the error. The process repeats until the mean and variance of FMV falls below the given threshold. Then corresponding affine parameters are used to describe camera motion.

2.1.3.1 Global Motion

Global motion information gives clues of camera work, and this often reveals some intents of producers. In [MLZL02]’s camera attention model, they discuss the possible effects of camera work based on global motion. Typical camera motion is categorized into 6 types, i.e. panning and tilting, rolling tracking and booming, dollying, zooming and still. Camera motion are characterized by the motion vectors. [KCKK00] use the affine parameters to classify video shots and understand camera motion characteristics. Here we adopt the parameters in Equation 2.4, and the 6 kinds of camera motion are defined

by

$$pan = a_1 \quad (2.6)$$

$$tilt = a_4 \quad (2.7)$$

$$zoom = \frac{1}{2}(a_2 + a_6) \quad (2.8)$$

$$rotate = \frac{1}{2}(a_5 - a_3) \quad (2.9)$$

$$hyp_1 = \frac{1}{2}(a_2 - a_6) \quad (2.10)$$

$$hyp_2 = \frac{1}{2}(a_3 + a_5) \quad (2.11)$$

In order to associate camera motion information with human perception, the authors provide several general camera-work rules:

- zooming and dollying are used to emphasize the important objects.
- panning makes audiences neglect some objects.
- frequent camera motion is thought to be random and unstable.

Then different importance-weighting factors are associated to subshots according to the corresponding camera work. For example, subshots with zooming are thought to be more important than those with panning. And unstable subshots are believed to have lower quality.

2.1.3.2 Local Motion

Once the foreground object motion information is available, different motion models are presented to describe the motion pattern. Typical models include three aspects of information:

- *Velocity*. [MLZL02] calls it intensity indicator. This item depicts the

fastness of object motion. [Wol96] compute it by the normalized motion vector magnitude

$$M_k = \frac{u_k^2 + v_k^2}{Max} \quad (2.12)$$

where Max represents the maximum of motion vector magnitude of all macroblocks. Based on varying applications, different models will decide whether to sum all the value up within one frame, or to use a matrix to present each frame's motion velocity information. In the former case, for example, [YLSL07] use the mean magnitude of one frame to represent the whole frame's velocity.

- *Spatial coherence indicator.* Spatial difference is used to describe the motion field smoothness. [MLZL02] compute the phase histogram distribution within a local window of each pixel. [YLSL07] use the standard variance in a local window at each macroblock to depict the spatial motion information.
- *Temporal coherence indicator.* The most straight-forward way is to compute motion vector field difference along the time dimension. [MLZL02] use the histogram distributions of pixel intensity along the time line (L frames). [YLSL07] use the average of correlations with proportional weights for all the macroblocks to model temporal motion correlation, i.e

$$M = mean(\frac{\sqrt{\|V_k - V'_k\|}}{6 \cdot \omega_k}) \quad (2.13)$$

where V_k is the motion vector of macroblock k and V'_k are the weighed average motion vectors of macroblocks near k .

2.1.4 Composition and Object Detection

Frame composition is the issue of properly arranging objects in the frame. Different regions in the frame have different levels of audience attention, and the visual importance weighing factors influence object composition of media clips.

[HLZG03] puts forward an automatic attention extraction scheme, based on seeded region growing. The J -map of the image is firstly computed. The attention seed areas are defined to be those with low J value but high local saliency value. For a given pixel P in the still image, let R denote its neighboring region. Then the average and standard deviation of J -map value in the region R are denoted as μ_J and σ_J respectively. Similarly, the average and standard deviation of saliency value in the region R are denoted as μ_S and σ_S . Thus the area attention model of P is given by

$$A_P = e^{-\mu_J + \sigma_J} - e^{-(\mu_S - \sigma_S)} \quad (2.14)$$

More typical object detection schemes are to find certain spatial objects such as human [WH06] [YLSL07], animals [YLSL07], and events [Div07]. The special scene model assumes that human will be more interested in certain video contents, which include human faces and captions. So this model detects the existence and location of such objects and assign different weighing factors to them. Based on the detected faces and captions, the perceptive face model

is given by [YLSL07]

$$P_{face} = \sum_{i \in \Omega_{face}} \omega_i \quad (2.15)$$

$$P_{caption} = \sum_{i \in \Omega_{caption}} \omega_i \quad (2.16)$$

$$P_{fact} = P_{face} + P_{caption} \quad (2.17)$$

where ω is the corresponding weighing factors of macroblocks.

In addition to the spatial composition, temporal compositional characteristics have also been considered. It refers to the combinational pattern of shots of different length, reflecting certain personal styles of directors [Dav10]. For example, clip duration in exciting videos will be shorter and in videos with negative emotions, the reverse is true [AYK06].

A video clip may contain several segments (shots or subshots) of different length. These differences in length contain a certain rhythm which is exploited by a statistical model [YLSL07]. According to neurobiological theory of perception formation, longer segments may attract more human attention, and the content changes between neighboring frames may influence the human perception to a certain degree. So the statistical rhythm of frames is given by

$$P_{rhy} = (1 - \frac{N_{intra}}{AMB}) \cdot (1 - \frac{E_{resi-error}}{M_{segment}}) \cdot L \quad (2.18)$$

where N_{intra} denotes the number of intra-coded macroblocks, $E_{resi-error}$ denotes the average energy of all residual error blocks, $M_{segment} = \sum E_{resi-error}$ is the sum of all residual error in the segment, and L is the length of the segment to which the current frame belongs.

2.1.5 Audio

Audio attention is an important part in an aesthetic content analysis framework. Speech is meaningful for human beings, and background music is used to create or emphasize the atmosphere in an artistic composition. It often conveys a certain emotion and enhances the impact of art itself. The nature of speech and background music is different. They are often independently modeled in a media attention framework.

[MLZL02] build an audio saliency attention model based on sound energy. Specifically speaking, human beings are more likely to be attracted by loud or sudden sound. Thus in their model, they build the audio attention model based on two attributes: loudness and suddenness. The former one is defined to be the average energy of an audio segment, which is related to the absolute loudness of sound. The latter is modeled by energy peaks in audio segments, which reveals the sudden drops or increase of audio loudness.

$$\bar{E}_a = E_{avr}/M_{ax}E_{avr} \quad (2.19)$$

$$\bar{E}_p = E_{peak}/M_{ax}E_{peak} \quad (2.20)$$

$$M_{as} = \bar{E}_a \cdot \bar{E}_p \quad (2.21)$$

where M_{as} is the audio saliency of the whole audio, E_{avr} is the average energy of each audio segment, and E_{peak} is the energy peak of each audio segment.

The audio stream also gives clues for the video content. [Div07] use a data training scheme to analyze the audio track. They divide the audio segments into several types based on the audio information. The content of the input audio track is classified into different types, for example, applause, cheering, music, speech etc. The system is trained with typical audio segments, and

it compares the likelihood of the audio track of input video clip with the database. Based on the different audio types, the excitement of video content can be inferred.

[FCG02] use audio self-similarity analysis. The self-similarity for the past and future region is estimated. Meanwhile, the cross-similarity for past and future regions is also computed. Interesting points are assumed to lie between regions of high self-similarity. In the experiment, they generate a matrix whose rows and columns are the normalized region value of the audio clip, to compute the similarity between each region.

In addition to analyzing the audio interestingness, the model can offer clues for segmentation. In [FCG02]’s work, they search the diagonal of the similarity matrix to find the salient audio changes. They use a checkerboard-like Gaussian filter to do the kernel correlation along the diagonal. The peaks are selected to be the segment boundaries. In their framework, they consider the signal of the audio track itself without taking assumptions about nature of genre.

2.1.6 Fusion

The above models describe videos from different aspects. In order to build a semantic description of the given media piece, the seemingly independent information needs to be integrated. The straight-forward fusion scheme is the linear average of all the attribute values. In the local motion saliency model [YLSL07], motion is modeled from three aspects: velocity (mean magnitude) PM_{mv} , spatial coherence (spatial variance) PM_{sc} , temporal correlation (temporal frequency) PM_{tc} . Based on the authors’ assumptions, continuous motion could attract human perception more than discontinuous motion. Thus

Attributes	Unsta.	Jerky	Infid.	Bright	Blur	Orient.
ω_i	2.2009	0.2009	0.1402	0.1449	0.1636	0.1495

Table 2.1: Weights for different factors in Equation . Unsta:unstable, Infid:infident, orient:orientation. [MZZH05]

they define the local motion saliency model based on the above three aspects:

$$P_{motion} = \frac{PM_{mv}}{0.4 \times PM_{sc} + 0.6 \times PM_{tc}} \quad (2.22)$$

The choice of weighing coefficients is the central issue in linear fusion techniques. And the parameters are often decided by empirical knowledge. [MZZH05] propose their linear average fusion model based on user study. Consider the linear function

$$Q(\theta) = \sum_i \omega_i F_i(\theta) \quad (2.23)$$

where F_i is the value of video attribute i , $\sum_i \omega_i = 1$ is the weight of the influence of the i th factor. The larger ω_i is, the higher influence the corresponding attribute will have on audiences. And the detailed parameters in their work are listed in Table 2.23

In this model, they mainly consider the different levels of influence of video artifacts on human being. A factor fusion model is given by [HLZ04a] [MZZH05]. Given an attribute vector $x = (x_1, x_2, x_3 \cdots x_n)$, where x_i represents the i th factor and n is the total number of video attributes. They make two assumptions of the attributes:

- the factor vector x with higher mean deviation has a higher unacceptable value

- the fusion function is a monotone increasing function.

so that their video attribute fusion model is

$$Q(\theta) = E(x) + \frac{1}{2(n-1) + n\lambda} \sum_{i=1}^n |x_i - E(x)| \quad (2.24)$$

where $E(x)$ is mean of x , $\lambda > 0$ is a predefined constant.

2.2 The Applications of Multimedia Aesthetics

Two of the most important applications of computational multimedia aesthetics are aesthetics evaluation and aesthetic processing. The former automatically rates the quality of media data, and differentiates the ones of low quality from those of high quality. Aside from the computational evaluation of media aesthetics, aesthetic computing also aims to apply elements of art and design to the field of computing. Aesthetic grammar defines series of rules to connect aesthetic media clips, including color, rhythm, space, motion, audio etc. Especially in the area of film production, all these grammar rules are defined in aesthetic domain. Efforts have been made to build computational models and apply these models on media processing. These processing techniques include enhancement, authoring and designs.

2.2.1 Aesthetic Evaluation

In the field of computational media aesthetics, most of the work is done in aesthetics assessment. After all, before we can apply media aesthetics on retrieval, enhancement, design and any other area, the very initial stage of work is to differentiate "good" from "bad". Actually, automatic non-reference image

Features	Description	Remarks
Object Position	<ul style="list-style-type: none"> • Center objects are more important. • The Rule of Thirds. • Golden Ratio. 	Relies on saliency estimation, and more aesthetic issues can be considered.
Spatial Features	<ul style="list-style-type: none"> • Color influences mood. • Contrast implies important content. 	Disciplines are abstract because the process of aesthetic interpretation is closely related to the computational modeling.
Motion	<ul style="list-style-type: none"> • Global Motion. Camera work gives clues for directors' intent. • Local Motion. <ul style="list-style-type: none"> – Velocity – Spatial Coherence. – Temporal Coherence. 	Relies on proper motion estimation.
Composit.	<p>The features include:</p> <ul style="list-style-type: none"> • Object detection. Human, faces, animals, etc. • Spatial composition. Color pairs, Rule of Third, etc. • Temporal composition. Rhythm, clip length. 	-
Audio	<p>The features include:</p> <ul style="list-style-type: none"> • Speech. Length, content understanding. • Music. Beat, genre, energy, ... 	Correlation between video stream and audio stream need to be considered.

Table 2.2: Media Representation Models

quality assessment is of high importance itself. Consider the situations that the search engines could incorporate photo quality into the ranking system and return the best looking photos. Family users can manage their photo book by the quality of their vacation pictures and the system could automatically decide the photos that can be shown to friends.

With the advancement of imaging, storage, and networking, multimedia production and sharing now has become a common daily practice. However, the quality of media data produced everyday varies dramatically due to various constraints for common users. Finding data of high quality, particularly professional photos, movies and TV shows, is important for a wide variety of applications, such as media sharing, copyright protection, media-based advertisement, and media ranking. Automatically finding high-quality data is difficult as media quality assessment is usually subjective and often requires semantic content understanding. Existing media quality assessment methods can be roughly categorized into two classes: from visual quality and from aesthetics. The former one considers the low-level visual features, while the latter focuses on relatively higher features.

Common quality assessment methods focus on the measurement of media quality degradations caused by compression ([DVKG⁺00]) and transmission ([SB10a]). These methods assess the image and video quality by measuring the low-level visual distortions, such as blocking, ringing, mosaic patterns, false contouring, blur, noise, ghosting, jerkiness, and so on. Generally speaking, these approaches can be further categorized into reference-based and nonreference-based methods. Reference-based methods require the original non-distorted media data for quality assessment ([WWS⁺06]). Although these methods can provide reasonable assessments, they require the references that

are often not available in practice. Non-reference-based quality assessment, also known as blind image quality assessment, is much more complex. [SBC05] serves as one of the successful examples of non-reference quality assessment. A comprehensive introduction to computational aesthetic evaluation can be found in [Gal12].

2.2.1.1 Aesthetics-Related Database

Classic automatic aesthetic evaluation approaches study a set of visual features describing various characteristics related to image quality and aesthetic values. Such approaches are consistent with the media aesthetics theories proposed by Zettl [Zet99] which begins with the study of basic aesthetic elements. These image indicators are then used to generate multidimensional feature spaces. The nature of aesthetics is highly subjective, and machine learning algorithms are developed to estimate the aesthetic scales of images based on the extracted features. Therefore, reliable ground truth data, which have been properly collected and annotated for aesthetics analysis, are highly important for the computational accuracy. Before going further into the issue of automatic media aesthetic analysis, we first briefly look into the publicly available databases containing aesthetic annotations.

- **Photo.net** [DJLW06] contains 3,581 images gathered from the social network Photo.net. Members are invited to score the aesthetics of images from 1 to 7, and the database contains the mean aesthetic score of each image.
- **Dpchallenge.com** [KTJ06a] contains 12,000 images, half of which are considered high quality and the rest labeled as low quality. Images are

scored by the community users of Dpchallenge.com from 1 to 10. The database is randomly generated by taking images whose scores fall into the top and bottom 10%.

- **CUHKPQ** [LWT11] consists of 17,613 images obtained from a variety of on-line communities. The images have been divided into 7 semantic categories and each of them is labeled as either high or low quality. Therefore this dataset consists of binary labels of very high consensus images.
- **Aesthetic Visual Analysis (AVA)** [MMP12] contains over 255,000 images which are also taken from Dpchallenge.com. These images cover a wide variety of subjects on 963 challenges. In addition to the scores given by Dpchallenge.com members, AVA provides three types of additional annotations: Aesthetics, semantics, and photographic style.

In addition is CLEF (Image CLEF: Visual Concept Detection and Annotation Task 2011) is also a large dataset introduced in the multimedia retrieval community. It contains 1 million images from Flickr with textual tags, aesthetic annotations (Flickr’s interestingness flag) and EXIF meta-data. But this dataset is not essentially rich in aesthetics-related annotations, and only the “interestingness” flag is somewhat related. A comparison between current aesthetic datasets is given in Table 2.3.

2.2.1.2 Image Assessment

Traditional media quality evaluation algorithms work well for low-level visual quality assessment. This thesis focuses on aesthetic-based media evaluation, and will not go further into the issue of the distortion assessment. Moreover,

Prop. \ D.S	PN	DP	CUHKPQ	AVA	CLEF
Large Scale				✓	
Score Distr.	✓			✓	
Rich Annotation		✓	✓	✓	✓
Semantic Labels			✓	✓	✓
Style Labels				✓	✓

Table 2.3: Comparison of the properties of current databases containing aesthetic annotations. PN: Photo.net [DJLW06], DP: Dpchallenge.com [KTJ06a], CUHKPQ [LWT11], Aesthetic Visual Analysis (AVA) [MMP12], CLEF: Visual Concept Detection and Annotation Task 2011

the quality of images and videos often cannot be perfectly measured only with respect to the low-level distortions only. An image or video that is free of the above artifacts, can still be aesthetically displeasing. Here comes the aesthetic rules that further help to evaluate the quality of media data. Studies have been done on the aesthetics-based image quality assessment [DJLW06], [LC09], [ZCJ⁺06].

Ke et al. [KTJ06b] considers the issue of distinguishing professional photographs and those taken by amateur users. In traditional approaches of assessing image quality, low-level features are fused in a way without any subjective guidance. Therefore, Ke’s work starts with identifying criteria that people use to rate photos, and then designs features to match people’s perception of photo quality. Some of their important extracted features are listed in Table 2.4. The challenge is the conversion from abstract aesthetic terms, such

Distortion	Aesthetics	
Blur	Composition	Color
Unlike professional images, amateur photos are often degraded by blur.	<ul style="list-style-type: none"> • Edge Distribution. • Hue Distribution. 	Quantized RGB Distribution

Table 2.4: Features of Ke et al. [KTJ06b]

as good composition, pleasant colors, suitable lighting, to concrete computable measures, which can be estimated by computers. They admit the variation between their approach and the real human perception. Also, they claim that their work is the two-class image classification problem, i.e. to differentiate "high quality" photos and "low quality" ones, not the finer differences between masterpieces.

Further improvement of the two-class categorization between aesthetic images and those of low-quality is to score the input image. Datta et al. [DJLW06] build an automated classifier based on support vector machines and classification trees using linear regression on polynomial terms of the features to infer numerical aesthetic ratings. They take emotion into consideration, and attempt to explore the relationship between emotions which pictures arouse in people, and their corresponding low-level content. The extracted features are listed in Table 2.5.

Afterwards, they extend their work in [DJLW06] and propose a machine-learning-based online system ACQUINE [DW10] – Aesthetic Quality Inference Engine - which is a publicly accessible system allowing users to upload photographs and have them rated automatically for aesthetic quality. They

Brightness	Color	Composit.	Basic proper- ties	Semantic feature
Exposure of light, colorfull- ness	Hue, satu- ration	Rule of Thirds, complex- ity based on color patches, depth of field, shape convexity	Smoothness, aspect ratio	The in- tegrated region matching (IRM) distance between images.

Table 2.5: Features of Datta et al. [DJLW06]

treat the problem of aesthetics inference as a standard two-class classification problem as well. The Support Vector Machine (SVM) classifier is trained based on the extracted features that are determined to have a good correlation with aesthetic quality. The classifier is trained so as to differentiate the extremely good images from plain ones. The optimal hyperplane of the SVM is then used to score the image with a sigmoid function. The distance from the image feature to the hyperplane is mapped to a 0-to-1 scale and the sigmoid function takes in the distance $dist$ and computes a score by:

$$score(dist) = \frac{1}{1 + \exp^{-dist}} \quad (2.25)$$

The statistic scoring results of ACQUINE is shown in Figure 2.1. The distribution curve follows the Gaussian distribution. Among these uploaded images, only about 1% of them get scores in excess of 90, thereby making getting very high ACQUINE scores a rare occurrence. It is the same case for scores lower than 10. And most images tend to fall into the 25-55 bracket, i.e.

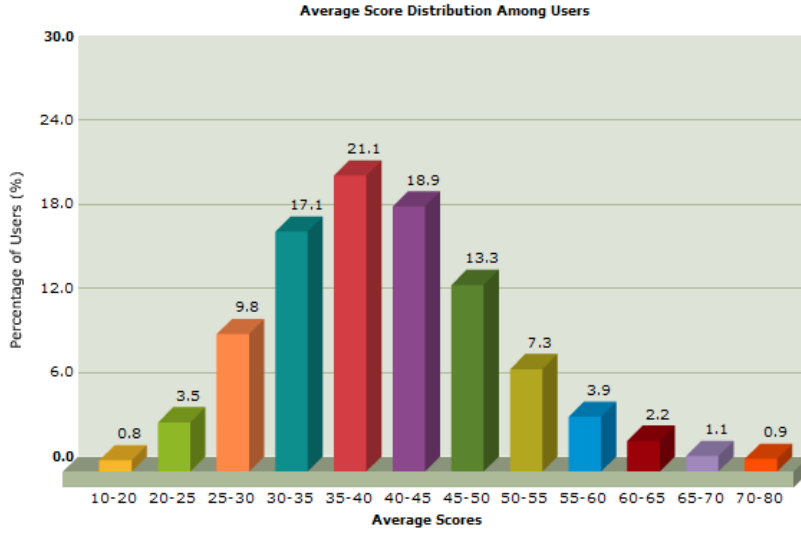


Figure 2.1: The statistic scoring results of ACQUINE [DW10].

most images are plain. Datta’s work provides a standard way to assess image aesthetics, and many other works follow almost the same way, i.e. feature extraction plus SVM-based classifiers [WBT10] [BSS10].

Traditional image aesthetic evaluation systems do not consider the theme of images. However, the generalized aesthetic rules are not really universal. Taking the Rule of Thirds as an example, which is a widely accepted rule adopted in evaluating compositional quality in many previous work [DLW08] [DJLW06]. However, the rule may be applied differently to a landscape photo and a portrait. Li et al. [LGLC10] propose an image aesthetic quality assessment system that deals with photos containing multiple faces. Widely used image dataset for image aesthetic evaluation include DPChallenge.com and Photo.net, in which most images are professional. Li’s team conducted an on-line survey to collect people’s opinions towards a set of consumer images, and it provides a larger image database with especially ordinary consumer photos. The extracted features can be found in Table 2.6. They adopt a high-level of

Content Features	Social Relationship Features	Perceptual Features
Foreground brightness contrast, color correlation, and clarity contrast. Background color simplicity.	The social relationship of people in the photo might emotionally affect the viewer's preferences. Corresponding features include face expression, face pose, and relative position.	Artistic rules including symmetry, composition, colorfulness, and consistency.

Table 2.6: Features of Li et al. [LGLC10]

interpretation of low-level features. The quality assessment process is defined to be a multiclass categorization problem and a Gaussian-kernel SVM is applied for classification. Experiments show an improvement for the face-image assessment as compared to ACQUINE [DW10].

Khan et al. [KV12] also try to narrow down the scope of the aesthetic evaluation problem. Instead of presenting an evaluator that can be applied to all kinds of pictures, they focus on photographic portraits of individuals. The growth of cell-phones and social media websites has allowed individual photographs a relevant space. Therefore the proposed problem focusing on portraits has reasonable applications for users. And the characteristics of individual portraits have made saliency detection more tractable. Unlike traditional bottom-up analysis of aesthetics, which starts from the calculation of features and correlate them with visual aesthetics, they adopt a top-down approach. They analyze the images mainly from the compositional point of view and the details of extracted features are listed in Table 2.7. In addition to the traditional visual features such as color, texture and statistical information,

Classic features	Compositional features
<ul style="list-style-type: none"> • Color, texture, statistic values • Face position, size • Horizontal line • Aspect ratio 	<ul style="list-style-type: none"> • Object Composition: Rule of thirds, golden ratio • Light/shadow composition: face illuminance, contrast, brightness

Table 2.7: Features of Khan et al. [KV12]

the top-down scheme attaches higher importance to the image composition. Human faces, as the most important salient region for portraits, enables an easy adoption of Rule of Thirds.

Instead of focusing only on individual portraits, Pere Obrador et al. [OSSO12] present a data-driven category-based approach to automatically assess the aesthetic appeal of photographs. In their proposed system, 7 popular image categories are considered: animals, architecture, cityscape, floral, landscape, portraiture and seascapes. Different categories have different dimensions of extracted features. For example, the model for animal category is composed of 22 features while that for architecture contains 24 features.

Similar to Li et al.’s work on portraits, Chen et al. [LC09] focus on the aesthetic evaluation of digital painting images. Table 2.2 lists the features they extracted. Actually the extracted features do not really reveal the characteristics of paintings, but some very classic images descriptors are extracted. But the issue of painting evaluation is interesting, and still an open problem.

Feature	Meaning of Feature	Characteristics	Feature	Meaning of Feature	Characteristics
f_1	Average hue across the whole image	Color	f_2	Average saturation across the whole image	Color
f_3	Number of quantized hues present in the image	Color	f_4	Number of pixels that belong to the most frequent hue	Color
f_5	Hue contrast across the whole image	Color	f_6	Hue model the painting fits with	Color
f_7	Saturation-Lightness model the painting fits with	Color	f_8	Arithmetic average brightness	Brightness
f_9	Logarithmic average brightness	Brightness	f_{10}	Brightness contrast across the whole image	Brightness
f_{11}	Blurring Effect across the whole image	Composition	f_{12}	Edge distribution metric	Composition
f_{13}	Horizontal coordinate of the mass center for the largest segment	Composition	f_{14}	Horizontal coordinate of the mass center for the largest segment	Composition
f_{15}	Horizontal coordinate of the mass center for the 3 rd largest segment	Composition	f_{16}	Vertical coordinate of the mass center for the largest segment	Composition
f_{17}	Vertical coordinate of the mass center for the 2 nd largest segment	Composition	f_{18}	Vertical coordinate of the mass center for the 3 rd largest segment	Composition
f_{19}	Mass variance for the largest segment	Composition	f_{20}	Mass variance for the 2 nd largest segment	Composition
f_{21}	Mass variance for the 3 rd largest segment	Composition	f_{22}	Mass skewness for the largest segment	Composition
f_{23}	Mass skewness for the 2 nd largest segment	Composition	f_{24}	Mass skewness for the 3 rd largest segment	Composition
f_{25}	Average hue for the largest segment	Color	f_{26}	Average hue for the 2 nd largest segment	Color
f_{27}	Average hue for the 3 rd largest segment	Color	f_{28}	Average saturation for the largest segment	Color
f_{29}	Average saturation for the 2 nd largest segment	Color	f_{30}	Average saturation for the 3 rd largest segment	Color
f_{31}	Average brightness for the largest segment	Brightness	f_{32}	Average brightness for the 2 nd largest segment	Brightness
f_{33}	Average brightness for the 3 rd largest segment	Brightness	f_{34}	Hue contrast between segments	Color / Comp
f_{35}	Saturation contrast between segments	Color / Comp	f_{36}	Brightness contrast between segments	Brightness / Comp
f_{37}	Blurring contrast between segments	Composition	f_{38}	Average hue for the focus region	Color
f_{39}	Average saturation for the focus region	Color	f_{40}	Average lightness for the focus region	Brightness

Figure 2.2: The extracted features of Chen et al. [LC09]

Photographs are sincere duplications of the reality, while paintings themselves contain highly-subjective aesthetic nature. The two have some criteria in common, such as color harmony, lighting condition, composition rules. However, paintings are the subjective interpretation of the reality and are much more abstract than photographs.

Su et al. [SCK⁺11] consider scenic photos only. They apply a bottom-up approach instead of top-down methods that make use of rule-specific features listed in the photography literature. Therefore, the proposed method can cover both implicit and explicit aesthetic features by a learning process. Details of their bag-of-aesthetics- preserving (BoAP) features are listed in Table 2.8. And Adaboost is utilized for analyzing the relation between the extracted BoAP features and landscape photos. As a bottom-up approach, they do not

Color	Texture	Saliency	Edges
Statistical features of HSV, LBP and saliency map.			Histogram of oriented gradient (HoG)

Table 2.8: Bag-of-aesthetics- preserving (BoAP) features [SCK⁺11].

really propose the characteristic features of landscape images, but rely on the regression procedure.

Luo et al. [LT08] take another approach of addressing aesthetic feature extraction and interpretation. Classic aesthetic assessment systems compute features from the whole image without considering the image characteristics of different themes. They treat all photos equally without considering the diversity in photo content. Category-based approaches target certain kinds of image classes and make the extracted features more tractable. They place constraints on the characteristics of foreground objects by narrowing down the problem scope. As an alternative approach, Luo’s group focus on the foreground objects directly. They assume that good photographers often treat the foreground subjects and the background very differently. The subject of the photo is differentiated from the background to highlight the topic of the photo. One of the well-known techniques is to use low-depth of field to segment the two. Therefore, their proposed work starts with blur detection to roughly identify the focus subject area. Several quantitative metrics are developed to evaluate human perception of photo qualities. The extracted features are listed in Table 2.9

Automatic image aesthetics evaluation based on content is studied in [LWT11]. They emphasize the importance of subject areas, which is assumed

Composition	Luminance	Focus Control	Color
Simplicity: color distribution. Rule of Thirds (RoT): the distance between centroid of foreground and the intersects of RoT	Brightness contrast between foreground and background	Clarity contrast between foreground and background	Color harmony: The ratio of hue, saturation and brightness between bright and dark regions.

Table 2.9: Features of Luo et al. [LT08]

to draw the most attention of human eyes. Actually, professional photographers may adopt different photographic techniques and may have different aesthetic criteria in mind when taking different types of photos (e.g. landscape versus portrait). Therefore, the authors divide the images into 7 categories and extract visual features in different ways according to the categorization of photo content. The seven classes are: animal, plant, static, architecture, landscape, human and night. Details of extracted features are listed in Table 2.10. Similar to the content-based aesthetic evaluation approach in [LWT11], Wong et al. [WL09] also attach high importance on the local salient regions, which contain photo subject. They compute the features of foreground and background regions separately and fuse to obtain an additional feature dimension.

The classic aesthetics-related features and their interpretation have been discussed in the previous section. Another problem in automatic scoring of aesthetics is to adaptively select proper features. Jiang et al. [JLC10] propose a regression method, named Diff- RankBoost, based on RankBoost and

Global features		Subject area features		
Hue Composition	Scene Composition	Dark Channel	Face-based Feature	Complexity
Harmonic template on the hue wheel	Locations and orientations of semantic Hough lines.	The dark channel prior serves as a combined measurement of clarity, saturation, and hue composition.	The ratio of face areas, the average lighting of faces, the ratio of shadow areas, and the face clarity	The ratio between the super-pixels in the subject area and that in the background.

Table 2.10: Features of Luo Wei et al. [LWT11]

support vector techniques to estimate fine-granularity aesthetic scores ranging from 0 to 100. Their work also categorizes images into different classes and each class is attached with a spacial classifier, but these classifiers are taken as the coarse-granularity of the initial scoring system. For one thing, the training data may be insufficient if the image classes taken individually. For another, users may not be really interested in the exact aesthetic values, i.e. users do not really care about if an image is 96% better or only 92% better than the rest images.

2.2.1.3 Video Assessment

In addition to the aesthetic assessment of single images, video quality has also been studied. The initial step still lies in differentiating professional videos from amateur ones. The importance of this assessment problem could be revealed in video retrieval, copyright protection, video-based advertisement,

video sharing and so on. For one thing, video management tools with models of aesthetic appeal can help users to navigate and enjoy their personal video collections. On the other hand, filtering and re-ranking video search results with a measured aesthetic value would probably improve the user experience. Moreover, it helps evaluate the videos and identify if they are “advertisement worthy” or not. Intuitively speaking, Video assessment can be taken as an extension of aesthetic evaluation of images. For example, Luo et al. [LT08] directly add descriptors to evaluate motion velocity and complexity as temporal information.

Niu et al. [NL12] addresses the video aesthetics rating problem by examining the discrepancy between how high-quality professional and low-quality amateur videos are created. A professional video not only tells good stories but also is aesthetically appealing, and viewers can easily differentiate professional videos from amateur-made ones at the first glance without considering the content. Therefore, their proposed system is purely based on pure visual features. In their framework, they further assume that the input videos are free of compression and transmission-related degradations. And the feature details are listed in 2.11. Features are categorized into two classes: the video features and the image features. More specifically speaking, they are the temporal features and spatial features. Spatial features are similar to those of still single images, while motion is the most important feature in the temporal scale.

The temporal axis makes videos different from single images. Therefore the aesthetic evaluation of videos is more than the naive extension of the work done for single images to a frame sequence. Yang et al. [YYC11] attach high priority to temporal information, and come up with more useful video-

Distortion		Aesthetics			
Noise	Blur	Video		Images	
	Depth of Field	camera motion	Shot length	Illumination	Color

Table 2.11: Features of Niu et al. [NL12]

Semantic-independent features	Semantic-dependent features
<ul style="list-style-type: none"> • Motion space measures the coherence between foreground motion and that of the background. • Hand shakes • Color harmony • Composition: Rule of Thirds, contrast between foreground and background, shape convexity 	<ul style="list-style-type: none"> • Motion Direction Entropy (MDE) measures the velocity of motion • Color Saturation and Value • Luminance

Table 2.12: Features of Yang et al. [YYC11]

based features such as motion space and motion direction entropy. They categorize the extracted features into semantic-independent and semantic-dependent ones. Details of these features are listed in Table 2.12 cues, instead of high-level content-based or emotion-based semantic analysis.

Moorthy et al. [MOO10] studies the aesthetic evaluation of consumer videos. This type of videos are not professionally generated, therefore low-level features related to distortions, such as illuminance, frame rate, blocky artifacts, can offer reasonable prediction of the quality. The authors have gathered their own video dataset which contains 1600 video clips from YouTube.

A 15-second segment is extracted from the middle part of each clip. These clips are scored on a 5-point scale based on their aesthetic appeal. On the issue of feature extraction, they follow the classic way of single image aesthetic assessment. But they propose a hierarchical pooling approach to collapse each of the features extracted on a frame-by-frame basis into a single value for the entire video. Here “pooling” is defined as the process of collapsing a set of features, either spatially or temporally:

1. Extract aesthetic features frame-by-frame.
2. Frame-level features are pooled within each subshot using 6 different pooling techniques, generating 6 subshot-level features.
3. The subshot-level features are pooled across the entire video and a set of 12 video-level features is generated for each of subshot.

2.2.1.4 Webpage Assessment

With the development of Internet, it has become an important part in people’s daily life. Web pages serve as the user interfaces of the Internet, and there is an increasing need to design visually appealing Web pages. Researchers in multiple disciplines have laid emphasis on the aesthetics of web pages. Michailidou et al. [MHB08] investigate into user perception of the visual complexity and aesthetic appearance of Web pages. The results show a strong and high correlation between users’ perception and aesthetic appearance of a Web page. Studies also show that visually appealing web pages are perceived to be easier to use and access, and aesthetic web pages are usually judged as having more credibility. Efforts have been made to automatically compute the visual quality and aesthetics of web pages. These generalized automatic

models have a wide range of web-based applications including

- Web search. Current search engines return results based on web pages' relevance. Corresponding factors include content relevance, user feedback, page-rank score and so on. However, since studies show that Web page of higher visual quality are more appealing to the viewers, search engines based on aesthetic evaluation could offer a more user-friendly scheme.
- Web design. An objective web page visual quality evaluation system is needed for designers which can help them to rate the aesthetic quality during the design stage. The feedback from the system can further help designers to reduce individual bias on the aesthetic impressions.
- Web advertisement. Visually appealing web pages are perceived to be easier to use and access. It should also provide useful information for publishers to decide on the advertising scheme. Advertisements placed on web pages of higher aesthetic quality can therefore be expected to bring in positive user responses.

Wu et al. [WCLH10] proposed the "Visual Quality" (VisQ) system to evaluate the aesthetics of web pages. Classic aesthetic evaluation systems adopt the learning framework, i.e extracting discriminative features and train the model for a classifier or regression function. VisQ follows a similar framework. Web pages are taken as semi-structured images, and extracted discriminative features are categorized into four classes. Details of these features are listed in Figure 2.13. The construction of the VisQ evaluation system is formalized into a multi-cost-sensitive learning problem in terms of classification and a multi-value regression problem in terms of scoring.

Layout	Text	Classical Visual Features	Visual Complexity
size, number and layers of blocks	number, area, and density of text blocks	Color, texture	the compressed size of screen-shots

Table 2.13: Features of VisQ [WCLH10].

Singh et al. [SB10b] focus on the aesthetic evaluation of web page interface. They do not take Web pages as an image, but the combination of objects (text, dialogue box, images, buttons, etc.) These objects are evaluated from 6 aesthetic-related aspects:

- **Balance** computes the difference between total weighting of objects on each side of the horizontal and vertical axis.
- **Equilibrium** computes the difference between the center of mass of the objects and the physical center of screen.
- **Symmetry** examines if the objects are placed symmetrically in three directions: vertical, horizontal, and diagonal.
- **Sequence** measures the level of coherence between information layout and the common reading pattern (upper to lower, left the right).
- **Rhythm** evaluate an interface by taking into account the number and dissimilarities of interface objects.
- **Order and Complexity** is the weighted sum of the above measures.

These features are linearly fused for the result rating. The evaluation system relies on the accurate recognition of Web page objects, which is not a plain task especially for web pages constructing using CSS.

	Appealing	Simple	Professional	Captivating
Balance	✓		✓	✓
Symmetry	✓			✓
Equilibrium		✓	✓	

Table 2.14: Statistically significant correlations between features and patterns [ZCLR09].

To focus on the aesthetic influence of low-level features on the web pages, Zheng et al. [ZCLR09] try to minimize any preconceptions on the web page content. Therefore, popular web pages or high traffic sites (such as apple.com, facebook.com) are excluded. Moreover, web pages containing emotional objects - such as a baby's face - or familiar objects - such as the iPhone - are also excluded from their database. Similar to Singh et al. [SB10b], they also consider balance, symmetry and equilibrium in addition to low level features of the webpages. And the experimental results of their features are listed in Tab 2.14.

2.2.1.5 Summary

A summary of all the automatic aesthetic assessment works described above is provided in Table 2.3. Most existing approaches do not consider the compression and distribution distortions of images. Other degradations, including blur and noise, are either taken as indicators of composition or assumed to be not existing. On one hand, the selection of database, the extraction and interpretation of aesthetic rules become the major differences between different approaches. On reliability of the classic features extracted by most assessment

approaches, Loui's group [CL09] has studied a few low-level attributes that are usually believed to be related to the perception of aesthetics. Their study provides some broad categories of the low level features, not even comprehensive. It appears that features related to image distortion, such as color and sharpness, are still considered as attributes related to artistic quality.

2.2.2 Aesthetic Enhancement

Media-quality assessment and enhancement are two closely-related areas in computing aesthetics. A direct extension of automatic aesthetic assessment is to enable users to improve the visual aesthetics of their media works under the guidance of media aesthetic theories. We will consider the issue from two aspects: authoring and enhancement. The former one applies media aesthetic theories to guide media production, while the latter tackles the problem of enhancing the quality of existing media works.

Applied media aesthetics begins at the analysis of basic aesthetic elements, extends to the understanding of their contextual functions, and aims to examine how they can effectively generate and intensify the impact of media products. Therefore, the ultimate objective of media aesthetics is to guide the media production. The most straight-forward interpretation of "intensifying the impact of media products" is to make the products more visually pleasant. From this point of view, most of the media processing techniques are related to computational aesthetics, such as painterly rendering [ZZXZ09], color style transfer [GH05], and image abstraction. Just like the painter tools in Adobe Photoshop, for example, these computational models studies how image features of professional works could be replicated to benefit the aesthetic enhancement of other media products. Their focus is on "how" instead

	Distortion		Aesthetic Features				Content	S.F.	Classifier	Training Source
	Blur	Noise	Composit.	Color	Bright	Contrast				
Ke [KTJ06b]	✓		✓	✓	✓	✓			Bayesian	DPChallenge
Niu [NL12]	✓	✓		✓	✓				Bayesian,	A shot collection
									SVM	
VisQ [NL12]			✓	✓	✓				SVM, SVR	Selected web pages
Khan [KV12]			✓	✓	✓	✓	✓		SVM	The human photo data set
Datta [DLW08]	✓		✓	✓	✓				SVM	DPChallenge
Yang [YYC11]			✓	✓	✓	✓		✓		
Luo [LT08]			✓	✓	✓	✓		✓	Bayesian, SVM, Ad-aBoost	DPChallenge.com
Wei Luo [LWT11]			✓	✓		✓	✓	✓	SVM	CUHKPQ

Comments:

- **Abbreviations:** S.F: Semantic-related features. Composit.:Composition.
- **References:** Ke [KTJ06b], Niu [NL12], VisQ [WCLH10], Khan [KV12], Datta [DLW08], Yang[YYC11], Luo [LT08], Wei Luo [LWT11]

Figure 2.3: A summary of the extracted aesthetic features in the media assessing systems.

of “why”. It differs from computational media aesthetics, which aims to interpret the function of aesthetic elements and the roles they play in manipulating our perceptual reactions.

Media assessment quantifies some aesthetics-related criteria, such as color combination, object composition, and the applications of special effects. Researchers apply these rules to evaluate different media, but how to use these abstract criteria to improve the aesthetic quality of products is a different problem. For example, based on the study of aesthetic assessment, it has been revealed that photographic compositions can trigger several psycho-visual stimuli, due to which the photograph is perceived to be of high quality. Given a consumer photograph, it is possible to detect the salient regions, either based on the semantic content or on visual features. Then we can apply aesthetic assessment to see if it follows the Rule of Thirds and decide the level of compositional quality. However, if we find that the photo fails to follow these compositional rules, how can we make changes and improve the aesthetics of the input photo? This is not a stand-alone problem of media aesthetics, but the integration of graphic techniques including image segmentation, matting, and inpainting.

Coming back to the problem of image aesthetics enhancement, Bhattacharya et al. [BSS10] propose a photo quality assessment framework together with the visual-aesthetics enhancement. They focus on outdoor photographic compositions with one or more foreground subjects or compositions with no dominant foreground subjects. They relocate the objects to a more aesthetically pleasing location to tackle the former, and crop or expand the photograph for an aesthetically pleasing balance between sky and land/sea when encountering the latter case. The quality of automatic object detection,

segmentation and inpainting is still problematic, especially prescribing current automatic approaches. They avoid this issue by allowing user-guided object segmentation and inpainting to ensure that the final results match the users' preference. They deal with the composition of one foreground object, and they further improve the system and facilitate it with interactive selection of more than one objects in [BSS11]. Their system shows a standard framework of media aesthetics application which could be summarized as follows:

1. Extract aesthetic-related features.
2. Apply machine learning to train models of aesthetic rules.
3. Measure the deviation of the features of input media data from those of the aesthetic ones.
4. Rectify the input data by aesthetic-related post-editing techniques, including segmentation, matting, inpainting.

Zhang et al. [ZCC11] consider the issue of aesthetic enhancement of landscape photographs. They rectify the characteristics of saturation and luminance to enhance the depth perception of scenes. The processing goes in the gradient domain of the LCH space, because the changes of the gradient field directly correspond to contrast. In their work, all paintings and photos are partitioned into regions corresponding to the foreground, the middle-ground, the background, and the sky region manually. Statistical features of the source images are rectified based on the reference ones. Their algorithm produces excellent results without severe artifacts. Essentially speaking, it is a promotion of image style transfer because they do not consider the inherit aesthetic impact of corresponding features.

In addition to single image processing, Adams et al. [AV05] propose an

algorithm for automatic content-sensitive injection and repair the important video aesthetic element, i.e. the visual tempo, which is directly related to the emotion and impact of video clips. Generally speaking, it is assumed to depend on the shot length and motion characteristics. Tempo rectification is an important post-processing technique especially for home videos, because amateur consumers often lack the time or knowledge required to fashion video compositions that faithfully communicate their experience of an event. In the proposed framework, extracted tempo is modeled as the integration of shot length and motion intensity. Professional movies are taken as the reference samples. The system automatically repairs the aesthetic elements related to the visual tempo and recover the intended signals based on the deviation of the input video tempo from the desirable models.

Lots of research about media quality enhancement is related to aesthetics. For example, contrast restoration [NN03] and color transfer [RAGS01]. Some focus on the replication of features that could benefit the aesthetic enhancement of current media products ([NN05]), and do not take the inherent aesthetic functionality into consideration. An ideal framework of media aesthetic enhancement is based on the aesthetic interpretations of corresponding elements. The output quality is often influenced by the accuracy of the adopted graphic algorithms, including segmentation, matting and inpainting. Such dependency is common especially for compositional rectification. As a result, many algorithms choose a semi-automatic solution to reduce the influence. Anyway, media post-processing based on computational media aesthetics depends on many other processing techniques, and these unfavorable conditions constrain the practicality and dramatically increase the difficulty of corresponding research.

Comparing with aesthetic media post-processing, more work has been done in media authoring based on media aesthetics. Because media authoring, which fortunately avoids these problematic issues, is a more straight forward application for media aesthetics. For example, Sandhaus et al. [SREB10] propose an application of media aesthetics which follows a very classic framework. They apply the widely-known compositional rule, The Rule of Thirds, to split pages of the output photo book into sub-areas. Their work shows how the basic elements of media aesthetics can benefit the quality enhancement of media products. For video authoring, Masahito Kumano *et al.* [KAA⁺02] selected appropriate shots and connected them together based on film grammar. They listed four video editing rules, including shot size, camera work, and combination criteria. Shots' contents were then detected.

Achanta *et al.* [AYK06] combined video editing and intent modeling. They used some basic media elements, such as color, contrast, brightness and camera motion, to model different video grammars. They did not try to build up any storyboard, but they manipulated media elements of video to map original video clips to four kinds of intent: cheer, serenity, gloom and excitement.

2.3 Discussions

Computational media aesthetics starts at the extraction and interpretation of basic aesthetic elements. We examine the corresponding feature analysis algorithms in the first section. Based on the extracted features, efforts made in aesthetic assessment become straightforward. These algorithms look into the functionality of media elements, and try to discover how they are incorporated to influence media aesthetics. The assessment systems differentiate media

of low aesthetic quality from those of higher quality, and the next stage of research topic is to apply these rules to either guide the producing process or to improve the aesthetics of given media products. These works have been discussed in the third section.

Since the thesis focuses on the applications of computational media aesthetics, we only consider the general issues that are related to the applications in this chapter: feature extraction, feature interpretation and the aesthetic modeling. We consider the most widely used features and the ways to extract/interpret them. To summarize, there is existing work on aesthetic-related feature extraction and interpretation, the models building for aesthetic evaluation and the aesthetics-based applications on media enhancement. However, some problems are still open and not well studied.

Generally speaking, there are some issues in the current research areas of computational multimedia aesthetics.

- *Build effective aesthetic models.* The very initial but very important step of media aesthetics is the extraction and interpretation of basic media elements. Various aesthetic assessment systems proposed at present are still grappling with this issue. Extensive efforts are still required for the aesthetic models that integrate media elements from different sources, for example the correlation between visual and audio information.
- *Widen the applications of media aesthetics.* Applying media aesthetics on the assessment and authoring of media products has been widely studied, but aesthetic enhancement is still difficult to handle, because it is highly dependent on other image processing techniques such as matting and inpainting.

- *How to make aesthetic criteria to benefit media processing.* Computational processing of media often requires intensive computational work. Since media aesthetics looks into the relationship and interpretation of basic media elements, the constraints placed by media aesthetics criteria can reduce the computational complexity in return.

The objective of this dissertation is to build up aesthetic models that address these problems by proposing several typical applications of media aesthetic in different areas.

Single Image Aesthetics: Hazy Image Enhancement based on the Full-Saturation Assumption

Haze is a common image degradation, which occurs when photos are taken on foggy days. In this chapter, we present an algorithm that can successfully improve the quality of hazy images and offer visually-pleasant haze-free results with vivid colors. The notion of “vivid colors” is related to the visual quality from an aesthetic point of view. We propose the full-saturation assumption (FSA) based on the aesthetic photographic effect: photos of vivid colors are visually pleasant, and first recover the degraded saturation layer. The depth image is also obtained as a by-product. We then apply an example-based approach to avoid over-saturation.

This chapter show how properly utilizing aesthetic theories can help to improve the solution of traditional media processing problem. The full saturation assumption is based on the artistic guidance for photography, which alleges that photos of vivid colors are more pleasant. Based on this assumption, we can dramatically simplify traditional ill-posed under-constraint image processing problems and offer acceptable results.



Figure 3.1: The left shows an image free of haze. The right one is taken on a foggy day and degraded by haze.

3.1 Introduction

Light gets absorbed and scattered while it travels through the air. Light absorption refers to the process that reduces the light intensity while it interacts with matter. Light scattering is the physical process that makes photons refract in different directions [Sha03]. Rays are attenuated by both effects while traveling through the atmosphere. The light attenuation effect is influenced by the wavelength of light, the size of particles and thickness of the matters. The ray reflected from the objects surface is attenuated and mixed with airlight when it reaches the observers. The more foggy the days are, the more serious light attenuation will be. Thus the quality of images taken under such air conditions is greatly degraded. The images become foggy, which could be modeled as a function of scene albedo and depth $d(\mathbf{x})$ [Fat08]. Image dehazing algorithms try to restore these images by recovering the color and details. Figure 3.1 shows a comparison between a hazy image and an image free of haze.

The observed ray can be divided into 2 components [SNS06]: the attenuated signal that reflects from the object L_{obj} (so-called direct transmission),

and the airlight L_{inf} :

$$L(\mathbf{x}) = L_{obj}(\mathbf{x})t(\mathbf{x}) + L_{inf}(1 - t(\mathbf{x})) \quad (3.1)$$

where

$$t(\mathbf{x}) = e^{-\rho(\mathbf{x})d(\mathbf{x})} \quad (3.2)$$

$\rho(\cdot)$ is known as the attenuation coefficient, which is often assumed to be a constant over the entire scene ([KN09]). $d(\cdot)$ is the scene depth, and \mathbf{x} is the scene point corresponding to an individual image coordinate. L_{obj} is the image taken under a clear, haze-free condition, and L_{inf} is assumed to be the value of airlight at a non-occluded horizon.

Equation 3.1 is the widely accepted haze model ([SNS06] [SA07] [Tan08] [Fat08] [HST09] [ZLY⁺10].) It describes the haze process as a linear multiplication of object reflecting rays L_{obj} and airlight L_{inf} .

For a given haze image I , image dehazing aims to recover the underlying haze-free image. According to Equation 3.1, it is an under-constrained problem. Different constraints have been put forward to reduce the parameter ambiguity and offer desirable results. And the first question we need to answer is - "What does a haze-free image look like", i.e. the features that need to be recovered from a hazy image. It can be used as a constraint when we try to solve Equation 3.1.

From Equation 3.1, the process of light scattering reduces image contrast and saturation. As a result, contrast recovery is closely related to image dehazing. Schemes have been proposed to enlarge image contrast so as to restore the degraded images. In the proposed dehazing algorithm, instead of considering contrast, we place constraints on the saturation properties of

haze-free images. The proposed algorithm is based on the observation that an image with bright and saturated colors is visually attractive to observers. So we assume that for a haze-free high quality image, the saturation of most color patches is high. Such assumption corresponds to the Dark Channel Prior [HST09], that assumes that most local patches of haze-free images have some pixels which have very low intensity values in at least one color layer in the RGB space. When the dark channel prior holds, $\min(R, G, B)$ is low for most patches, which indicate a high saturation value according to Equation 3.1. One of the advantages of such assumption is that saturation provides a visual quality measurement.

In this chapter, we present a novel method for single image dehazing. Instead of considering the properties of transmission rays, which could help decompose the perceived rays, we place constraints on the visual features of the desirable haze-free images. A simple image upsampling is performed to estimate the transmission map. We test the proposed algorithm on a number of foggy images, and the results show pleasant haze-free outputs.

3.2 Previous Work

The dehazing issue arouse from the desire to see objects in bad weather with early dehazing approaches requiring a set of input images under different haze conditions ([NN00] [NN03].) [SNS06] [SA07] use a pair of polarized images which are captured using polarized filters. As the dehazing problem is inherently under-constrained, researchers have utilized constraints from different images of the same scene. These algorithms offer acceptable results but the availability of input data is not a trivial task.

In the context of computational photography, a recent trend is to restore degraded images and extract other meaningful quantities with minimal input data. There has been significant progress in the area of single image dehazing. Since haze removal is a highly ill-posed problem, prior assumptions on the haze-free image play an important role.

[Fat08] considers the single image dehazing problem based on independent component analysis. Fattal maps the observed color vector onto the direction parallel to the airlight and the other direction which is orthogonal to the airlight. The albedo value is assumed to be locally constant. Therefore, the airlight-albedo ambiguity problem is reduced to determining a single scale for the entire image. He clarifies this ambiguity by assuming that the object shading and scene transmission are locally uncorrelated.

Tan’s work [Tan08] aims to enhance contrast and improve visibility of an input hazy image. He proposed two observations: the clear image ought to have higher contrast than the haze-degraded image, and the variance of airlight tends to be smooth. Based on these two observations, he maximizes the contrast in a local window under the framework of markov random field.

An interesting prior based on the statistics of haze-free images, namely, the dark channel prior has been proposed in [HST09]. He et al. assume that for most non-sky haze-free natural images, at least one color channel has very low intensity at some pixels. Then the minimal intensity value of all the channels within a local patch is brought in by the airlight. They perform an initial rough estimation of the transmission map based on the dark channel prior. Then they use soft matting to refine the initial results.

[CH09] considers the influence of bright objects when estimating the transmission map. They assume that neighboring objects should have similar

depths and objects on the bottom are nearer than those on the top.

[ZLY⁺10] uses a bilateral filter to refine the transmission map. They assume that chroma variance over large scale is resulted from transmission while local chroma variance is due to scene albedo. Thus the transmission map is blurred out over large scale while the sharp edges are preserved which indicate local scene albedo changes.

However, the above mentioned techniques do not ensure the completeness of haze removal. Without a clear description of the haze-free image features, these methods can efficiently remove the haze to some extent, but they cannot ensure the completeness of haze removal in the output image. In He's work [HST09], for example, they use the dark channel prior to get a coarse estimation of the transmission map. But when they refine the initial map, a global smoothness constraint is adopted, which makes the output result divergent from the dark channel prior.

Our main contribution is the simple criterion, which can easily produce haze-free images of good quality and the same criteria can be used to evaluate the quality of dehazing process.

3.3 The HSI Color Space and the Dehazing Problem

The HSI color space is based on the cylindrical-coordinate, which re-arranges RGB values from a more perceptual relevance, where H, S and I layer represent hue, saturation and intensity respectively. The conversion from RGB color

space to HSI is given by [GW07]

$$\begin{cases} H &= \arccos\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\} \\ S &= 1 - \frac{3}{R+G+B}\min(R, G, B) \\ I &= \frac{1}{3}(R + G + B) \end{cases} \quad (3.3)$$

To remove the singular points in the HSI space, we adopt the solution:

$$H \stackrel{\triangle}{=}_{R=G=B} 0 \quad (3.4)$$

$$S \stackrel{\triangle}{=}_{R=G=B=0} 0 \quad (3.5)$$

We rewrite Equation 3.1 and denote the standard observation model for the formation of a hazy image as

$$\mathbf{I}(\mathbf{x}) = \mathbf{J}(\mathbf{x})(1 - t(\mathbf{x})) + \mathbf{A}t(\mathbf{x}) \quad (3.6)$$

where \mathbf{x} denotes the pixel coordinate, \mathbf{I} is the observed intensity of the input foggy image, \mathbf{J} is the underlying haze-free image, \mathbf{A} is the airlight vector and t denotes the transmission coefficient. The above model is specified in the RGB color space. We map it onto the HSI color space. Let H_I, S_I, I_I denote the observed pixel value of the hazy image, and H_J, S_J, I_J denote the underlying haze-free image value in the HSI space.

To simplify the problem, we assume that the airlight A is achromatic. Otherwise, we can rewrite the problem 3.6 in another form:

$$\bar{\mathbf{I}}(\mathbf{x}) = \bar{\mathbf{J}}(\mathbf{x})(1 - t(\mathbf{x})) + \bar{\mathbf{A}}t(\mathbf{x}) \quad (3.7)$$

where

$$\begin{cases} \tilde{\mathbf{I}}(i) &= \frac{\mathbf{I}(i)}{\mathbf{A}(i)}, i \text{ is the R/G/B layer} \\ \bar{\mathbf{I}} &= \frac{\tilde{\mathbf{I}}}{\max(\tilde{\mathbf{I}})} \\ \bar{A} &= \frac{1}{\max(\tilde{\mathbf{I}})} \end{cases} \quad (3.8)$$

The second equation in Eqn. 3.8 ensures the correctness of mapping from RGB color space to HSI color space, resulting the new airlight \bar{A} being achromatic.

According to the dehazing model (Equation 3.6) and the conversion equations between RGB and HSI (3.3), we get

$$\begin{cases} H_I(\mathbf{x}) &= H_J(\mathbf{x}) \\ S_I(\mathbf{x}) &= \mathbf{t}(\mathbf{x}) \frac{I_J(\mathbf{x})}{I_I(\mathbf{x})} S_J(\mathbf{x}) \\ I_I(\mathbf{x}) &= \mathbf{t}(\mathbf{x}) I_J(\mathbf{x}) + (1 - \mathbf{t}(\mathbf{x})) \bar{A} \end{cases} \quad (3.9)$$

When airlight is achromatic, i.e. the values of A in the three RGB layer are the same, the hue layer remains the same after the haze degradation. For the saturation layer,

$$\frac{\mathbf{t}(\mathbf{x}) I_J}{I_I} = \frac{\mathbf{t}(\mathbf{x}) I_J}{\mathbf{t} I_J + (1 - \mathbf{t}) \bar{A}} \leq 1 \quad (3.10)$$

so that $S_I \leq S_J$. Therefore, haze does not influence the hue of the image, but degrades the saturation and intensity layers of the original image and results in low saturation and contrast. The fact that the haze artifact reduces the saturation level coincides with our intuitive understanding. When pure colors (full-saturated) are mixed with achromatic ones, the saturation decreases. For the intensity layer I_I , it follows the standard haze model.

3.4 Full-Saturation Assumption

When we talk about the visual pleasure of color images, *vivid colors* are often desirable. [JRW97] discussed the necessity of producing vivid colors for higher visual pleasure. We think images with vivid colors are pleasant because human perception re-constructs the visual objects with vivid colors. Thus in our proposed algorithm, we assume that the underlying haze-free image is of vivid color. However, the definition of vivid colors is vague and subjective. Generally speaking, bright colors with high saturation are thought to be vivid. Consider the functionality of saturation from a perceptual point of view, it influences the level of purity and vividness of a color [Sha03]. Moreover, a desaturated image is said to be dull. Figure 3.2 shows an example of natural vivid image and most of its saturation values are close to one. Our proposed assumption is only suitable for natural images, and works especially well for natural outdoor scene. Figure 3.3.(a) shows the distribution of the local maximum saturation value for 1000 clear landscape images from DpChallenge dataset. The saturation value for most macroblocks are either close to 1 or close to 0. Figure 3.3.(b) shows a counter example. The 1500 images are indoor still objects with post-processed color effects. And the full saturation assumption fails in this case.

3.5 Relations with Dark Channel Prior

As mentioned in the previous sections, there are some existing approaches for image dehazing. Here we will discuss the relations with previous solutions, especially the dark channel prior [HST09].

The dark channel prior assumes that for some pixels in a haze-free natural



Figure 3.2: A sample natural image of vivid color. (a). The natural image. (b). The saturation layer.

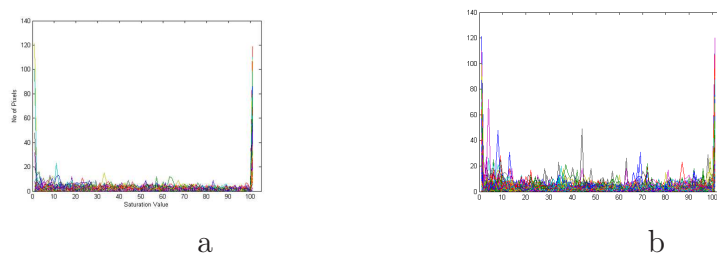


Figure 3.3: Distribution of local maximum saturation. (a). The natural outdoor scene. (b). Indoor objects with post-processed color effects.

image, at least one color channel will have very low value, i.e. for these pixels, $\min(R, G, B) \leq \varepsilon$.

The saturation value in the HSI color space (Equation 3.3) is given by

$$S = 1 - \frac{3}{R + G + B} \min(R, G, B) \quad (3.11)$$

When the full-saturation assumption holds, for every image patch Ω , there will be at least one pixel whose saturation value is close to 1

$$S_J^{max}(\Omega) = \max_{\mathbf{x} \in \Omega} S_J(\mathbf{x}) \approx 1 \quad (3.12)$$

The alternative expression of Equation 3.12 is $\min(R, G, B) \approx 0$ (according to Equation 3.11), and dark channel prior holds. However, from Equation 3.11, we can find that low value of one channel alone can not ensure a high saturation value of a pixel. It is not the minimal value but the ratio between the minimum and the mean of three RGB layers (the intensity layer in HSI color space as in Equation 3.3). Therefore, the proposed full saturation assumption is actually a stronger assumption than the dark channel prior, which takes the visual pleasure of output images into consideration.

Now we can see the impact on the output dehazed images under the two assumptions. Figure 3.4 gives an example how intensity influences our perception on chroma information. Image a1 and a2 are two colors (0.1,0.1,0) and (0.8,0.8,0.3) in the RGB color space. According to the numeric definition of saturation, a1 is fully saturated while a2 is unsaturated. But the intensity of a1 is too low to present any chromatic information, or show its “vivid color”. a2, on the contrary, appears more colorful and eye-striking to the viewers. So when we try to restore the chroma of degraded hazy images, intensity is

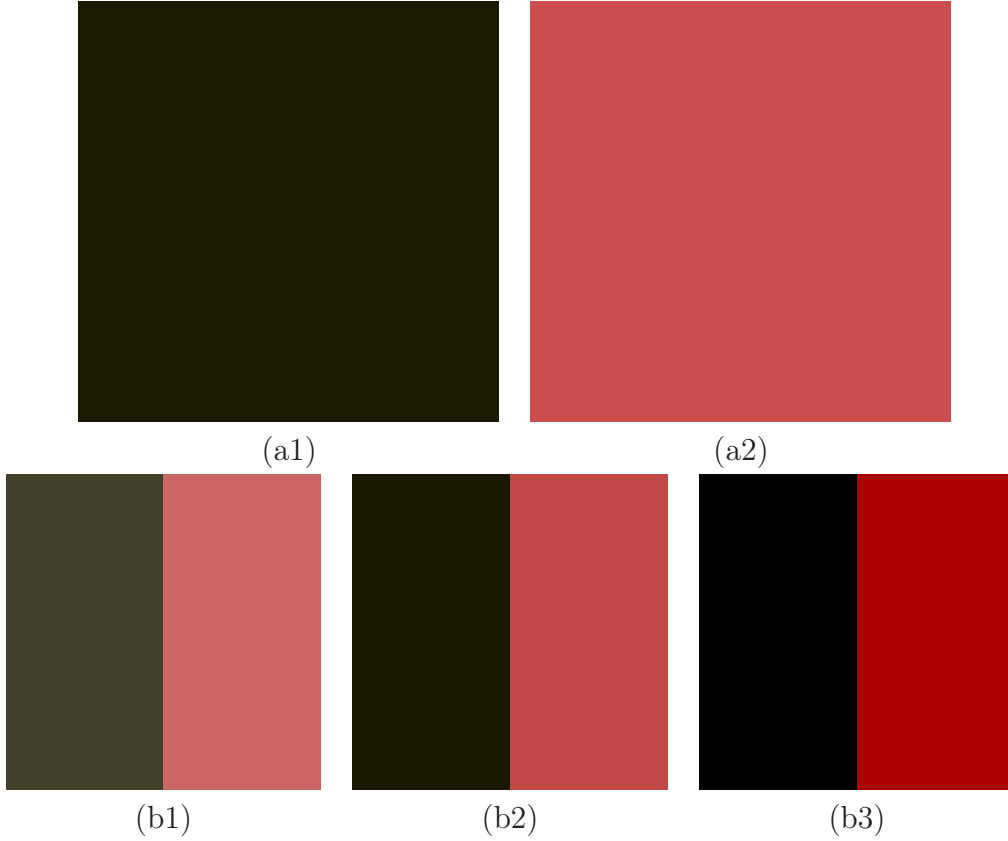


Figure 3.4: Color saturation under different Intensity.

another factor in the coarse transmission map generation.

Image (b1)-(b3) in Figure 3.4 shows a further comparison between the proposed assumption and the dark channel prior. b1 is a synthesized hazy image with color $(0.25, 0.25, 0.17)$ on the left and color $(0.8, 0.8, 0.4)$ on the right. According to the dark channel prior, the initial guess of the transmission coefficient t is 0.17, which gives the dehazed image b2. The saturation of b2 is 0 on the left and 0.367 on the right. Even though the dark color is fully-saturated, the red color on the right is under-saturated. Actually, the red color has higher intensity and is more eye-catching than that on the left, so it still seems to be dull and impure. b3 shows the result under the proposed full saturation assumption, and it produces a more saturated output color,

especially on the left (algorithm details will be discussed in the next section). The saturation of b_3 is 1 on both side, and under the proposed assumption, haze has been completely removed.

Therefore, the proposed full-saturation assumption takes visual pleasure into consideration, and is a stronger assumption than dark channel prior. It ensures more vivid output color patches, which are of higher saturation. We do not allege that we try to recover the true underlying haze-free image, but one that is visually pleasant. The proposed assumption is made for image enhancement, not for restoration. Moreover, photography itself does not always truly, sincerely record the real world. The photos produced by Nikon and Canon may be quite different - the former one is well-known for producing bright and vivid colors.

3.6 Our Example-based Approach

To restate our assumption in a more formal way, let Ω denote a macroblock of an image,

$$\max_{\mathbf{x} \in \Omega} S_J(\mathbf{x}) \in \{0, 1 - \varepsilon\} \quad (3.13)$$

In a vivid, haze-free image, most chromatic macroblocks have pixels which are fully saturated. In our experiments, we set ε to 0.001 for system numerical stability.

We solve the transmission coefficient map t from Equation 3.9

$$t(\mathbf{x}) = 1 - (I_I(\mathbf{x}) - \frac{S_I(\mathbf{x})}{S_J(\mathbf{x})} I_I(\mathbf{x})) / A \quad (3.14)$$

The equation is under-constrained, because there are two unknowns. As long

as we can obtain a good approximation for S_J , Equation 3.14 could be solved immediately. Based on the full-saturation assumption (Equation 3.12), we can assume the colors of haze-free images are vivid. For each macro-block Ω ,

$$\max_{\mathbf{x} \in \Omega} S_J(\mathbf{x}) = 1 - \varepsilon \quad (3.15)$$

where ε is a small number to avoid over-saturation of the recovered colors. This is the target value of the output maximal saturation. Bright colors, such as green, red, and orange, are more likely to be over-saturated than cold colors such as blue. Therefore, we choose different ε for different colors. In practice, we divide the RGB color wheel into 12 clusters, clockwise from red-orange to red. For each color cluster, we find the corresponding ε from the characteristics of sample images. We have collected around 400 color images which are labeled vivid, and estimated the maximal saturation of different color clusters in these images. We compare the color distribution of the input hazy image (hue is assumed to be un-degraded) and those of the images in the dataset. The optimal ε is chosen to be the one that has the most similar color distribution with the input image.

Haze degradation reduces image saturation (Equation 3.10), so in the macroblock $\forall \mathbf{x} \in \Omega$, $S_I(\mathbf{x}) \leq \eta$. The corresponding pixel indices are given by

$$\tilde{\mathbf{x}}_\Omega = \operatorname{argmax}_{\mathbf{x} \in \Omega} S_I(\mathbf{x}) \quad (3.16)$$

We can then solve t from Equation 3.14 at pixel $\tilde{\mathbf{x}}_\Omega$

$$t(\tilde{\mathbf{x}}_\Omega) = 1 - (I_I(\tilde{\mathbf{x}}_\Omega) - \frac{S_I(\tilde{\mathbf{x}}_\Omega)}{1 - \varepsilon} I_I(\tilde{\mathbf{x}}_\Omega)) / A \quad (3.17)$$

For each macro-block of the input image, we can solve $t(\tilde{\mathbf{x}}_\Omega)$. It is a down-sampled transmission map. Notice that the saturation value at the singular point $(0, 0, 0)$ is intentionally set to 1. Therefore, these pixels inherently satisfy the full-saturation assumption and are selected as the downsampling targets with priority. We upsample $t(\tilde{\mathbf{x}}_\Omega)$ to get the original map by the joint bilateral filter [KCLU07], which is efficient in image-upsampling with sharp edges.

3.7 Experimental Results

In the proposed dehazing scheme, the value of airlight A is assumed to be known. A lot of airlight estimation schemes have been given in the previous image dehazing methods, such as [Fat08], [SNS06] and so on. In our experiment, we select a haze-opaque range and use its mean value as the airlight A . Several important parameters of our method are listed in Table 3.1. Their values have been tested to work for most images in our experiments.

We perform the proposed dehazing scheme on several real hazy images and compare the results with other methods. Figure 3.5 shows a general process of the proposed scheme. The input hazy image (a) is converted to the HSI color space, while (b) shows its saturation value. The objects near the camera are of high saturation, while saturation value of the further objects are lower because of haze, which corresponds to our assumption. The whole image is segmented into 16×16 macro-blocks. For each macro-block, the pixel of the highest saturation is selected and the corresponding transmission value t is computed. We get a down-sampled transmission map, which is shown in (c) and these selected pixels and their transmission value t are shown in (d). The joint bilateral filter is applied to upsample the image (c), i.e. to compute

Parameter	Value	Description
Ω	16×16	The size of color patches. It decides the downsampling ratio of initial transmission maps.
ε	0.001	Control the level of saturation and avoid over-saturation. Higher value reduces the global saturation in the output image.
δ_1	0.001	Control the smoothness of transmission map. Higher value can smooth the transmission map but results in hazing pixels near sharp boundaries.
δ_2	20	Help to preserve sharp boundaries in the transmission map.
Λ	20×20	The size of the joint bilateral filter. Higher value increases computational complexity, but it helps to recover hazy areas whose depth is different from surrounding, for example, the holes of background between leaves.

Table 3.1: Related Parameters

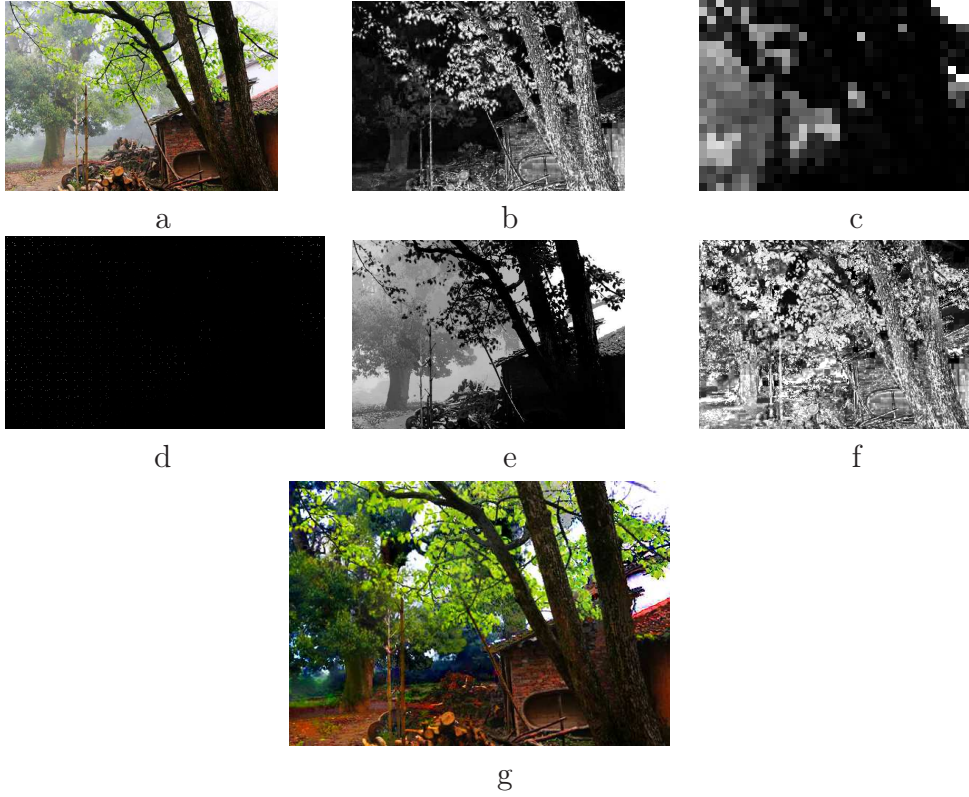


Figure 3.5: Haze removal result. (a) Input hazy image. (b) The saturation layer of the original image in the HSI color space. (c) The initial downsampled transmission map. (d) The corresponding pixel index of downsampled transmission map in the up-sampled map. The joint bilateral filter is performed on (d), and the estimated transmission map is shown in (e). (f) The saturation layer of the dehazed image. (g) The output haze-free image.

the value of blank pixels in (d). The result is shown in (e) while (g) is the output image. The objects in (g) are of high saturation which is justified by its saturation layer value in the HSI color space (f). Except the achromatic sky regions, the rest of the areas have their local maximal saturation value close to 1. Thus, according to the full saturation assumption, image (g) is a haze-free image. More experimental results are shown in Figure 3.6. The vivid colors have been successfully recovered in the output images.

Figure 3.7, Figure 3.8, and Figure 3.9 show the comparisons between the proposed dehazing method and those of previous methods. In Figure 3.7, we

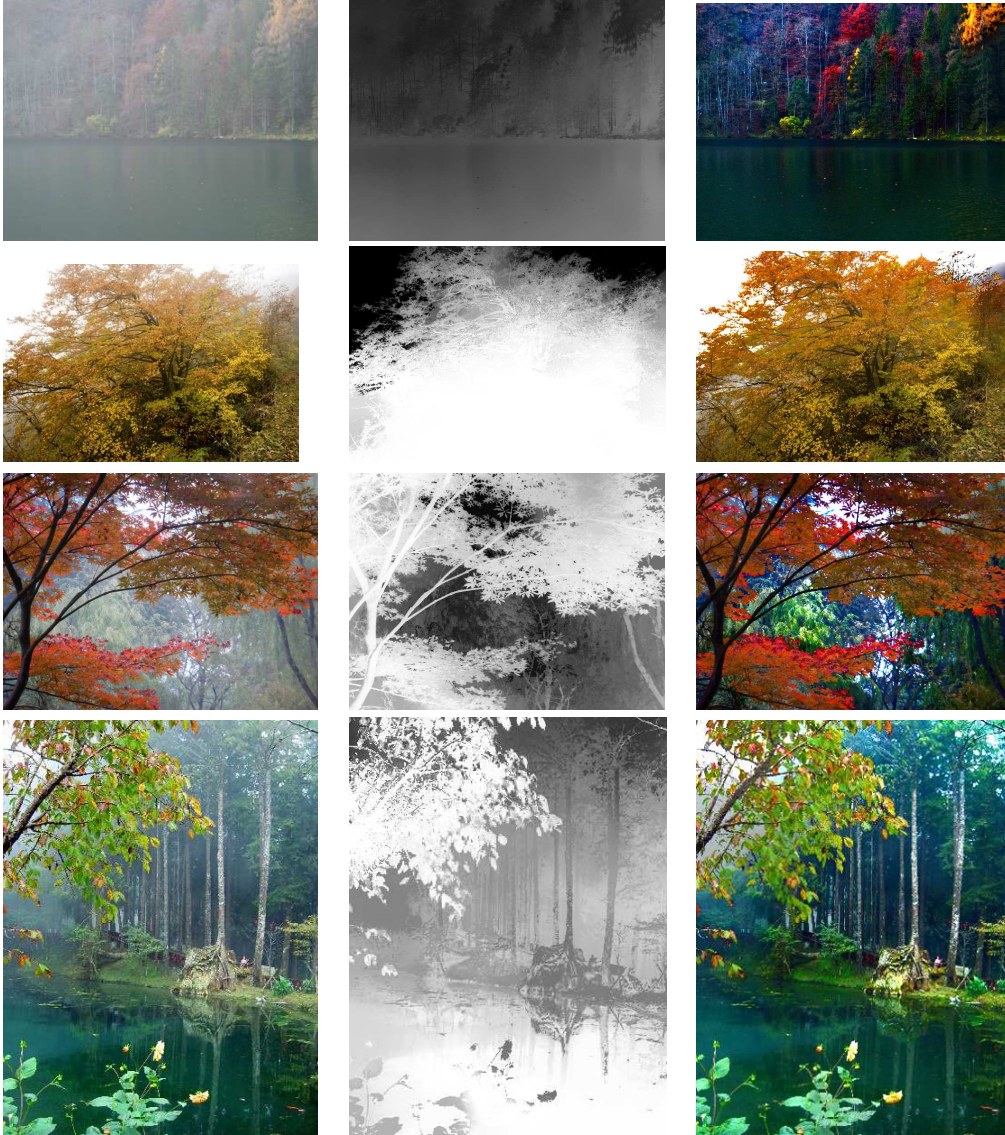


Figure 3.6: Haze removal results. First column: input hazy images. Second column: the transmission map. Third column: Output haze-free images.

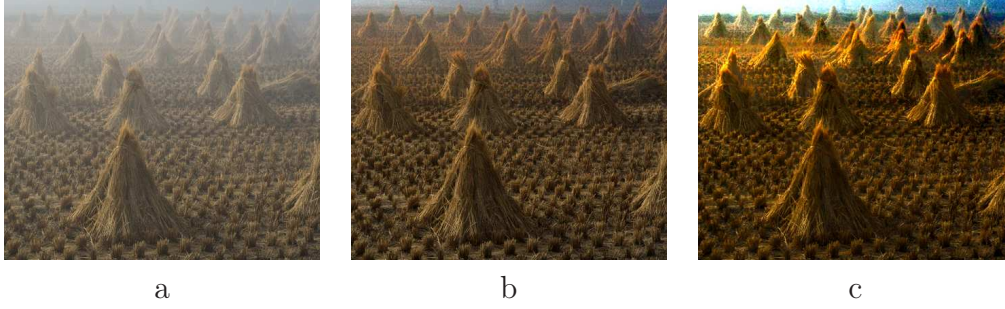


Figure 3.7: Comparison with He et al’s work [HST09]. (a) The input hazy image. (b) Dark channel prior. (c) Our result.

compare our result with Fattal’s and the dark channel prior. Our result is comparable to theirs, but ours is a little more saturated. On the upper right corner, the proposed algorithm recovers the dark regions a little better than the rest two methods, for the details are brighter and clearer. In Figure 3.8, we compare our result with the dark channel prior. The colors in our result are more vivid. And our image is of higher contrast.

We compare the proposed approach with the recent work by Zhang et al [ZLY⁺10] in Figure 3.9. Our result is more saturated, and has no apparent hazy degradation anymore. Observing the transmission map, the proposed method does a good job in recovering the depth information at the holes among foreground objects (upper-right corner) and the isolated foreground objects (the branch of leaves on the upper-left corner.) Moreover, our result successfully preserves the sharp boundaries between nearer objects and those further away.

Figure 3.10 shows a synthetic experimental result, where the proposed scheme can handle color patches well. The dense haze in the far away areas has been completely removed. But there is some color distortion on the foreground white walls.

Figure 3.11 shows a failing case of the proposed method. The color of



Figure 3.8: Comparison with others' work. (a) The input hazy image. (b) Fattal's result [Fat08]. (c) Dark channel prior [HST09]. (d) Our result.

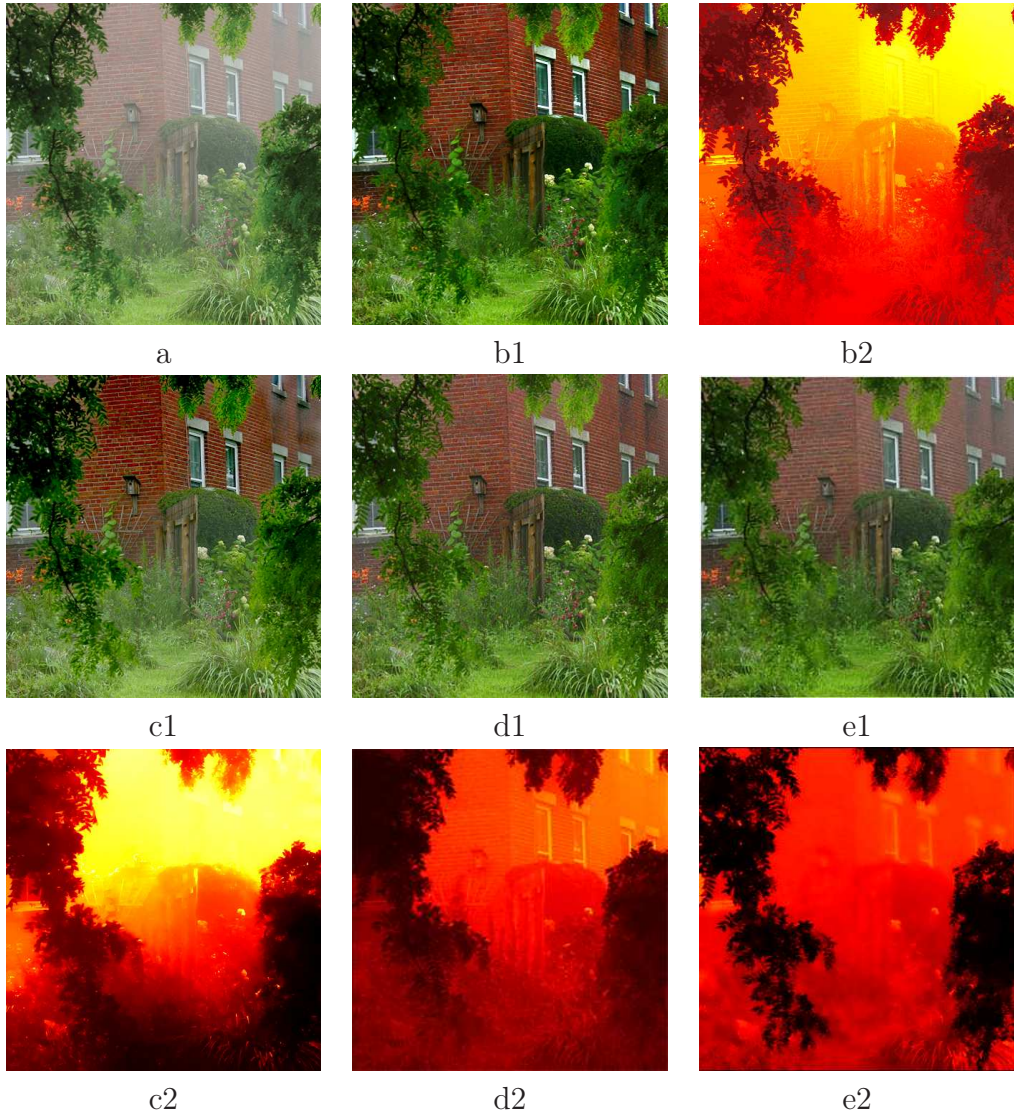


Figure 3.9: More comparisons with other work. (a) The input hazy image. (b) Our results. (c) Fattal’s results [Fat08]. (d) Dark channel prior [HST09], (e) Zhang’s results [ZLY⁺10]

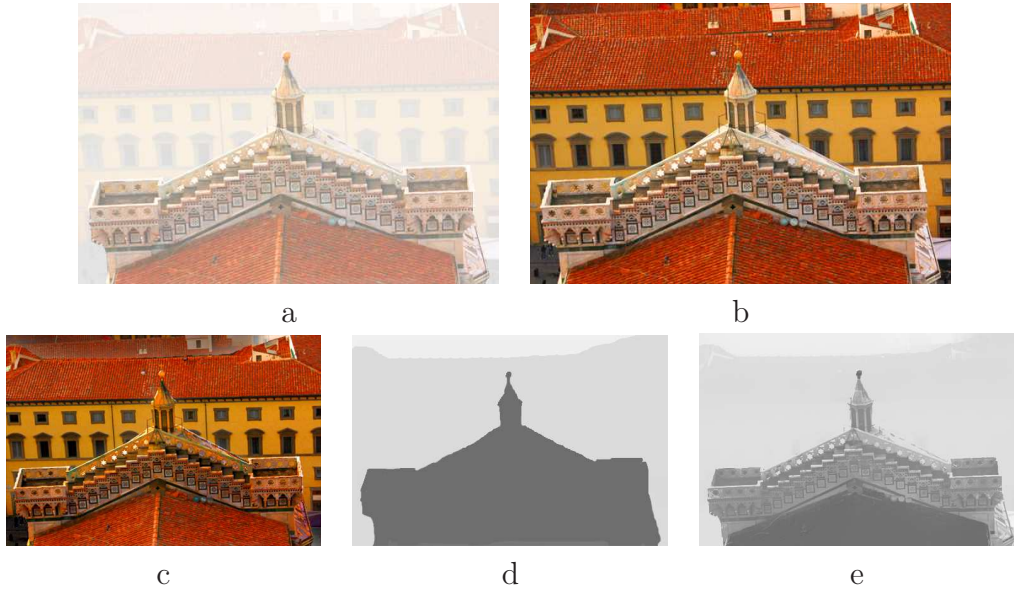


Figure 3.10: A synthetic experimental result. (a) the synthetic hazy image. (b) the ground truth image. (c) output haze-free image. (d) the estimated transmission map. (e) the ground truth map.



Figure 3.11: A failure case of the proposed algorithm. (a) Input hazy image. (b) Output image.

the cliffs is over-saturated and seems to be unnatural. Actually, the full-saturation assumption places constraints on the desirable properties of the output images, and transfer the hazy image to such condition. When the real underlying haze-free image does not satisfy such kind of assumption, the algorithm fails. The color of cliff in Figure 3.11 is undersaturated even if it is taken in a haze-free condition. Thus our assumption that any macro-block has some fully-saturated pixels is invalid, which results in an incorrect output result.

Such artifacts may be taken as color shifts as well. Color shifts in our experimental results are resulted from two factors: 1) over-casting airlight (3.5); and 2) object properties (3.11). On one hand, in our experiments, we assume that the haze is achromatic. Actually, the airlight is more or less bluish and the colors of objects in the background farway are corrupted by the bluish airlight to some extent. This color corruption is subtle and may be not obvious in most cases (as in 3.5) when the haze exists and acts as the dominant degradation factor. We have discussed this issue in the previous section: when brightness is too high or too low, hue and saturation shifts become less obvious. When we remove the haze, however, the color shifts become obvious and seemly exaggerated. To overcome this issue, proper color correction may be performed before we apply the full saturation assumption.

3.8 Discussions

In this chapter, we present a simple algorithm that can successfully improve the quality of hazy images and offer visually-pleasant haze-free images with vivid colors. The algorithm is based on the photographic assumption that a

natural haze-free image ought to have vivid colors. Applying this assumption, we can first recover the degraded saturation layer in the HSI color space and the dehazing process becomes direct and simple.

The assumption could also be used as an evaluation of the completeness of haze-removal. As shown in Figure 3.5, most pixel values of the saturation layer are very close to 1. According to the full-saturation assumption, the output image is vivid and haze-free.

The haze we discuss here is the artifact that caused by poor lighting conditions. However, haze can be also used as a special effect in photograph production, which could reveal 3-dimensional distance or help to create mysterious atmosphere. Our proposed dehazing approach is based on the aesthetic criterion of saturation, and its target is to remove haze for a natural clear image. Therefore, the output image may not be necessarily aesthetic from the special-effect point of view.

Just like the dark channel prior, the full-saturation assumption is a statistic, and it may not work for all images. When the underlying image does not satisfy such properties, the assumption becomes invalid, such as Figure 3.11.

Another drawback of the proposed method is that the estimation of the transmission map is highly correlated with the image features. It contains too many details of the image, which contradicts the fact that the map reveals the depth of objects and ought to be smooth. Improving this is an interesting problem for further research. Moreover, the haze imaging model (Equation 3.6) assumes a global constant airlight A , which may be invalid. In Figure 3.5, we can see that the trees at the back are almost blue in the restored image, which indicates that their hue has been seriously degraded by the airlight near the horizon. Therefore, hazing model that could handle more

complicated phenomena might be desirable to produce better results.

Aesthetics for Image Ensembles: A Synaesthetic Approach for Image Slideshow Generation

Image slideshow is a way to present a series of images. To make the slideshow visually pleasant, images should be selected and arranged under certain aesthetic criteria. In this chapter, we present a novel automatic image slideshow system that explores a new medium between images and music. It can be regarded as a new image selection and slideshow composition criterion. Based on the idea of “hearing colors, seeing sounds” from the art of music visualization, equal importance is assigned to image features and audio properties for better synchronization. We minimize the aesthetic energy distance between visual and audio features. Given a set of images, a subset is selected by correlating image features with the input audio properties. The selected images are then synchronized with the music subclips by their audio-visual distance. The inductive image displaying approach has been introduced for common displaying devices.

4.1 Introduction

To properly manage the large set of consumer photographs, one of the interesting problems is to select suitable images and exhibit them in a more enjoyable way. Photo album summarization aims to select a subset of photos from the entire collection. Many criteria have been put forward for a desirable selection scheme. Another issue is to introduce appropriate motion and neatly display the selected images. The artificial camera motion is known as the Ken Burns effect [Tho95]. It is a type of video production effect which uses panning and zooming to generate videos from still images. Many software packages have been equipped with this feature. iMovie, for example, displays still images by introducing zooming and panning, and adds transitions between frames. Muvee [muv09] offers an automatic solution for image slideshow production.

The generated image-slide by purely image features is an independent visual structure and isolated from the audio information. To make it more interesting, background music is considered. Actually, in the field of applied media aesthetics, sound is an indispensable element in media production [Zet99]. Music is a unique language that is universally resonant. And it is one of the most direct ways to establish moods. Happy music can enhance the happiness of the visual scene, even if the context of the scene is of a neutral atmosphere. Sad music contributes to the opposite. [Alt02] lists 12 basic functions of music, ranging from content interpretation to mood control. 3 of them being part of our photo slideshow generation procedure:

- *Setting Pace.* Music influences the pace of video through tempo and rhythm. Different tempo exhibits different moods to the audiences. For example, slow motion implies dullness or dignity. On the contrary, fast

tempo shows agility, happiness or other strong emotions.

- *Unifying Transition.* It provides audio transitions between scenes. For example, overlapping clips of music help the transition of one scene to the next continuously. Fade in/fade out effects introduce definite breaks.
- *Evoking emotion.* Music is believed to be the most effective way of creating atmosphere, delivering feelings, and evoking mood. In the field of art, music can be an analogue to almost all conditions and emotions.

Setting pace considers music/visual alignment, the second is related to shot composition and the third discusses emotional mapping between music and visual scenes. The connections between images and audio is direct and intuitive, as stated by synaesthesia theories[Alt02]. "Hearing Colors and Seeing Sounds" is a core idea in music visualization [McD07]. [JI] discusses the characteristics of the art of Visual Music, which is defined by three aspects: 1) It deals with composition of music and moving images; 2) Equal importance is given to both images and sounds; 3) It builds up close music-image relationship based on the sense of synaesthesia. Instead of treating background music as non-essential, or a supplementary part of videos, artists of visual music believe that audio is of equal importance as of visual objects. The temporal architecture of images and audio is typically non-narrative and non-representative[Eva05]. The relationship is so important that some visual music artists believe a muted video is better than one whose background music is not related to the visual objects[McD07]. Even though in the traditional scope of visual music, both music and visual objects benefit from abstract forms, we are inspired by the power of "visual ear", and try to expand the dimension of photo slideshow by a higher level of audio-visual synchronization.

There are widely accepted colorful emotion descriptions: blue means heart-breaks, red means anger, and green means envy. Colors influence human perception and emotions [Zet99], just like music does. We use musical terminology to describe colors and use colors to name music scales. It seems natural for human beings to combine color and music [DeW87].

The underlying relationship between colors and music offers a new way for image selection and displaying arrangement. Instead of using image features as the only independent landmark in image selection, we want to assign equal importance to audio features for a better synchronization in the output video. Given a music clip, its emotion places constraints on image features that could offer good matches.

In the current stage of work, we consider colors, which is an excellent tool to introduce and enhance event moods [Zet99]. Unfortunately, DeWitt [DeW87] pointed out the fact that there are no universal rules to correlate music and colors. Attempts have been made to match visual colors and sound in an aesthetic system, but most of them have failed. The emotional impact of color patches are contextual, so is music. In order to avoid this semantic gap between feature interpretations and global impacts, Zettl [Zet99] suggests a more reasonable way: to match color and sound by their *aesthetic energy*. Aesthetic energy is the relative aesthetic impact on human beings [Zet99].

In this chapter, we present our image slideshow scheme based on the aesthetic energy correlation between images and music. It could serve as a new image selection criterion. Rather than independent image feature analysis, we equalize the importance of music and images and match the aesthetic impacts of background music and selected images. The characteristics of given music decide the global emotion of the selected image subsets.

4.2 Previous Work

The challenge of automatic photo slideshow generation has been addressed previously. There are two underlying problems: image selection and image composition. Feature analysis and content interpretation are two common approaches.

Common low-level features include color [LS07], texture [LS07], and contrast [HLZ04b]. Feature selection also considers quality assessment [HLZ04b] and image similarity [HLZ03b] [HLZ04b]. Images with serious degradations are believed to be unsuitable for the final image album. Image similarity is used to cluster images and remove duplication, while content interpretation uses higher levels of information (human faces [HLZ03c] [SPJ09], indoor/outdoor scene). To further understand the images dataset, annotation is used for semantic analysis [LS07]. [RSB10] mines the social network of users to generate social stories.

Image composition considers sequencing and layout of selected images in the final output slideshow. One intuitive objective is to generate a story like [HLZ03c] and [HLZ04b], which makes use of timestamps and face annotation to produce a rough storyline. Each selected image is assigned with a certain camera motion pattern. [RSB10] looks into the user's network. Photo selection is based on event detection. The individual social importance ranks tagged photos. [HLZ03b] makes use of time, scene and similarity(GoS) to group images and only one image is selected from a GoS group. Rather than story generation, [LS07] extracts images based on theme and highlighting faces. They use time, color and faces to measure the theme similarity between images and cluster them according to theme distance. The emotion-

based slideshow generation is proposed in [LS07]. They consider impressionistic images which seldom contain semantic relations between each other. They examine color and gray-scale information and associate them with painting emotions by affinity graph. Images are clustered based on emotion and a corresponding music clip is recommended according to emotion association.

In these slideshow frameworks, image properties are attached with higher priority in selection and composition. Music is treated as an accompaniment, and mostly only beat synchronization is considered. However, the central idea of music visualization is how a color art can be analogous to the art of music [Cui02]. McDonnell [McD07] gives a thorough introduction on visual music. Dionysios and his group carry out series of experiments in determining and visualizing musical chromatic index [PM03]. They consider music chroma χ based on scales. 12 colors are assigned to different χ values, where white is of the lowest chroma and black is of the highest emotional impact. [PMtH07] builds up the connection between hue and timbre. Hue in the HSV color space is quantized into 8 categories and 8 corresponding timbral sounds are mapped to the color space. [Abb88] builds up the maps between timbre/shape and loudness/brightness. The Stanford group tries to translate pixel color and brightness of an image to the pitch and loudness of a small component of a given sound [LBY04]. The audio-visual relationship could also be found in the mappings from color to timbre [Abb88], tone harmony [CKS10], tone/color [PKW08], timbre/shape [BKMY08], and frequency [CNSF10].

The mappings built in visual music studies are used to compose music from images [PMtH07][PKW08] and generate animation to match music [Abb88][CKS10][BKMY08]. We want to synchronize image analysis and audio analysis by audio-visual mappings. Instead of image subset selection based on purely image features,

we want to obtain a better compatibility between the selected home photo subset and the given background music. In our work, we try to build up the correlation between natural images and music. Equal importance is assigned to images and sounds[J1]. Instead of using abstract objects to interpret the given music clip, which is the concern of traditional visual music, we use a sequence of selected natural images. Audio-visual matching is built based on aesthetic energy correlation between colors and music. To generate an impressive photo slideshow from the chosen images, we employ the Ken Burns effect [Eff]. Artificial camera work enhances synchronization between images and music clips. Our major contributions include the audio-visual mapping model based on aesthetic energy and camera work generation based on music and images. Zoom in/out pattern is designed for videos suitable for common electronic devices. Music-guided camera motion is produced so as to make the image display more interesting.

4.3 Color and Sound Matching

Colors have a special influence on our perceptions and emotions, which is quite similar to what music does on human beings. They are powerful tools for establishing mood of an event. Thus it seems natural to associate colors with music. But the correlation is so subjective that there is hardly any scientific evidence to give a reliable direct match. Zettl proposed another possibility. Instead of matching colors and sound by individual note and hue, a more sensible approach could be the relative aesthetic energy [Zet99].

Attribute	Property	Energy
Hue	Warm	High
	Cold	Low
Saturation	High	High
	Low	Low
Brightness	High	High
	High	Low
Contrast	High	High
	High	Low

Figure 4.1: Aesthetic Energy of Colors

4.3.1 Aesthetic Energy of Images

The aesthetic energy is the relative aesthetic impacts of colors on human beings. Generally speaking, hue, saturation, brightness and contrast are important factors that influence the color energy. Contributions of these attributes to aesthetic energy are given in Table 4.1. Saturation is the most important attribute of color energy of the four. A simple example is given by the fact that cold colors with high saturation is more emotionally powerful than warm colors with low saturation. Based on the criteria in Table 4.1, we propose our model for a quantitative evaluation of aesthetic energy for a given image. Among the 4 color features, hue and saturation determine the chroma properties, while brightness and contrast are the gray scale information. In our scheme, the former two are computed in HSV space, while brightness and contrast are estimated in the L*a*b space, because the model coincides with eyes' reaction to different colors.

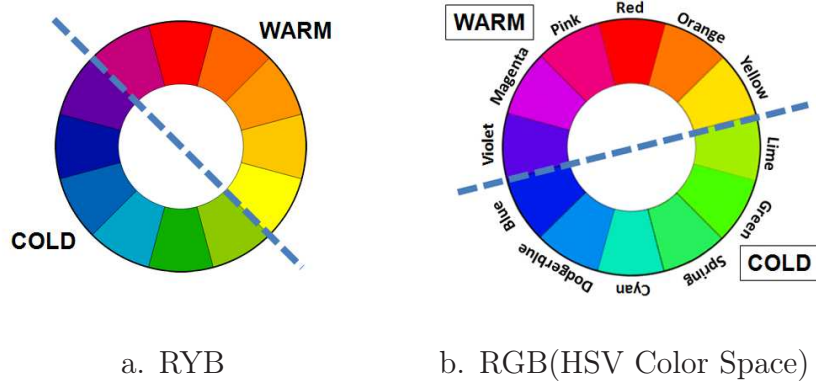


Figure 4.2: . (a) The color wheel under Red-Yellow-Blue(RYB) model. (b) The color wheel under RGB model (in the HSV color space).

4.3.1.1 Hue (H)

The warmth of color is directly related to mood. According to color theories, warm colors are energetic and vivid, while cold colors make people calm and the impression is comforting. Meanwhile, achromatic colors (white, black and gray) are considered to be neutral. Color wheel is a useful tool to segment warm and cold colors [Boy01]. Figure 4.2 shows two color wheel which could be used to segment warm and cold colors. The RYB color wheel is widely used in artistic domain because they are based on primary pigments. While in the computer science domain, the RGB color wheel is more common. Figure 4.2.b shows the RGB color wheel in the HSV color space. Segmentation is performed on both color wheels (dashed lines). The most significant difference between the two segmentation schemes is the placement of violet. In RYB model, violet is categorized as a cold color, while in our segmentation under RGB model, it is warm. We know that violet is the combination of red (warm) and blue (cold), thus its color warmth ought to lie between them. We try to reduce the influence of the displacement, which will be discussed later.

Associating the emotional impact and hue's influence on color aesthetic

energy, we let the energy coefficient \mathbb{H} range from -1 to 1. For cold colors, the corresponding \mathbb{H} is below 0, and \mathbb{H} for warm colors is positive. The zero point is for achromatic colors, whose emotion is neutral, so is their contribution to the aesthetic energy. Table 4.3 shows our hue energy evaluation scheme. We choose the HSV color space over the RGB model. In the HSV space, the first layer contains hue information. We categorize all the colors into 12 types by quantization. Colors whose value falls into $[0, 1/24] \cup [23/24, 1]$ are categorized as red. Similarly for all the other colors, quantization is performed to segment them to a corresponding class and aesthetic energy coefficients are assigned to them.

When computing the energy coefficient for each color, we carried out a small survey. The Participants agreed on the assumption that red is of the highest energy while cyan is the lowest. Thus in our color warmth model, energy coefficient for red is 1, the highest, while that of cyan is -1, the lowest. Observing that the two colors are exactly on the diagonal of the color wheel, we assign coefficients to other colors based on the following criteria: 1) \mathbb{H} for warm colors is positive, \mathbb{H} for cold colors is negative, and achroma colors have zero- \mathbb{H} ; 2) \mathbb{H} of color pairs on the diagonal are opposite numbers. 3) Colors that lie closer to the segmentation boundary are of smaller $|\mathbb{H}|$.

Based on the above criteria, we can reduce the influence of incorrect segmentation of violet. Violet lies on the boundaries in RYB and RGB color space. According to Rule.3, the absolute value of its energy coefficient \mathbb{H} shall be very close to 0 (in our scheme, it is set to be $1/4$, whose absolute value is the lowest.) So when we compute the aesthetic energy, the displacement of violet will not cause significant value shifts on the outcomes.

Color	H	Color	H
Red	1	Cyan	-1
Orange	3/4	Dodgerblue	-3/4
Yellow	1/2	Blue	-1/2
Pink	3/4	Spring	-3/4
Magenta	1/2	Green	-1/2
Violet	1/4	Lime	-1/4

Figure 4.3: The assigned energy coefficients for different colors.

Color	Value	Color	Value
Red	$[0, 1/24) \cup [23/24, 1]$	Cyan	$[11/24, 13/24)$
Orange	$[1/24, 1/8)$	Dodgerblue	$[13/24, 5/8)$
Yellow	$[1/8, 5/25)$	Blue	$[5/8, 17/24)$
Pink	$[7/8, 23/24)$	Spring	$[3/8, 11/24)$
Magenta	$[19/24, 7/8)$	Green	$[7/24, 3/8)$
Violet	$[17/24, 19/24)$	Lime	$[5/24, 7/24)$

Figure 4.4: Color quantization for categorization.

4.3.1.2 Saturation (H)

Saturation is the most important attribute when we evaluate color aesthetic energy[Zet99]. In HSV color space, the second layer, ranging from 0 to 1, contains saturation information. We use mean saturation over the entire image to evaluate the saturation energy coefficient $\$$

$$\$ = \frac{1}{MN} \sum_{m,n} S(m, n) \quad (4.1)$$

To further improve the reliability of saturation energy coefficients, we compute the mean saturation of a certain subset of the image instead of using the global average directly. According to the dark prior channel assumption [HST09], any macro-block of a haze-free natural image contains at least a pixel, one color layer of which is close to zero in the RGB space. The saturation of such pixels is close to 1 in the HSV color space. It is known that colors of higher saturation have higher aesthetic energy [Zet99]. We come to the assumption that for a natural image of high quality, most of its pixel saturation will be close to 1. Moreover, contrast between foreground and background is another attribute in color aesthetic energy evaluation. Image editors might intentionally reduce the background saturation and exaggerate the impact of fully saturated foreground. In order to rule out the influence of less-important low-saturated pixels, we only consider those chromatic pixels, i.e

$$\$ = \frac{1}{|\Omega|} \sum_{(m,n) \in \Omega} S(m, n), \quad \Omega = \{(m, n) | S(m, n) > 0\}. \quad (4.2)$$

The special case comes when the image is a single-layer one. Such images are gray-scale and contain no chroma information at all. According to previous



a. Original Image b. HSI intensity I c. HSV value V d. CIELAB L*

Figure 4.5: . The gray scale images in different color spaces.

discussion, the achroma colors are regarded as neutral, and their hue and saturation energy coefficients are all set to 0.

4.3.1.3 Brightness (B) & Contrast (C)

Brightness and contrast are gray-scale attributes, so we convert the color image to a grayscale one. In practice, instead of using the value layer of HSV space, we choose L* of CIELAB color space, because L*a*b* color is designed to approximate human vision based on perceptual uniformity. The L* component of CIELAB closely matches human perception of lightness. Figure 4.5 shows a comparison between the two color spaces. CIELAB is more competent in maintaining luminance perception of human eyes.

Let $L(m, n)$ denote the L* value at pixel (m, n) , the brightness energy coefficient B is defined as the Square Mean Root over the whole image, and the contrast brightness coefficient C is the standard deviation

$$B = \sqrt{\frac{1}{MN} \sum_{m,n} L(m, n)^2} \quad (4.3)$$

$$C = \sqrt{\frac{1}{MN} \sum_{m,n} (L(m, n) - \bar{L})^2} \quad (4.4)$$

The above 4 attributes are directly related to the color aesthetic energy. Finally, we introduce another aesthetic energy attribute, color energy E, for



		
Hue H	-0.49	0.81
Saturation S	0.60	0.76
Brightness B	0.44	0.45
Contrast C	0.31	0.19
Energy E	0.50	0.64

Figure 4.6: Color aesthetic energy for two test images.

the audio-image mapping in the later discussion. It is the weighing combination of the other 4 attributes.

$$E = \omega_1 |H| + \omega_2 S + \omega_3 B + \omega_4 C \quad (4.5)$$

Saturation is the most important attribute, so the highest weighing value is assigned to it. Hue is directly related to mood perception, so the second largest weighing value is assigned to hue. Thus in our implementation, the corresponding weighing factors are set to be 0.3, 0.4, 0.15, and 0.15. Table 4.6 shows the corresponding coefficient values for the two test images. The dominant color of the chess image is cold, and that of the sunset is warm. The hue coefficients coincide with human subjective perception. The saturation of chromatic areas are relatively high for both images, so that S are above average.

4.3.2 Aesthetic Energy of Audio

Sound are made up of basic audio elements related to computational aesthetic analysis as listed in [MKYH03]. Each element contains certain characteristics that influence our perception of sound. Table 4.7 shows several selected basic

elements. In our audio analysis process, we use the MIRtoolbox [LTE08] to extract corresponding audio features.

4.3.2.1 Pitch \mathcal{P}

Pitch is the highness or lowness of a sound, which is measured by frequency. The generally accepted pitch standard is called A prime, which is 440 Hz [Zet99]. The lowest sound frequency that could be detected by human beings is 20Hz. Studies shows that the upper frequency limit of tones that can be judged to be musical is 4000-5000Hz [TH93]. Thus in our scheme, we set the boundary of music pitch to be 30 - 4000Hz. According to Table 4.7, music pieces of higher pitch are believed to be more exciting. Denoting the detected music pitch as P , the aesthetic energy \mathcal{P} is defined as

$$\mathcal{P} = \begin{cases} \frac{1}{2} \cdot \frac{\log(P) - \log(440)}{\log(440) - \log(30)}, & P \in [30, 440] \\ \frac{1}{2} \cdot \frac{\log(P) - \log(440)}{\log(4000) - \log(440)}, & P \in (440, 4000] \end{cases} \quad (4.6)$$

When the detected pitch is 440Hz, the corresponding \mathcal{P} is 0.5, which represents the neutral energy.

4.3.2.2 Dynamics \mathcal{D}

Dynamics, also called loudness, describes how loud or soft the music piece is. It is often evaluated by the magnitude of the temporal wave. Direct magnitude comparison is not always reasonable. Louder music piece may show higher energy impacts on the audiences. But in our audio-visual mapping scheme, it is not so reasonable because the inherent properties of music remain the same by volume adjustment. Thus in our approach, we use *low-energy* to depict dynamics \mathcal{D} . *Low-energy* shows the energy distribution over time, and

Attribute	Property	Effect
Pitch	High	Bright
	Low	Peaceful
Dynamics	High	Strength
	Low	Weakness
Timbre	Brassy	Harsh
	Reedy	Lonely
Tempo	High	Excitement
	Low	Dignity
Attack	Major	Fast
	Minor	Slow

Figure 4.7: Sound elements and their effects on perception.

is defined to be a percentage of frames showing less-than-average energy. If most frames of a given audio clip contain high energy, it results in a lower low-energy rate LE , and the dynamics energy is high for such music clips. Otherwise, higher low-energy rate means low dynamics energy. Thus we have $\mathcal{D} = 1 - LE$

4.3.2.3 Tempo \mathcal{T}_1

Tempo is the speed of an audio clip, which is evaluated by beats per minutes (BPM). A certain value is assigned to a music piece, which specifies the number of beats that need to be played in a minute. Faster tempo shows excitement, while slower tempo may suggest control [Alt02]. Tempo marks are used to describe the tempo of music. Common marks define music tempo ranging from 40bpm to 200bpm [Blo]. Most beat detectors also assume that the tempo is roughly between 70bpm - 160 bpm [ZW07]. Thus in our model, tempo is set to fall within the range from 60bpm to 180bpm, with 120bpm the neutral-energy

state. The aesthetic energy \mathcal{T} of the neutral music pieces is set to be 0.5. For those music clips whose tempo is above 120bpm, their \mathcal{T} will be higher than 0.5, otherwise, \mathcal{T} shall be lower.

4.3.2.4 Attack \mathcal{A}

Attack is the speed of a sound that reaches a certain level of loudness. In addition to attack, decay defines the time range starting from the tone point to the time when we cannot perceive the sound any more. We use fast/slow to describe attack or decay. The sound reaches its peak very quickly under a fast attack. Such fast attack suggests excitement and sharpness. Otherwise, it is slow for the sound to go up to its maximal point. Such change is soft, and the aesthetic feeling is gentle. Denote the detected attack time of an onset by A , the attack energy is given by

$$\mathcal{A} = -\log_{10} A \quad (4.7)$$

and \mathcal{A} is normalized to the interval $[0,1]$.

4.3.2.5 Timbre \mathcal{T}_2

ADSR stands for attack, decay, sustain, and release. It is commonly used in modeling the timbre of sound. Each instrument has its own envelope that could be described by the 4 phases. Unfortunately, even though the ADSR model can partially profile different families of musical instruments, it is still not an easy task to distinguish the sound timbre. There are some widely applicable rules for distinguishing timbre [SZL09]: 1)Reeds have short attack-decay phase; 2)String instruments have longer attack-decay phase, and the

Image Attribute	Property	Music Attribute	Property
Brightness	High	Dynamics	Loud
	Low		Soft
Saturation	High	Timbre	Brass,String
	Low		Flutes,Reeds
Hue	Warm	Pitch	High
	Cold		Low
Energy	High	Tempo	Fast
	Low		Slow
Contrast	High	Attack	Fast
	Low		Soft

Figure 4.8: Structural transition from images to music for audio matching.

sustain phase is extremely long as well. We are not interested in the exact timbre classification but the aesthetic energy. Thus according to Table 4.8, \mathcal{T}_2 is simply given by a linear combination of the 4 attributes in the ADSR model.

4.3.3 Color-Sound Matching

Suppose we have an image dataset Ω . Consider a given piece of audio clip M , how can we select an image subset to match the music clip and create an audio-visual slideshow? A brief structure mapping between images and music is listed in Table 4.8. Zetl [Zet99] points out that in order to synthesize a meaningful audio-visual structure, it is not acceptable to process the images and music independently. On the contrary, the two must be combined in a way such that we can "visualize the sound" and "hear the image event". However, there is no maximally effective answer to the meaningful matching between

visual objects and music. We used a linear model and assigned empirically coefficients to different features in [YK12]. The limitation of the linear model comes that direct audio-visual feature mapping relies on accurate feature extraction techniques. Therefore the quality of mapping results rely heavily on the correctness of feature preprocessing. Another drawback results from the empirical coefficients.

Therefore, in our current work, we use a learning-based audio-visual mapping. We have built up the aesthetic energy description model for images and audio in the above sections. [MKYH03] built a correlation between video features and audio. But they admit that even though we have the correlation map, not all features are extractable. At the current stage of our work, we follow their feature connection framework, and build our solution based on aesthetic energy. By matching the mood between images and music. We extract color information from images, because color is directly related to mood. Meanwhile, music itself is one of the most effective tools to deliver emotion. The audio clip is segmented into subclips based on detected onsets. In our current experiment, we use professional musical video as the training dataset. The music video clips include: professional image slideshow, scenery music video, and historical education video. Specifically speaking, we do not use commercial music video because these videos have too much emphasis on the content, and not on the audio-visual features. The proposed slideshow algorithms, on the contrary, use low-level information to model the correlation between audio-visual features.

The aesthetic energy value ranges from 0 to 1 (for hue and pitch, the interval is $[-1,1]$ with 0 the neutral point.) Figure 4.9 shows our audio-image mapping scheme. In our experiment, each of the training video clips lasts

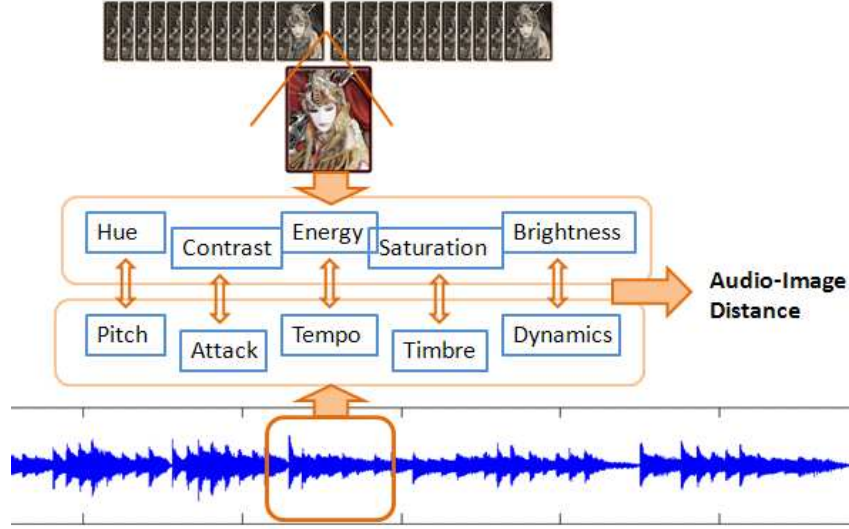


Figure 4.9: A brief description of our audio-visual mapping scheme.

about 10 seconds. The clips are segmented based their on-sets. For each subclip x , the audio-visual distance ($D_{av}(x, y)$) is defined to be

$$D_{av}(x, y) = \sum_i \omega_i \chi_i(x, y) \quad (4.8)$$

where $\{\chi_i(x, y)\}$ is the set of aesthetic energy distance between audio-visual pairs: $|\mathcal{B}(y) - \mathcal{D}(x)|$, $|\mathcal{S}(y) - \mathcal{T}_1(x)|$, $|\mathcal{H}(y) - \mathcal{P}(x)|$, $|\mathcal{E}(y) - \mathcal{T}_2(x)|$ and $|\mathcal{C}(y) - \mathcal{A}(x)|$.

Our task in the audio-visual feature training process is to decide on the audio-visual mapping parameters ω_i . Each video clip is segmented based on their onsets, and for each subclip, the corresponding audio-visual feature distance are denoted as a 5-dimensional vector $\{\chi_i\}$, and each dimension represents the distance between the corresponding audio-visual feature pairs, as shown in Figure 4.8. The weighing coefficients ω_i are decided by a Bayesian classifier. The optimal audio-image match is defined to be the pair with lowest

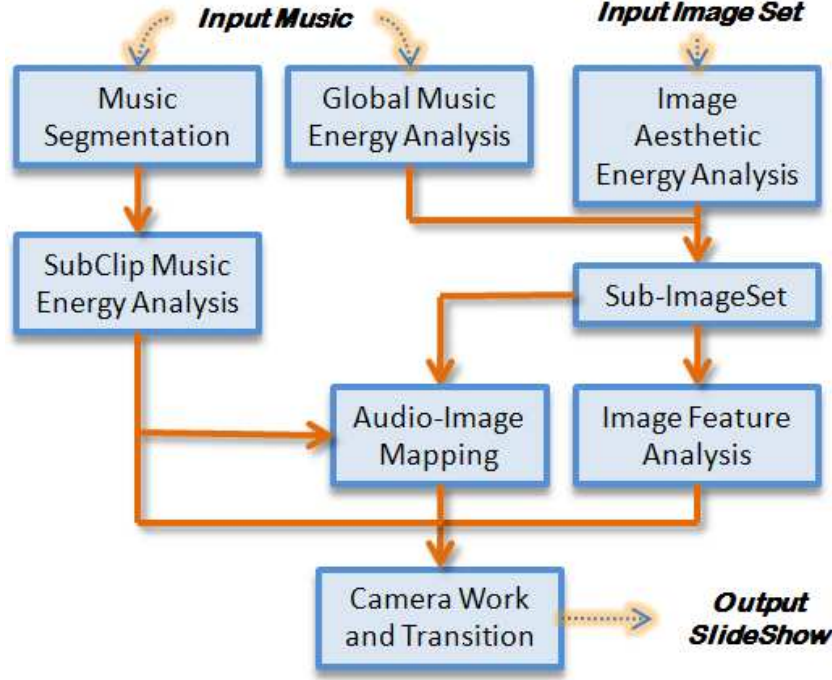


Figure 4.10: . The flowchart of our proposed music-photo SlideShow scheme.

D_{av} value

$$\Phi(x) = \operatorname{argmin}_{y \in \Omega} D_{av}(x, y) \quad (4.9)$$

4.4 Our Photo SlideShow

The overall framework of our proposed image slideshow is given in Figure 4.10. Related features and audio-visual mappings have been discussed in the previous section.

4.4.1 Image Pre-Selection

The personal image collection may be very large for common users. To improve the efficiency and quality of the output slideshow, a pre-selection is performed. The pre-selection process mainly tackles two issues: 1) choose the

appropriate theme of the image subset to better match the given music piece;
 2) reduce the size of the image candidates to improve efficiency.

To begin with, we perform an off-line process of the personal image dataset. The images are first classified into 3 categories based on their valence features. In our current work, we use the model proposed in [MH10]. In their model, they classify the images into 8 kinds of emotions. We select only 3 of them: happy, excitement and neutral. Theoretically speaking, valence includes both negative and positive emotions. But in practice, users want an image slideshow for sharing, and the contents are tend to be positive. Therefore, we choose only the 2 positive categories and the rest are left as neutral. Based on the dominant mood of the given music clip, a corresponding image subset is selected.

Then a further step of subset pre-selection is based on feature connections between image features and the global audio features. We extract dynamics, timbre, pitch, tempo, and attack information from the global audio clip. Based on the feature matching criteria discussed in Section 4.3, we get the first N images of the lowest D_{av} as pre-selected images subsets.

4.4.2 Audio-Image Mapping

At this stage, the audio clip is segmented into subclips. A unique image from the pre-selected subset is assigned to each subclip. We use onsets to segment music, which is the same as in [HLZ04b]. The onset series are extracted first. Considering that the interval between neighboring onsets might be too short for image display, we set the minimal duration of each subclip to be proportional to the global music tempo. The detected onsets in the series are sorted in a list according to their relative strength. Then from the first onset

onwards, the interval whose length is above the minimal duration is set as a subclip. The onsets that are included in this subclip are removed from the onset list. The same process is performed on all the onset series until there are no onsets left in the list.

Once we have the audio clip segmented, these subclips are analyzed and the aesthetic energy features are extracted. We want to build up the mapping between images and the audio subclips. Let K denote the total number of subclips, and the image subset is denoted as Λ . Let $\varphi(\cdot)$ be a mapping scheme, in which $\varphi(i) = j$ means that the j -th image is mapped to the i -th audio subclip. In Section 3, we have put forward the mapping criterion that an optimal mapping between image and audio should minimize the audio-image distance function D_{av} . Furthermore, we hope that the hue of neighboring images should be close for visual pleasure. Abrupt changes from cold images to warm ones disturb the viewing continuity. Thus the optimal mapping is given by

$$\min_{\varphi \subset \Lambda} \sum_{k=1}^K D_{av}(\varphi(k), k) + \sum_{k=1}^{K-1} (\mathbb{H}(\varphi(k+1)) - \mathbb{H}(\varphi(k)))^2 \quad (4.10)$$

Ideally speaking, the solution to Equation 4.10 gives us a unique image selection and audio-visual matching scheme. However, considering the fact that the correlation between images and audio clips is rather subjective, we try to introduce a certain level of randomness in the mapping process. We do not try to solve Equation 4.10 directly. Instead, one audio subclip k_0 is randomly selected among the K subclips. An optimal matched image $\bar{\varphi}(k_0)$ could be easily selected by finding the one with minimal $D_{av}(\bar{\varphi}(k_0), k_0)$. Then

image $\bar{\varphi}(k_0)$ is removed from the image list $\tilde{\Lambda}$. Starting from subclip k_0 , we move leftward and rightward independently, and find the image that minimizes

$$\min_{\bar{\varphi}(k) \in \tilde{\Lambda}} D_{av}(\bar{\varphi}(k), k) + (\mathbb{H}(\bar{\varphi}(k+1)) - \mathbb{H}(\bar{\varphi}(k)))^2 \quad (4.11)$$

Take the leftward part as an example. The value of $\bar{\varphi}(k+1)$ has been decided in the previous step. So to solve Equation 4.11, we just need to select the optimal image $\bar{\varphi}(k)$ from $\tilde{\Lambda}$. $\tilde{\Lambda}$ is a subset of the original images, so the computational complexity is not high even for an exhaustive search. Equation 4.11 considers the minimization between neighboring subclips, and the solution is a sub-optimal one, when compared to Equation 4.10. The randomness lies in the fact that different initial seed k_0 could bring forward different match sequence, because we have removed selected images from the image list. In practice, the level of randomness is controllable by the number of initial seeds. The more the number of initial seeds, the higher will be the randomness of the audio-visual matching outcome.

4.4.3 Image Saliency

Let the selected images that matches audio subclips be the set $\bar{\Lambda}$. These images are further analyzed for their saliency map S_{RoI} . The saliency map is two dimensional and is used to represent the visual interest of a given image. In our current scheme, we use the saliency detection scheme in [HZ07]. Moreover, human faces are regarded as the most important features in the images. We use the efficient face detection scheme in [RSA03]. The detected face regions are assigned with high saliency value.

Figure 4.11: Music Structure and Camera Motion.

Motion	Property	Music Attribute	Property
Motion Vectors	High	Tempo	High
	Low		Low
Zooms	Fast	Attack	Fast
	Slow		Slow
Vector Continuity	Good	Rhythmic	Even
	Bad	Continuity	Uneven
Transition	Abrupt	Key Changes	Extreme
	Gradual		Conservative

4.4.4 Camera Work

To convert image sequence into videos, we introduce the third dimension, time, to the image sequence. The corresponding important feature related to time is motion. As a visual attribute, motion is related to music as well. Zettl [Zet99] gives a brief correlation between motion attributes and aesthetic audio clip features (Table 4.11). Our camera work model is based on the correlation between music and motion.

4.4.4.1 Zoom Factors

Let t denote the duration of an audio subclip, and ω is the zoom factor. In our implementation, we use the standard 640×480 as the output video resolution. In most cases, the image resolution is not of the 4 : 3 aspect ratio. So the image is firstly rescaled. For an image of the size $w \times h$, the scaling factor ω_c is defined as

$$\omega_c = \begin{cases} \frac{640}{w}, & \frac{w}{h} \leq \frac{4}{3} \\ \frac{480}{h}, & \frac{w}{h} > \frac{4}{3} \end{cases} \quad (4.12)$$

According to Table 4.11, the attack time is positively correlated with zoom speed. In other words, the aesthetic energy of attack \mathcal{A} is negatively correlated with zoom speed $\Delta\omega$

$$\Delta\omega = \frac{\lambda}{\mathcal{A}} \quad (4.13)$$

Meanwhile, the average changing speed $\Delta\omega$ is the change speed of ω from the initial value ω_0 to 1 within time t

$$\Delta\omega = \frac{1 - \omega_0}{t} \quad (4.14)$$

According to Equation 4.12, 4.13 and 4.15, the initial zoom factor ω_0 is given by

$$\omega_0 = 1 - \frac{\lambda t}{\mathcal{A}} \quad (4.15)$$

Consequently, when we have the coefficient λ determined, the initial zoom factor ω_0 is also decided. In the implementation, since the attack aesthetic energy value ranges from 0 to 1 with 0.5 as the neutral state, we set the zoom factor for a neutral subclip to be 0.75. To decide the position of initial window, we use the same scheme in [XK10b]. The window that catches the highest saliency is thought to be the optimal initial window. The corresponding saliency is denoted as $S_{RoI}(k)$, where k is the subclip index. In the next step, we need to decide the displaying pattern scheme. [XK10b] discussed two different displaying approaches: the inductive approach and the deductive ap-

proach. The deductive approach shows certain targets from the general to the locally specific and the inductive approach exhibits objects from local detail and offer the audience a general overview in the end. The home produced video clips are of low resolution, thus the inductive approach is more appropriate in our current situation. However, it is boring if all of the images are displayed in an inductive approach, i.e. give a close shot and then zoom out. Instead, we want some changes between neighboring subclips. Let $\chi(\cdot) \in \{0, 1\}$ be a zoom pattern, where $\chi(k) = 1$ means an inductive approach and $\chi(k) = 0$ means a deductive approach. χ is given as an optimization problem

$$\operatorname{argmax}_{\chi \in \{0,1\}} \sum_{k=1}^K S_{RoI}(k) \chi(k) + \alpha \sum_{k=1}^{K-1} (\chi(k) - \chi(k+1))^2 \quad (4.16)$$

The optimal zoom pattern is given by the one that maximize the above equation. Since the inductive approach is the more desirable displaying scheme, $S_{RoI}(k)$ is regarded as the weighing factor for subclip k . If the RoI is of higher saliency, this part is more important and need to be emphasized by a more powerful displaying scheme, i.e. the inductive approach, so that $\chi(k)$ is more probably to be 1. The latter part of Equation 4.16 places constraints on the neighboring subclip displaying patterns. Different approaches for neighboring frames are more desirable. To get the optimal zoom pattern defined by Equation 4.16, we only need to solve a non-homogeneous Markov system [Ste94].

Let $P(\chi(k) = \rho)$ denote the probability of $\chi(k) = \rho, \rho \in \{0, 1\}$. The conditional probability of choosing the displaying pattern from state ρ_k to

state ρ_{k+1} is given by

$$\begin{aligned} & P(\chi(k+1) = \rho_{k+1} | \chi(1) = \rho_1, \chi(2) = \rho_2, \dots, \chi(k) = \rho_k) \\ &= P(\chi(k+1) = \rho_{k+1} | \chi(k) = \rho_k) \end{aligned}$$

where the nonhomogeneous chain matrix $P_{ij}(k) = \text{Prob}(\chi(k+1) = i | \chi(k) = j)$ is

$$p_{ij}(k) = \begin{pmatrix} 1 & 0 \\ \frac{S_{RoI}(k)}{2S_{RoI}(k)+\alpha} & \frac{S_{RoI}(k)+\alpha}{2S_{RoI}(k)+\alpha} \end{pmatrix} \quad (4.17)$$

In the previous discussion, the inductive approach is more appropriate for subclip display, i.e. $\chi(k) = 1, \forall k \leq K$, so 1 is chosen to be the initial seed for the iteration.

4.4.4.2 Camera Path

When we have determined the initial zoom factors (destination zoom factor for deductive approach) and the initial windows, the next step is to decide the camera path. Actually, the deductive approach is exactly the temporal inverse of inductive method. So in the following discussion, we consider only the inductive approach for simplicity. At the initial window, there are 4 potential directions for camera panning. In order to catch the maximum amount of image information along camera paths, the direction whose distance from the initial window to the image boundary is the furthest is selected. Figure 4.12.1 shows the four directions and the selected one is labeled by a solid line. The change of zoom factor $\Delta\omega$ is given by Equation 4.15. The window moves at the selected direction at speed v_1 , meanwhile the size increases according to

the zoom factor $\omega = \omega_0 + (t - 1)\Delta\omega$, where t is the frame index of the subclip. When the window touches one boundary (Figure 4.12.2), camera changes the direction at a new speed v_2 (Figure 4.12.3). After the window touches two boundaries (Figure 4.12.3), a new zoom speed $\Delta\bar{\omega}$ is introduced to show the complete image in one frame (Figure 4.12.4). In our implementation, v_1, v_2 and $\Delta\bar{\omega}$ are decided by the onsets within the subclip.

The most important part is the selection of initial directions. Ideally speaking, the direction of the largest distance to the boundary is selected. In practice, the slide is boring if too many neighboring subclips share the same path pattern. So we constrain the similarity between initial directions of neighboring subclips. Table 4.11 shows the relationship between continuity of motion vectors and that of rhythm. We use four vectors $d(1) = (1, 0)$, $d(2) = (0, 1)$, $d(3) = (-1, 0)$, $d(4) = (0, -1)$ to denote four directions: right, up, left and down. The audio rhythmic continuity between subclip k and $k + 1$ is denoted by $RC(k)$, and the distance between initial window to four image boundaries of frame k is denoted as $Dist(k, i)$, $1 \leq k \leq K, i = 1, 2, 3, 4$. Similar to displaying approach designing, the optimal path pattern ψ (on the discrete range $\{1, 2, 3, 4\}$) of all the subclips is given by

$$\max_{\psi} \sum_{k=1}^K d(\psi(k)) Dist(k, \psi(k)) + \gamma \sum_{k=1}^{K-1} RC(k) |d(\psi(k)) \dot{d}(\psi(k+1))'| \quad (4.18)$$

The solution scheme of the above equation is exactly the same as that used in solving Equation 4.16.

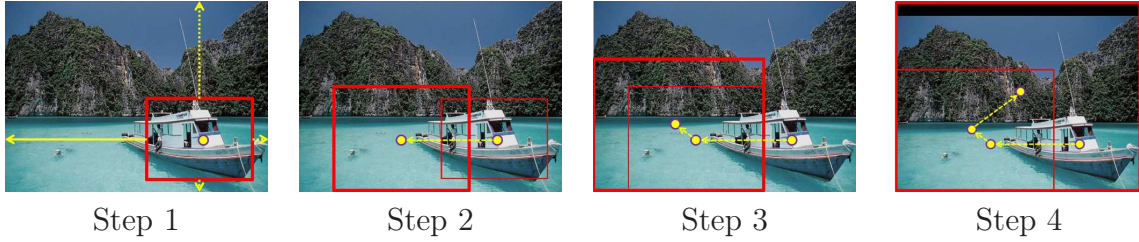


Figure 4.12: An example of the camera path.

4.4.5 Transition

Sequence motion, also known as tertiary motion, is the relationship of vector fields from one shot to the next [Zet99]. Transition devices contribute to the formation of relational rhythm. Cut and dissolve are two common transitions. Actually, cut can be thought to be a fast dissolve whose duration is 0. Table 4.11 associates the duration of transitions to the change of key between neighboring subclips. In our scheme, the duration of transitions is set to be proportional to the jump of keynote between neighboring subclips. Moreover, in our current stage of work, we only consider cut and dissolve. Even though special transitional effects, such as tilting, freeze, and tumble, can be visually exciting, judicious application of these effects such that they help to intensify the impacts to the audience [Zet99] is beyond our current scope of work.

4.5 Experimental Results

Objective evaluation of the proposed scheme is difficult, so a subjective user study is performed. We carried out the experiments on a dataset of 600 image, including 3 categories and 200 image for each. Sample images are shown in Figure 4.13. Because the proposed system does not consider the storyboard generation, the images have no semantic relationship with each other. 15

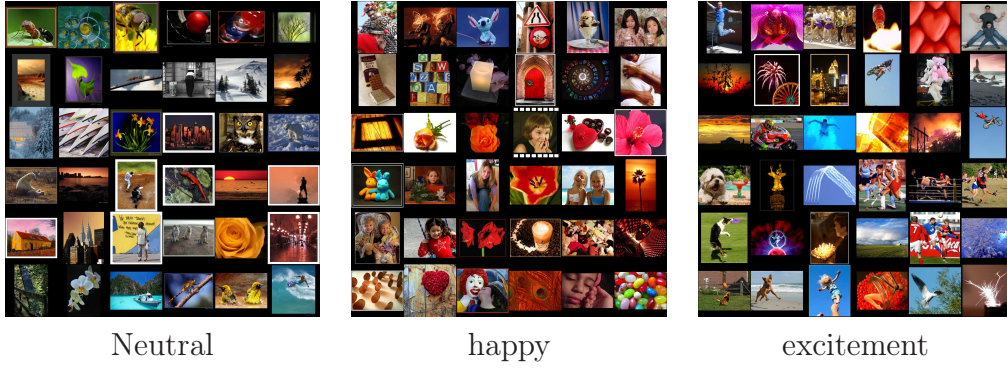


Figure 4.13: . Sample images of the experimental image dataset. Each group contains 200 images and 36 random images of each group are displayed in the figure.

audio clips of different instruments are selected for the experiment, and the experiment itself contains 24 slideshow outputs. We carry out 3 groups of experiments and aim to test the proposed framework from several aspects. For each group, at least 3 audio clips are given and each is associated with one image subset. 17 participants are invited to do the user study. They are asked to score the given videos from several aspects. The score ranges from 1 to 5, with 5 the best and 1 the worst.

4.5.1 Scheme Comparison

The first experimental group contains 3 sets of videos. This part aims to evaluate the results of different slideshow schemes. The 3 pieces of input music are selected to fit the 3 emotion categories (happy, excitement and neutral) of the image dataset. There are 3 video clips in each experimental set. In each set, the first clip is randomly generated. A given number of images are randomly selected from the corresponding dataset. They are displayed along with the audio clip accordingly. The duration is the same for all the images and no audio-visual synchronization is considered. The second clip

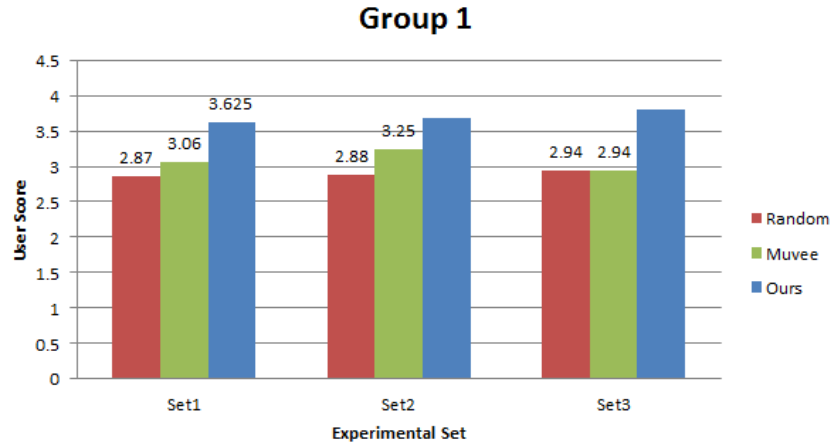


Figure 4.14: . User Evaluation of Group 1.

is generated by Muvee Reveal[muv09] (Seagate Edition). All the images in the corresponding image subset are taken as inputs, and the speed is set to normal (the default settings). Images are rearranged by Muvee. The third is the result of the proposed scheme.

Users are asked to score the overall viewing pleasure of different slideshows (Figure 4.14). Our output is quantitatively better (score 3.6, 3.5 and 3.8) than the other two schemes. In the randomly generated slideshow, no synchronization is considered and there are perceptual shifts between music and images. In addition, the camera motion is dull and image duplication occurs. For the second slideshows, the software-estimated duration of each image is too short for the audiences. In addition to the audio-visual synchronization and displaying issues, proper duration might be another issue that influences the viewing pleasure. In our proposed scheme, the displaying speed is automatically set to be proportional to the music tempo. But in the random and Muvee results, the inappropriate duration significantly hampers the viewing pleasure. To further compare our proposed scheme with Muvee, we manually adjust the pace settings and reduce the speed of image in the Muvee output. According

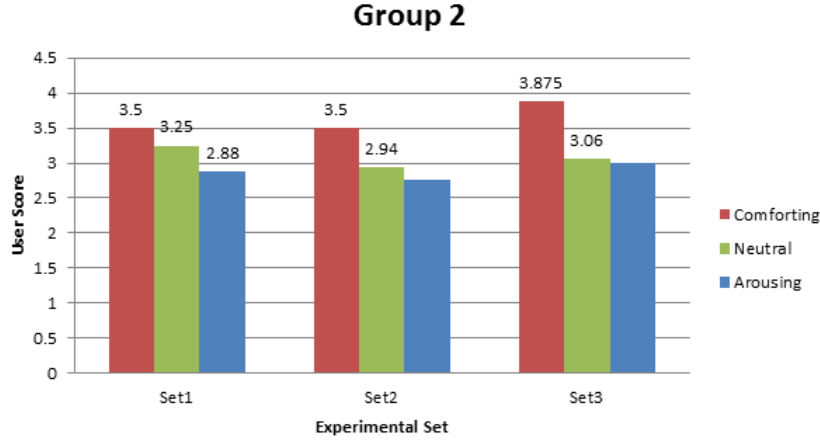


Figure 4.15: . User Evaluation of Group 2.

to the feedbacks from the users, the quality of the two are comparable. Our results offer better synchronization between the audio and video tracks, while Muvee gives better post-processing techniques. Their rendering results are much better than ours.

4.5.2 Comparison between Different Input Audio

Group 2 contains 3 sets of videos. In this part, we want to compare the impressional consistency with different music. The proposed framework has preselected image subset to fit the given audio clip, but for a given audio, the arousal level of different segments may be different. In this group of experiments, we aim to test to what extent the proposed system could be adaptive to the valence of input music. As the experimental design in the first group, each image subset of the 3 types of valence is assigned to the corresponding testing set. 3 audio clips are given in each set. The two audio clips are taken from the same song but different sections. Thus they are similar but with slightly different arousal.

Figure 4.15 shows the user study results. We can find that for the same

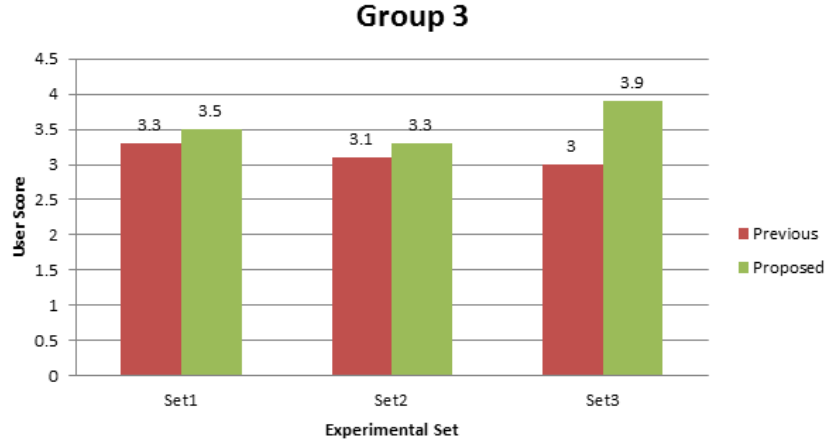


Figure 4.16: . User Evaluation of Group 3.

dominant valence, the arousal of music clip plays an important role in the user experience. In our current system, the comforting music pieces over-performs music pieces with higher arousal value (in the arousal/valence coordinate system). User evaluation shows that slideshows generated from relatively comforting music pieces are better than the others (score 3.5, 3.5, and 3.8), meanwhile the results of the most arousing music get the relatively lowest score (2.8, 2.7, and 3.0)

4.5.3 Comparison with the previous results

In the group of experiments, we compare the performance of the proposed slideshow system and that of the previous version [YK12]. 3 music clips are given and each one of them is assigned with the corresponding image subset. For the results of the former system, the whole image dataset is taken into consideration without the segmentation of image valence. User scores are shown in 4.16.

We can find that the current system over-performs the previous version (User score 3.5/3.3, 3.3/3.1, and 3.9/3.0). Still, the user score of the neutral

image dataset is much higher than the other two (3.9), and that of the image dataset labeled as excitement is relatively lower (3.1). Actually, our current audio feature extraction system is not very effective in dealing with audio clips with high tempo. So that the mapping between visual features and audio properties is not well built. Improvements could be made by better audio analysis algorithms. Moreover, semantic interpretation can contribute to this mis-synchronization.

4.6 Discussions

In this chapter, we present a novel automatic image slideshow system based on the aesthetic energy correlation between images and music. Equal importance is assigned to both image features and audio properties for better synchronization. We are inspired by the idea “hearing colors, seeing sounds” from the art of music visualization. For given image series, a subset is selected based on the features of the input audio clip. The selected images are synchronized with the music subclips by their audio-visual distance. Our scheme can be regarded as a new image selection criterion for slideshow generation. Instead of considering the semantic relationship between images, we minimize the aesthetic energy distance between visual and audio features. The inductive image displaying approach is introduced for family users. Artificial camera motion aims to catch image saliency. Moreover, our algorithm is fully automatic, it is convenient for users to adopt. For automatic slideshow generation, our current work only consider one aspect of the problem: given a selected music piece, how to select an image sequence. The inverse one, how to select a music clip for given image data, is out of our current scope. We may work on it in

the next stage of our work.

At our current stage of work, only chromatic information is extracted from images. Actually, higher levels of image understanding are more desirable. For example, a smiling face is emotionally positive even if the atmosphere is of neutral chroma. For feature extraction, [MKYH03] lists 43 attributes, but in our current scheme, only 10 of them are used. Future research can build up further correlation between other features. Our audio-visual feature matching schemes are essentially given in [Zet99]. Better aesthetic decisions could be made if other mixing heuristics are possible. Special effects, such as digital transition devices and camera work, could be considered so as to enhance the impacts and interestingness of the output slideshow. Proper matching criteria are needed to aid rather than hinder the videos. The semantic understanding of photos can further benefit the slideshow generation. Further work can introduce themes into consideration. Theme-oriented slideshow generation could further explore the semantic relationship between images and audio pieces.

Videos Aesthetics: Automatic Retargeting and Reprojection for Editing Home Videos

Video post-editing is closely related to art. Artistic theories have put forward many criteria for the good piece of video production. Under the guidance of these theories, media aesthetics aims to extract elements and quantize the abstract artistic criteria. In this chapter, we present an automated post-processing method for home produced videos based on frame “interest”. The input single video clip is treated as a long take, and film editing operations for sequence shot are performed. The proposed system automatically adjusts the distribution of interest, both spatially and temporally, in the video clip. We use the idea of video retargeting to introduce fake camera work and manipulate spatial interest, then we perform video re-projection to introduce motion rhythm and modify the temporal distribution of interest.

5.1 Introduction

The development and availability of home entertainment gadgets has boosted the production of home videos. Compared to complex and expensive profes-

sional equipments, common family video recorders are advantageous because of their portability, multi-purpose functionality and economy. However, the quality of most home produced video clips is not always satisfactory. It is affected by user skills and the ambient conditions. Commercial filmmakers use advanced devices and carefully design the scene settings (so called *mise-en-scene*) [Dav10], and enjoy full control over the whole production process. However, for ordinary family users, the available equipment is usually of lower quality, and the ambient environment is uncontrolled as there are no sets. Therefore, post-editing techniques are important to make up for such limitations.

According to film theories [Dav10], proper camera work and shot duration are important factors in story telling.

- *Camera Work*, known as the secondary motion in film theories, is an important element in video creation. Proper camera work has a special role of providing hints to the viewers that make them indispensable [Zet99]. However, one of the most common limitation of home produced videos is the lack of proper camera work. To overcome these deficiencies, we want to bring in synthetic camera work in video post editing.
- *Projection Velocity*. The time in a video is different from the objective time a clock tells, and can be purposely manipulated by the editor by adjusting the projecting speed of the recorded raw video. To the viewers, it shows as fast/slow motion. Neatly controlled projection velocity can offer special aesthetic effects. Sophisticated producers manipulate motion to exaggerate emotional impact and control audiences' perception of events. For example, *The Matrix: Reloaded* (2003) uses a freeze

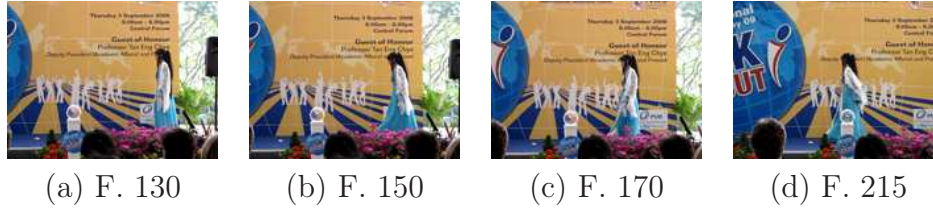


Figure 5.1: The four frames (a)-(d) from a stage performance video clip. This segment lasts more than 4 seconds.

scene with a 360-degree rotation to give the global view and exaggerate a burst of conflict. It signals an increased speed and increases velocity reversely.

An important video editing technique is known as the *sequence shot*, which is a combination of *long takes* and sophisticated camera motion [Dav10]. Video shots have measurable screen duration, and some shots may last extremely long. These shots are called *long takes*, and they are thought to be a “powerful creative resource” [Zet99]. Studies by film critics reveal that long take itself heavily relies on camera motion, such as panning, tilting and zoom in/out. Proper camera motion can guide the viewers to segment a long-take shot into several subshots.

When ordinary users record their everyday life, it is highly probable that they will not stop recording until a certain event comes to an end. Such video clips can last very long. The advantage of such clip is that they record an event from beginning to the end and enjoy temporal continuity. It contains enough information to tell a complete story: who, when, where and what. But without careful *mise-en-scene* and proper recording skills, information redundancy leads to boring subclips and reduces the viewing pleasure of the clip. One example is shown in Figure 5.1. The actor slowly walks from one side of the stage to the other. Even though the camera has tracked the

moving figure, the walking event lasts several seconds without introducing new information to the viewers, which results in a boring subshot.

Guided by the theories of sequence shot, we improve the viewing pleasure of home videos by treating them as a long shot and introducing appropriate camera work. For the long shot, since some subshots contains temporal redundancy, re-projection is introduced to correct the temporal distribution of interesting parts. Thus, the sequence shot technique is adopted in our home video editing as the combination of re-projection and re-targeting.

The proposed automated home video post-editing scheme is based on *frame interest*. While watching a video clip, viewers are more often attracted by foreground objects. Thus in the proposed method, we use the foreground saliency to evaluate frame interest. Works have been done to detect video saliency based on temporal and spatial cues [ZS06]. Motion is the most obvious and direct sign of time and one of the basic video grammar elements which serves to deliver intent and amplify media impact [Zet99]. When motion is not significant, spatial features are assumed to dominate the saliency space. However, common low-level spatial cues, such as color, contrast, and texture, fail in common home video clips quite frequently. Take Figure 5.1 as an example. The colors are quite similar between the dancer costumes and the background poster. Actually, the contrast (of color, sharpness, or brightness) assumption between foreground and background features can hardly hold without proper *mise-en-scene* and camera setups. These, unfortunately, are exactly what home videos lack. Therefore, in our current work, we use temporal motion information as the key cue to estimate foreground saliency. Spatial features are treated as supplementary information when motion fails to offer reliable estimates.

The post-editing process is briefly described as follows. Firstly, the frame foreground is detected and for each frame, we assign a certain value to evaluate its interest based on spatial motion vector correlation and temporal motion continuity. Then we estimate the cropping window and duration of each frame. Finally, we render the new image sequence based on each frame's new duration.

In this paper, we present a novel video editing technique combining retargeting and re-projection, which utilizes the sequence shot editing to enhance the aesthetic impact of home produced videos. User study shows interest improvement under our system. Our main contribution is the adaptive fusion approach based on spatial-temporal saliency. The adaptive fusion of spatial-temporal cues also shows improvement in frame interest estimation.

5.2 Previous Work

Masahito Kumano *et al.* [KAA⁺02] selected appropriate shots and connected them together based on film grammar. They listed four video editing rules, including shot size, camera work, and combination criteria. Instead of simply considering low level information, Brett *et al.* [AV03] built up a storyboard and tried to deliver video intent. They used predefined narrative templates and video genre to automatically generate a storyboard. Media elements were aesthetically combined together to maintain consistence and deliver intents. Madhwacharyula *et al.* [MMK04] tried to cover the gap between low-level signals and high-level metadata. They proposed a content-based semantic video editing mechanism. Schemes mentioned above are mainly based on clips instead of frames, i.e. none of them consider the problem of internal

video clip manipulation. Achanta *et al.* [AYK06] combined video editing and intent modeling. They used some basic media elements, such as color, contrast, brightness and camera motion, to model different video grammars. They did not try to build up any storyboard, but they manipulated media elements of video to map original video clips to four kinds of intent: cheer, serenity, gloom and excitement.

Traditional digital video editing system is an adaptation of the classic analog film editing mechanism, i.e. detect key frames, partition video clips and assemble them. Softwares, such as Windows Movie Maker [Mak], Sony Vegas [Veg], Corel VideoStudio [Vid] and Adobe Premiere [CS4], all follow such routine. Users are able to cut, select, edit and assemble video clips manually to compose desirable output videos. Some semi-automated and automated video editing systems that adopt this mechanism can be found in [HLZ04a] [GBC⁺00] [HLZ03a] [XLY⁺07] and [WH06]. Hua *et al.* put forward their automated home video editing system (AVE) based on optimization in [HLZ04a]. Shots were segmented according to frame difference; important subshots were detected by attention model; narrative and music information was analyzed. These video editing approaches are clip-based editing, without performing frame-level modification.

We are inspired by the idea of retargeting for creating fake camera motion, aimed at improving the video aesthetic interest in post editing process. The original objection of video retargeting is to make videos adapt to devices with smaller screens. The most direct way is to squeeze the video or to crop the boundary to make the video resolution suitable for the display devices. Such schemes result in considerable information loss. A lot of work has been done to reduce important information loss during

the resolution reduction. Current video retargeting algorithms can be categorized into two classes. The first class is seam carving, i.e. use non-uniform sampling schemes to preserve salient pixels and removes pixels with low saliency [RSA09],[GMZ08],[WGCO07] [KKK09]. The second class detect the most important rectangle and crop the regions outside the rectangle [DDN08] [WRL⁺04]. Li et al. [LTY⁺10] treated the scaling and cropping problems as an optimization process. Based on the visual attention shifting theory, they searched for the optimal trajectory along which the local window could catch the most amount of saliency and they solved it as a graph problem. Present video retargeting algorithms have been implemented efficiently to introduce video adaptation of different projection resolutions.

Previous work concerning video duration modification can be found in video summarization and abstraction [SK00], [YKM03], [SXM⁺06]. The most significant difference between traditional video summarization and the proposed video re-projection lies in the fact that they have completely different objectives. As the name implies, video summarization aims to offer an easily interpreted synopsis of given media data and has been used to preview media, highlight videos and detect events in surveillance videos. It tries to deliver the maximum amount of information to the users within a specific time. Truong *et al.* [TV07] has given a comprehensive review on the problem. The output may either be a series of still images [Dir00] [KH00] or a video clip made up from segments [SH04]. In other words, the output does not maintain the temporal integrity. In the proposed video re-projection, the output video clip is assumed to be a visually pleasant one. The "unimportant" segments are accelerated instead of being left out. They occupy some amount of temporal duration and help to maintain the completeness of the story.

Frame content analysis is an important process in automatic video editing system design. A lot of work has been done in this area. Action measure-based key frame detection estimates the amount of visual content variance and is widely used in video summarization. Wolf [Wol96] used the magnitude of optical flow to measure the importance of each frame. It fails when the motion is smooth and optical flow difference is not significant. Ma *et al.* [MLZL02] built up an attention model to detect key frames. The model combines visual saliency and audio saliency. The value is used to measure key-frame importance. Junyong *et al.* [YLSL07] put forward a perceptive analysis based on multiple visual models. Instead of representing content by low-level information such as color and motion, they simulated the human perceptive process based on video visual cues, i.e, a higher-level semantic understanding. Their visual model combines local motion, contrast and some special features, such as human faces, to measure the frame importance and fuse them to reflect audiences' perception changes.

To summarize, automatic video editing frameworks emphasize clip segmentation and re-assembling to make up a story and deliver certain intents, while less attention has been attached to frame-based editing. In our previous work, we considered video retargeting [XK10b] and re-projection [XK10a] as two independent issues. Both of them aim to manipulate the distribution of interest and enhance viewing pleasure, but the former considers spatial manipulation by adding synthetic camera work, while the latter considers temporal editing by changing frame duration. The problem arises from the fact that in practice, the two post-editing techniques are not independent. Editors choose appropriate special effects to achieve higher visual pleasure for different video clips. To further improve the proposed automatic video editing framework

and benefit amateur users, we want to adaptively integrate the two.

In our current work, we aim to produce a video that emphasizes the interesting sections, uses camera work to highlight important regions, and maintains temporal coherency at the same time. Therefore, the architecture of the proposed system contains 3 aspects: aesthetic retargeting, reprojection and the adaptive fusion of the above two. The basic differences between our schemes and corresponding traditional approaches are given as:

- Retargeting. Properly crop videos to make them adaptive to devices with small screens;
- Aesthetic Retargeting. Introduce artificial camera work to make videos aesthetically pleasant;
- Video Summarization. Produce a shorter clip containing important information;
- Reprojection. Adjust the duration of a clip (shorten or lengthen) by changing the duration of each frame.

In the proposed scheme, we fusion the two based on the guidance of sequence shot editing for a better and more sophisticated home video editing framework.

5.3 Our Approach

Consider a single-shot video clip with K frames, whose resolution is $R \times C$. Let (r, c) denote the corresponding macroblock indices. For each frame $k \in \{1, 2, \dots, K\}$, we want to decide its duration $a(k)$ and a cropping window $(r_0(k), c_0(k), \omega(k))$. $(r_0(k), c_0(k))$ represents the center of the cropping win-

dow. $\omega(k)$ is the ratio between the cropping window size and the original window size, i.e. the inverse of the *zooming factor*. In our following discussion, we will take it as the zooming factor for notational simplicity.

5.3.1 Frame Saliency

The saliency of each frame is computed by the temporal information and the spatial features. Temporal information is based on the motion vectors, and for spatial saliency estimation, we adapt the algorithm described in [ZS06]. According to the previous discussion in Section 1, motion vector is treated as the critical interest cue while spatial features are used as the supplementary information.

Motion is modeled by the combination of velocity M_v , motion complexity M_s and motion continuity M_t . Block-based motion vectors, $MV_x(r, c, k)$ and $MV_y(r, c, k)$ ($k \in [1, K], r \in [1, R], c \in [1, C]$) are adopted here, where K is the total number of frames, R and C are the resolution of the frame.

Velocity is defined to be the magnitude of foreground motion vectors. Motion with higher velocity may catch more audiences' attention.

$$M_v(r, c, k) = Mv_x(r, c, k)^2 + Mv_y(r, c, k)^2 \quad (5.1)$$

The local variance of vectors describes the *motion complexity* and *coherence* of each macro-block.

$$M_s(r, c, k) = \text{var}_{\Pi}(Mv_x) + \text{var}_{\Pi}(Mv_y) \quad (5.2)$$

where $\text{var}(Mv_x)$ is the variance of motion vector projection on the x -direction.

Π is a local window at pixel (r, c) , while M_s reveals the motion pattern. When the motion is complicated, it becomes difficult to predict motion by neighboring blocks. In such a case the neighboring blocks may not be coherent and have quite different motion patterns. Thus, a higher M_s value signals more complex local motion.

Temporal motion difference is used to define the *motion continuity and complexity*.

$$M_t(r, c, k) = (\nabla_k Mv_x(r, c, k))^2 + (\nabla_k Mv_y(r, c, k))^2 \quad (5.3)$$

$$\nabla_k Mv_x = Mv_x(r, c, k) - Mv_x(r, c, k - 1) \quad (5.4)$$

$$\nabla_k Mv_y = Mv_y(r, c, k) - Mv_y(r, c, k - 1) \quad (5.5)$$

when M_t is high, motion is not temporally smooth. When it is difficult to predict future motion based on current frame, higher saliency values are assigned to these pixels.

Compensation residue $I_{residue}$ is the normalized square difference between frame k and the predicted image based on frame $k - 1$. The temporal saliency is given by

$$M(r, c, k) = (\omega_1 * M_v + \omega_2 * M_s + \omega_3 * M_t) \times I_{residue} \quad (5.6)$$

where $\omega_1 + \omega_2 + \omega_3 = 1$ are weighing coefficients. The three weighing coefficients are empirically set to 0.5, 0.3 and 0.2.

We adapt the algorithm proposed in [ZS06] to estimate the spatial saliency.

To make it consistent with motion saliency, the frames are divided into macro blocks. Spatial saliency $S(r, c, k)$ is given by

$$S(r, c, k) = \sum_{r', c'} ||Color(r, c, k) - Color(r', c', k)|| \quad (5.7)$$

where $Color(r, c, k)$ denotes the mean color value of macro-block (r, c) in frame k .

Once we have obtained the spatial and temporal saliency maps, the frame can be segmented into foreground and background regions. In our current work, the foreground regions are assumed to be the macro-blocks with high salient value, and then sort the block saliency value in descending order. The first several blocks whose accumulative sum reaches a given threshold of the total saliency value are labeled as an initial foreground estimation. The threshold is set to be proportional to the variance, and the initial results are rectified by erosion and dilation, which is the simplest segmentation approach. More sophisticated algorithms (such as [CZM⁺11]) can offer much better results but with higher computational complexity. In our current work, an approximate foreground can give acceptable results in the next stage of processing.

Let F_s denote the foreground regions estimated by spatial saliency, while F_t denote regions from motion information. It is known that motion within a frame is made up of foreground object motion and camera motion. Since we are only interested in the foreground motion, the camera motion is removed. The global motion is modeled by affine morphing: $\mathbf{x}^* = A\mathbf{x} + b$. \mathbf{x} and \mathbf{x}^* are pixel indices before and after morphing respectively. Morphing parameters are computed by the motion vectors of macroblocks outside F_t . Once global

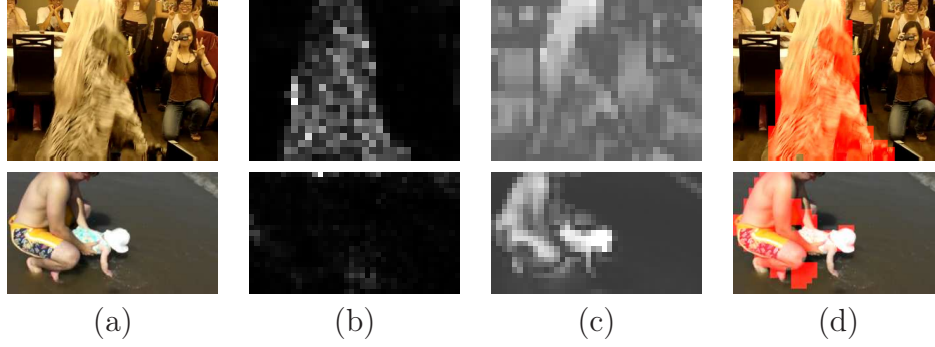


Figure 5.2: Saliency and detected foreground. Column(a) Original frames; Column (b) motion saliency; Column (c) spatial saliency; Column (d) fused foreground.

motion has been removed from foreground motion vectors, the motion saliency map is rectified.

To fuse the two results, we assign *trust values* to the temporal and spatial saliency respectively. The trust value is the confidence of the estimated salient information. Let $Rec_s(k)$ and $Rec_t(k)$ denote the minimal rectangular region that covers $F_s(k)$ and $F_t(k)$. The trust values are given by

$$\gamma_s(k) = \frac{|F_s(k)|}{|Rec_s(k)|}, \gamma_t(k) = \frac{|F_t(k)|}{|Rec_t(k)|} \quad (5.8)$$

where $|\cdot|$ computes the number of macroblocks within the region. We can see that if the salient regions are isolated and sparse, it will have a relatively lower trust value. According to previous discussions, temporal saliency is taken as the critical cue for interest estimation. Due to the limitation of home videos, spatial saliency is often not reliable. Admittedly, motion fails to offer informative foreground detection when foreground motion is subtle. However, in practice, users seldom use videos to record still objects. A photograph will be the more suitable choice in such circumstance. Therefore, temporal saliency is treated as the priority cue when we fuse the two.

$$SM(r, c, k) = \begin{cases} M(r, c, k), \gamma_t(k) > \theta_1 \\ S(r, c, k), \gamma_t(k) < \theta_1 \& \gamma_s(k) > \theta_2 \\ \gamma' M(r, c, k) + (1 - \gamma') S(r, c, k), \text{others} \end{cases} \quad (5.9)$$

where $\gamma' = \frac{\gamma_t(k)}{\gamma_t(k) + \gamma_s(k)}$. The fused foreground regions $\Omega(k)$ are recomputed based on $SM(r, c, k)$. Figure 5.2 shows the two saliency maps and the detected foreground after fusion. The first row shows a case when temporal saliency is more reliable with the trust value $\gamma_t=0.89$. Only temporal information is considered in the fusion process. The second row shows the opposite case when $\gamma_t = 0.04$, $\gamma_s = 0.51$. The two saliency maps are fused for a better result.

5.3.2 Subclip Segmentation

Now we want to categorize the frames into interesting and un-interesting ones for the subsequent processing steps. Generally speaking, foreground information is more important and attractive to the viewers. Foreground saliency is used to evaluate frame interest.

$$MI(k) = \frac{1}{RC} \sum_{(r,c) \in \Omega(k)} SM(r, c, k) \quad (5.10)$$

where $\Omega(k)$ is the set of the foreground macro-blocks of frame k . Frames are categorized as interesting or ordinary based on their MI . To remove the influence of noise and unintended camera shakes, we use the least square

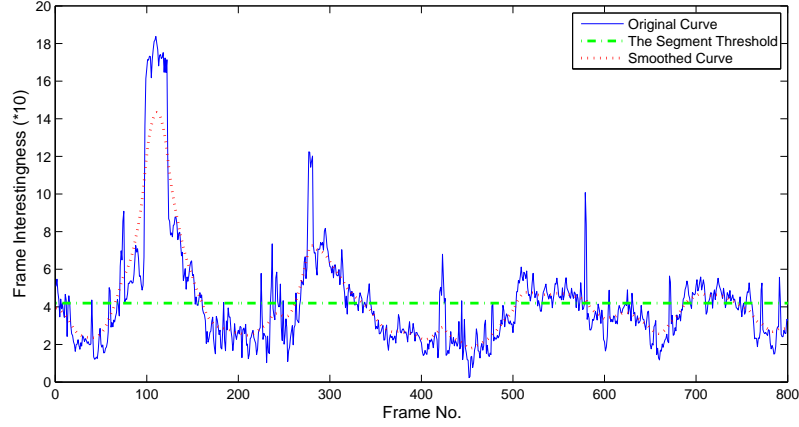


Figure 5.3: Frame Interest.

scheme to smooth the MI curve. Specifically speaking, we want to minimize

$$\sum_k (MI(k) - MI^*(k))^2 + \lambda (\nabla MI^*(k))^2 \quad (5.11)$$

where λ is the smoothness constraint parameter which controls the smoothness of the estimated interest curve.

Figure 5.3 shows the original and smoothed frame interest curve MI . The dotted line shows the estimated $MI(k)$ for each frame. It fluctuates due to noise and the unintended camera shakes. The solid lines shows the smoothed curve. We used the least square scheme to detect the underlying smooth curve $MI^*(k)$. A threshold parameter δ (shown as the dashed line) is empirically chosen to segment the video clip into interesting and ordinary. In our experiment, δ is set to be proportional to the average interest value. The frames that lie on the boundaries of subshots are denoted as $f_1, f_2 \cdots f_N$. The set of interesting subshots is denoted as I , while that of the uninteresting ones are denoted as U .

5.3.3 Retargeting, Reprojection and The Fusion

Follow the notations in the previous sections. For a given frame k , $a(k)$ is the projection duration of frame k , $(r_0(k), c_0(k))$ denotes the center of the cropping window and $\omega(k)$ represents the zooming factor. Let $\mathcal{P}(\cdot)$ denote the re-projecting process and $\mathcal{R}(\cdot)$ denote the retargeting process. $\mathcal{P}(a)$ displays the frame with duration a . $\mathcal{R}(r_0, c_0; \omega)$ sets the cropping window center at (r_0, c_0) and the ratio between the size of the window and that of the original frame is ω . Curves $a(k)$ and $\omega(k)$ are computed based on the mean saliency of clip segments $f_1, f_2 \cdots f_N$ ([XK10a] [XK10b]). To estimate the cropping window center (r_0, c_0) , we consider only the frames with high trust values $\gamma_t > \theta_1$ or $\gamma_s > \theta_2$. According to Equation 5.9, the frame saliency MI is decided solely by temporal saliency M or spatial saliency S under these conditions. Given ω , the optimal cropping window is the one that captures the highest frame saliency. (r_0, c_0) are interpolated for the rest frames.

The zooming factors of retargeting and frame durations of reprojection are decided by subclip-based interest. But direct multiplication of the two brings in unpleasant outcomes. Even though close-shot and slow motion can both highlight interesting events, their dual simultaneous effects may be disastrous. Just imagine the extreme close-shot with very slow projection motion. A little wind kindles, much puts out the fire! Thus the two should be fused for more reasonable outputs. Moreover, the projection velocity manipulation can be regarded as one of the many special effects that help to make videos more interesting. Professional directors alter the projection velocity at certain time to maximize the visual impact, but not every time. Re-projection makes viewers to feel that motion in the video is accelerated/slowed down, which makes it abnormal. Continuous abnormal motion patterns make the video

wired. Thus, the fusion process also helps to suppress the effect of projection adaptively.

The fusion scheme $\mathcal{F}(\cdot)$ can be modeled as

$$\mathcal{F}(k) = \mathcal{P}(\alpha_1(n)a(k)) + \mathcal{R}(\alpha_2(n)\omega(k)) \quad (5.12)$$

where n is the number of the subclip that frame k belongs to, while $\alpha_1(n) + \alpha_2(n) = 1$ are the weighing factors. Considering that the retargeting process redistributes the spatial saliency while reprojecting concerns temporal saliency, it is reasonable to relate the weighing factors to corresponding saliency value. $S(k)$ denotes the spatial saliency and $M(k)$ represents the temporal saliency of frame k . The trust values are γ_s and γ_t respectively. The fusion process is based on the observation that when motion is more important than spatial properties, re-projection takes over retargeting to exhibit motion details. Otherwise, when there is no detected reliable interesting motion and the spatial features are attractive, retargeting is applied. To ensure smooth curves for mapping parameters, Equation 5.12 is rewritten as

$$\mathcal{F}(k) = \begin{cases} \mathcal{P}(a(k)), \bar{\gamma}_t(n) > \theta_1 \\ \mathcal{R}(z(k)), \bar{\gamma}_t(n) < \theta_1 \& \bar{\gamma}_s(n) > \theta_2 \\ \mathcal{P}(\gamma_1''(n)\Theta(a(k)) + 1) + \mathcal{R}(\gamma_2''(n)\Theta(z(k)) + 1), \\ others \end{cases} \quad (5.13)$$

where $\Theta(x) = x - 1$, $\gamma_1''(n) = \frac{\bar{\gamma}_t(n)}{\bar{\gamma}_t(n) + \bar{\gamma}_s(n)}$, $\gamma_2''(n) = 1 - \gamma_1''(n)$ and $\bar{\gamma}$ is the corresponding mean value of γ in the subclip. Here the cropping window

position (r_0, c_0) is omitted, because it is completely spatial modification. Once the zooming factor ω has been decided, the cropping window can be computed based on spatial saliency only.

5.3.4 Frame Re-Rendering

Video rendering is essentially to crop the frames and resize them to retarget, and to add or delete certain number of frames to lengthen or shorten projection duration. But duration function $a(k)$ is not integral at every point, therefore we could not manipulate each frame directly based on $a(k)$. As a compromise, let every j frames be a set, and for each set of the subclip, we estimate the number of frames to be added according to $a(k)$. In our implementation, we empirically set j to be 5 which works well for a variety of videos. Thus, for every 5 continuous frames in I ,

$$\{\sum_m^{m+4} a(k) - 5\} \quad (5.14)$$

decides the amount of frames to be added among them, where $\{x\}$ represents the minimal integer that is not less than x . According to Rule 1, we want to control the output video length. To rule out the influence of uncontrollable mantissa, we choose the minimal integer that is larger than real duration function accumulation $\sum_m^{m+4} a(k)$ instead of rounding it. So for each subclip in I , frames are added from the beginning, and when the total number of newly interpolated frames reaches

$$\sum_{f_i}^{f_{i+1}} a(k) - (f_{i+1} - f_i) \quad (5.15)$$

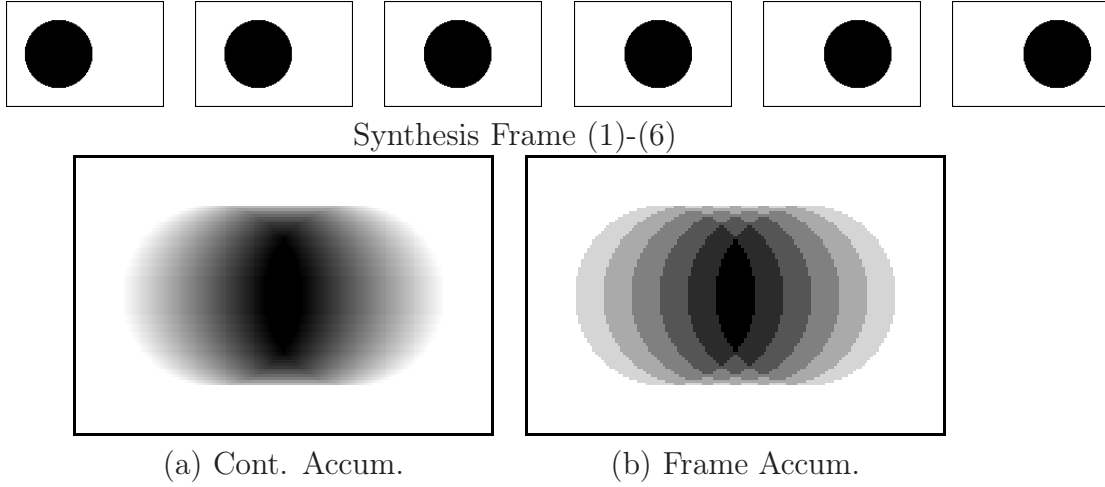


Figure 5.4: The synthesis example for the accelerated frame generation. Frame (1)-(6) are 6 continuous frames. The object motion velocity seems to increase by reducing the projection time. Within the same exposure time, the trace of moving object is longer and results in more noticeable motion blur. Figure (a) shows the ideal continuous combination of the 6 frames. In our implementation, we use the weighting combination in Equation 5.19 to accumulate temporal information (b).

the remaining frames of the subclip are left untouched. Since we have chosen the integer ceilings at each set, the quantity of interpolated frames will reach the target number before the last set of the subclip.

Subclips in U have to be accelerated, i.e. the number of frames have to be reduced. For each frame in the uninteresting subclip, its projection time is reduced to $a(k)$. While being projected, the frames are updated at a constant velocity (for example $1/25\text{s}$). Thus when projection duration of original frames is reduced, newly rendered frames will contain information of more than one original frame.

$$F_{new} = \sum_{\sum_k a(k)=1} a(k) F_{old}(k) \quad (5.16)$$

where $F_{old}(k)$ represents the k -th original frame. Thus we linearly com-

bined frames, and $a(k)$ are the weighing coefficients, as shown in Figure 5.4. Here we do not adopt any motion compensation scheme when we assemble the neighboring frames. In the case that object motion accelerates, frames will be blurred. So we artificially introduce motion blur by linearly assembling original frames. The sum of several continuous $a(k)$ might not always be exactly 1. In this case, the coefficient of last frame is separated into two:

$$F_{new}(m) = \bar{a}(k_{m-1})F_{old}(k_{m-1}) \quad (5.17)$$

$$+ \sum_{k_{m-1}+1}^{k_m} a(k)F_{old}(k) - \bar{a}(k_m)F_{old}(k_m) \quad (5.18)$$

where $\bar{a}(k) \geq 0$ and

$$\bar{a}(k_{m-1}) + \sum_{k_{m-1}+1}^{k_m} a(k) - \bar{a}(k_m) = 1 \quad (5.19)$$

5.4 Experimental Results

In our experiments, we choose 13 home-produced video clips of different content, ranging from daily family life, stage performance, sleeping pets and sports¹. Some of the videos are personal collection while the rest are downloaded from YouTube. 19 users are invited to watch the video clips and then score them from different aspects. The subjective score ranges from -5 to 5. Details of the user study are shown in Table 5.1. The results are shown in Figure 5.6.

¹The experimental results can be found at <http://www.youtube.com/playlist?list=PL704AB6A7E6039DC4>

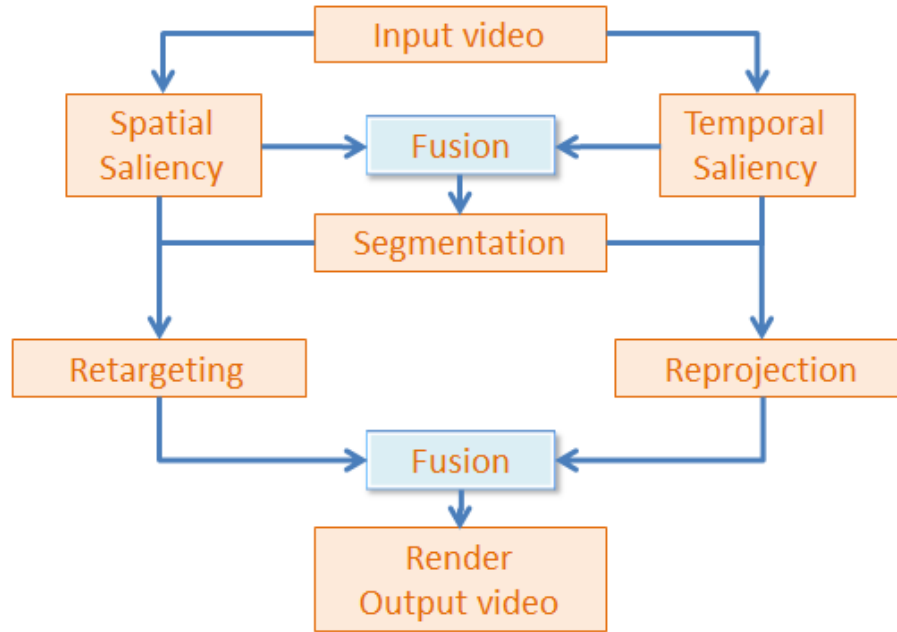


Figure 5.5: The flowchart of the whole system.

Abrv.	Score (-5 to 5)	Explanation
SD.	-5(Unacceptable) to 5(Excellent)	Segments Detection: Are the detected interesting segments successful?
CW.	-5(Much worse) to 0(Comparable) to 5(Much Better)	Camera Work: How is the camera work in the output video compared with the input one?
PS.	-5(Much worse) to 0(Comparable) to 5(Much Better)	Projection Speed: How is the altered projection speed compared with the input one?
FR.	-5(Much worse) to 0(Comparable) to 5(Much Better)	The Fusion Results: How do you like the output video compared with the input one?

Table 5.1: Details of User Study.

During the experiment, the original input video is shown at the very beginning, and a small window showing the detected interesting segments is placed at the lower-right corner with the interesting segments are labeled in red. The participants score the correctness of segment detection. According to the user study (Figure 5.6), the scores of most videos are above 2.5. The highest score is obtained at clip 11 (3.68) with the lowest standard deviation (0.82). The video is segmented into 2 parts. Since the content shows a sudden spatial and temporal change between the two segments, the segment boundary is intuitively straight-forward. Thus experiment and user study shows the consistence between proposed segmentation scheme and users' perception. The lowest score (1.74) occurs at Clip 13 with the highest standard deviation (1.82). Users respond quite differently to the interest segmentation towards this video. Examining the video itself, we find that the spatial saliency is almost constant while the change of the motion pattern is not significant. It seems difficult to offer a persuasive segmentation based on the proposed scheme.

Then the output video is shown right after the input clip, and the users are invited to score the output video from 3 aspects: the camera work, the projection speed and the fusion results. They are asked to compare the output video with the previously-shown input one, and score the improvement from -5 (much worse than the input one) to 0 (almost comparable to the input one) to 5 (much better than the original one). As shown in Figure 5.6, almost all the videos show score increments from separate re-projection/re-targeting to the fusion results and standard deviation decreases as well. We could arrive at the conclusion that users feel better to fuse the two scheme instead of performing either one independently.

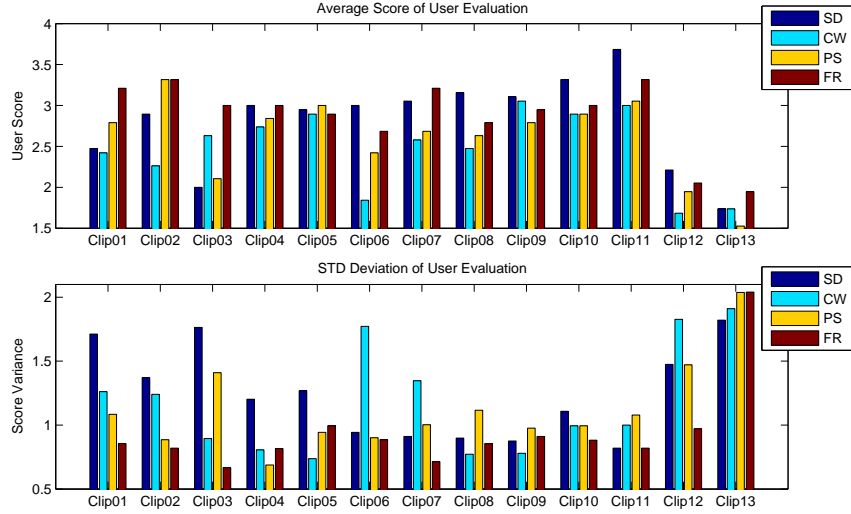


Figure 5.6: Subjective User Evaluation. SD: segment detection. CW: camera work. PS: projection speed, FR: fusion result.

To discuss our results in details, we take the video clip of cycling (Test02) as an example, the parameters of which are shown in Table 5.2. This is a simple example because the input video is segmented into only 2 subclips. Since the background is static, the trust values of spatial saliency are almost constant over the whole clip. Table 5.2 shows the same value 0.35 in both segments. The temporal saliency trust value, however, drops from 0.17 to 0.06, for there is almost no motion in the second segments. Thus the average MI over the two segments show dramatic difference. \bar{MI} of the interesting segments (the first one) is almost 8 times higher than that of the uninteresting subclip (the second). In the output video, the projection speed is increased by over 50% in the first segment and decreased by 50% in the uninteresting segment. Compared with the noticeable difference in projection speed, the zooming factor ω does not change that much, only 1.15 for the interesting part. This is due to the fact that the change of spatial saliency (though seems to be higher in γ_s) is much lower than that of the temporal saliency, and thus

Segments	γ_s	γ_t	Ave. MI	Frm. Ratio	Ave. ω
1	0.35	0.17	197.47	162%	1.15
2	0.35	0.06	24.76	50%	1

Table 5.2: Output Rendering Parameters of Clip 02.

the result shows the same trend. According to the user evaluation, this video enjoys the score 3.21 for the fusion result and a standard deviation 0.85.

The proposed method relies on low-level features, i.e. color and motion. The two channels of information ensures a reliable interest estimation when one channel fails. For example, video Test02 is of low spatial saliency but high temporal saliency, while video Test03 shows the opposite case. However, when both channels fail to offer useful information, the proposed method fails. Video Test13 shows a negative example. In this video, the still background is of complex texture and bright colors. The spatial contrast between foreground and background is low. Meanwhile, dancer’s motion pattern shows no significant differences over the whole video. Therefore, neither the spatial saliency nor the temporal saliency can offer trust-worthy information for segmentation, re-projection or re-targeting. The users’ evaluation (Figure 5.6) shows the lowest scores for both segmentation and fusion process. Interestingly, the users’ attitudes also show high divergence (the highest standard deviation for all questions). It seems that even for human beings, this is a difficult case to handle.

5.5 Discussions

In this chapter, we present a post-editing scheme for home produced video clips. It fuses retargeting and re-projection based on sequence shot technique, aims to enhance the aesthetic interest. Videos of a single shot are treated

as long takes. We discussed the applicable long take editing schemes and proposed a computational model of the fusion. We also proposed a perceptive model to evaluate frame interest and segment a single-shot video clip based on frame interest. Our contributions include the adaptive fusion model of re-targeting and re-projection, the model of video interest estimation and manipulation. We targeted home produced videos, because camera work and temporal interest distribution are two factors that most home videos lack.

In the current model, we try to extract the motion information frame by frame, and integrate the frame-based information by global optimization to rule out noise. On one hand, the frame-based approach is adopted to interpret the global motion information. This is not an optimal approach, but a compromise because of the current motion analysis techniques as well as computational efficiency. On the other hand, since our proposed scheme is based on frame information, we manipulate the local properties of the input videos. This is different from current video softwares (such as Muvee), which perform only global modifications, including the boundaries, the transitions and global special effects. In our experiments, we do not compare our results with their results, because the focus is completely different.

The next stage of work may also include a video analysis framework based on high level features. Human judgement about what is interesting or uninteresting are based on a combination of factors. Moreover, we do not take the audio track into consideration when we build the projection duration adjustment model. In home music video production, it is more complicated to match projection rhythm and background music tempo. Therefore, we want to look into the issue of frame re-projection based on the synchronization between the audio and video tracks.

Finally, our proposed scheme serves as one additional option for post-processing. In practice, it may be more reasonable to invite users to decide if they would like to utilize this special effect in certain clips or not. At present, since we take the shot as a long take, every frame will be processed for either re-targeting or re-projection. To make our system more user-friendly, an adaptive adjustment may be considered in the next stage of work.

Aesthetics for Non-Traditional Medium: Force-Model Based Aesthetic Online Advertisement Selection

Online advertising aims to insert advertisements in an efficient and non-intrusive way so as to attract consumers. In this chapter, we present a web page advertisement selection strategy based on the force model. It refines the results of contextual advertisement selection by introducing aesthetic criteria. The web page is semantically segmented into blocks, and each block is an element in the two-dimensional screen. Aesthetic theories on the screen balancing are adopted in the proposed system. We compute the graphic weights of blocks and treat them as vertices in a graph. Weighted graph edges are the forces between the elements. The aesthetically optimal advertisement is the one that balances the force system. Subjective experimental evaluations show visual improvements when compared with randomly selected advertisements from the contextually related candidates.

6.1 Introduction

Online advertising has become an essential component of the modern internet. It is superior over traditional advertising medium for its wide coverage and low cost. The increasing number of internet users has attracted many companies to promote their products via the internet. Their advertisements are delivered to users from different channels, webpage advertising, search engine advertising, in-image/video advertising and so on.

Among the various types of online advertising, we are specially interested in webpage advertisements, because browsing the web is the most important daily activity for many internet users. The most important form of these advertising, also called display advertising, is web banners. In the traditional advertising medium, advertisements may be ignored by customers because the content is not related to the activities users are performing. Online advertising, on the contrary, can link the advertising activity to the information it supports.

Contextual advertising aims to utilize local context and sentiments for identifying relevant advertisements. It makes use of the users' immediate interests on the third party web sites, and advertisements are displayed based the content of the web pages. However, contextual advertising schemes alone can not assure an optimal retrieved advertisement candidate, because the pool of advertisements provided by the advertisers is so large that many advertisements are contextually similar to each other. To further improve the efficiency of automatic online advertising, advertisements that are visually compatible with the web site are selected by introducing appropriate aesthetic criteria.

The essential features of online advertising are [HMH10]: effective target,

scalability, non-intrusiveness, and attractiveness. The first two consider the issue of contextual advertising schemes, and the remaining characteristics necessitate adding of aesthetic requirements to the advertising strategies. In this chapter, the advertising scheme that satisfies the non-intrusiveness and attractiveness requirements is defined to be the integration strategy that creates the highest visual pleasure.

Online advertising strategies aim to make the advertisements adaptive to the supported web page, maximize the efficiency of attracting consumers and minimize the intrusiveness at the same time. They act as supplementary parts of a web page, which, by the nature of advertising, are not expected by the users when they open a web page. Therefore, traditional advertising schemes passively add the advertisements to the rest of the web page on aesthetics, without considering what kind of contribution an inserted advertisement can make to the supported web page.

It has been suggested that aesthetics has important implications on the users and could enhance the feelings of web applications positively [SN00]. Therefore, efforts have been devoted to improve the web page aesthetics. The inserted advertisement, acting as an element of the web page, should not stand alone. From the aesthetic point of view, images are of much higher visual importance than text, which is often the main body of a web page. Therefore, if inserted appropriately, advertisements can serve as an embellishment and help to improve the aesthetics of the web pages. In an ideal case, advertisements can actively adjust the web page layouts and make them visually pleasant.

Taking a web page as a semi-structured image (the common practice as in [SN00],[LMNN10]), we consider the screen structure theories in aesthetics [Zet99]. There are forces between elements within the screen, and a success-

ful visual product requires a reasonable layout of these pictorial elements to balance the forces. We find it quite similar to the forced-based graph drawing algorithms, which also try to arrange the spatial elements and strive towards a stable state of a given force-based system.

Graph layout problem is related to the area of information visualization. Typically speaking, it is a pictorial representation of vertices and edges. Algorithms aim to derive a two-dimensional depiction of graphs which look visually pleasant and aesthetic. Force-based systems are one of the various approaches to solve the graph drawing problem. The forces are defined based on some physical metaphors, such as springs and electric charges. Attractive forces are assumed to exist between neighboring nodes and repulsive forces between all pairs of vertices. The system continuously modifies an initial vertex layout state until it reaches an energy-minima state. Visually speaking, the desirable layout shall be of short and even edge lengths and well-separated vertices.

In this chapter, we propose an aesthetic online advertisements selection algorithm, which is based on the graph layout theory for aesthetic visual pleasure. The idea of vertices layout is extended to the webpage block layout. A webpage contains several blocks, which can be regarded as the nodes as in the graph systems. However, the edges in the graph drawing problems do not really exist in the web page, so what we are really interested in are the interactions between web page blocks, as the physical forces built in the force-based graph drawing problems. Here comes the major differences between the proposed advertising strategy and traditional graph drawing problems, which have been listed in Table 6.1. In the current work, all the node positions have been pre-fixed by the publishers, while in graph drawing problems the positions are unknowns. We aim to find advertisements which make the force-

	Parameters	Unknowns	Targets
Graph Drawing	Inner Forces	Node positions	Arrange the layout of nodes
advertisement Selection	Block Position	Forces	Select an optimal advertisement

Table 6.1: Comparison between graph drawing and the proposed advertisement selection framework.

system stable, while they stabilize the system by arranging the placement of nodes.

Our contributions include the formulation of the aesthetic force system and the solution for an optimal, aesthetically pleasant advertisement selection problem. To the best of our knowledge, this is the first attempt to adapt the graph layout theory in online advertising for a visually pleasant advertisement selection.

The rest of the chapter is organized as follows. Section 2 gives an overview on the related works. In Section 3, we define the problem of aesthetical advertising based on art theories. The proposed force-based system is introduced in Section 4. The experimental results are shown in Section 5. Finally, Section 6 gives the conclusions.

6.2 Previous Work

Efforts that contribute to the online advertising studies can be categorized into two dimensions: relevance matching and position detection [LMNN10]. The first one concerns the issue of choosing advertisements that are related to their publishing platform, while the second one aims to find proper places to insert the selected advertisements. The latter one is especially important for in-image/video advertising ([MHL08] [MHYL07]). In our current work, we

restrict our scope to the relevance matching only, since the location and size of advertisement are assumed to be preserved and fixed by the web publishers. In practice, when applying online advertising services, publishers have pre-designed their webpage architectures and embed advertisements by API, which is provided by the advertisers.

Research works on relevance matching can be further classified into 3 directions: by keywords, by content and by user information. Conventional keyword advertising analyzes the salient keywords of webpages and match them with the advertisements keywords provided by advertisers [MSVV07]. Google AdWords [AdW10] and advertisementsense [AdS10] treat image and video advertisements as general text. advertisements are selected and displayed based on the contextual relevance between the webpage and the advertisements themselves.

Since context plays an important role in the way that advertisements are perceived, efforts have been made to enhance the user experience in online advertising. More sophisticated work has been done in ImageSense [MLHL12] and VideoSense [MHYL07]. They embed advertisements at appropriate positions within images and videos respectively. In addition to purely text information, visual similarity is also taken into consideration in these advertising frameworks.

Style-wise advertising is mainly considered in image/video advertising because visual elements are closely related to the selection and placement of advertisements. [LMNN10] presents a style-wise advertising framework for web pages. The system automatically detects blank areas of the webpage and embeds advertisements that are style-consistent. In their work, only color information is considered. The style consistence is assumed to be the color

similarity, so that users may perceive the advertisements as a natural part of the original page.

In this chapter, we present an advertisement selection solution from computational aesthetic point of view. The idea of *Computational Media Aesthetics* was put forward by Dorai et al. in 2001 [DV01]. It makes computational analysis of media elements combination, and aims to make use of such information to create tools that could add perceivable intent into media data.

Models are built to predict the aesthetics of a given media piece. One approach is to build up series of rules for aesthetic assessment. [FNH08] concerns with image harmony, which is an important aspect of aesthetics. The authors conducted a series of subjective experiments to see how low level features can predict harmony in an image.

However, from the aesthetic point of view, users' perception often affects the mood and emotion, which makes the problem difficult. Low-level features are insufficient to characterize high-level perceptions of aesthetics. Datta et al. [DLW08] put forward related questions to the gap between image aesthetics and visual content. They also propose a machine learning scheme that infers image aesthetics by the visual content. They still make use of low-level features, but they try to explore the relationship between emotions and these low-level information. [KTJ06b] uses high level semantic features to measure the quality of photos. They classify between the professional photos and snapshots of low quality. [JLC10] builds an automatic aesthetic assessment system for photographic images. Two learning machines, fine-granularity and coarse-granularity score predictions, are designed based on a series of visual features.

Other than the studies on computational aesthetics of images, which aim

to evaluate their appeal, our work is also related to the aesthetic studies of web page layout. In addition to the quality of inserted advertisements, users also attach importance to the design of web pages. A beautifully designed web page layout will definitely facilitate users' experience when they are browsing the pages ([SJ00], [TDG01] and [TCKS06]). [MHB08] explores the relationship between subjective aesthetic appearance and the visual complexity. To evaluate the visual pleasure of web pages, [WCLH10] proposes a learning-based computational aesthetic approach. In their model, the page layout include 4 aspects of features: text, image, background and the layout. Based on the subjective user evaluation of different webpages, they build a cost-sensitive SVM as the classification model.

Our proposed advertisements selection system is also closely related to the forced-based graph drawing algorithms. The force-based model is designed for drawing undirected graphs by Eades [Ead]. The graph is modeled as a physical system which is made up of steel rings and springs. Starting from some initial state of the vertices layout, the spring forces make the steel rings to move until the system reaches a minimal energy state. But his work did not incorporate Hooke's law. Thomas et al. [FR91] improves the spring-embed model. The assumptions that: 1) vertices that are neighbors attract each other; 2) all vertices repel each other, simplify the original n-body problem. Moreover, they use forces to induce acceleration, which results in static equilibria rather than dynamic equilibria.

6.3 Aesthetic Advertising

Fern’s team has carried out a systematic data-driven experiment studying the correlation between the visual appearance of advertisement and whether matters for propensity of user response [JA12]. Their investigation conclusively demonstrates the non-trivial impact of visual appearance on CTR(click-through-rate). In their experiment, the model taking the visual quality into consideration predicts CTR more than 3.7 times better than that purely based on CTR distribution. They quantitatively justify the traditional belief “visually appealing display advertisement can perform better in attracting online users”.

With respect to aesthetic advertising, efforts have been made to detect proper regions for advertisement placement. In our current scope of work, however, the position and size of advertisements are assumed to be fixed. It is the common practice for the publishers, who have pre-designed the web page layout and use API provided by the advertisers to insert advertisement. Based on this assumption, the proposed advertisement selection problem is given by

Problem Definition(I): Given a webpage W and fixed advertising locations, find the optimal advertisement A from the set of advertisement candidates S , which are contextually related to W , such that the integration of web page W and the selected advertisement A reaches the highest visual pleasure.

A web page is essentially a two dimensional space. When artists structure the two dimensional field, they work with various screen forces and try to arrange the static spatial elements so that they look right to the viewers. These

forces clarify and intensify events within the spatial field [Zet99]. The arranging activity is the artistic compositional process, i.e. a pleasing arrangement of essential static pictorial elements within an image. If we treat the web page as an image, the aesthetic selection of advertisements becomes an image structuring problem.

Balancing the spatial forces is the most important scheme of stabilizing the two dimensional field. Typically speaking, there are two kinds of forces: *graphic forces* and *frame magnetism* [Zet99]. Every object in the image carries its own graphic weight, which can be determined by: 1) the relative size; 2) shape; 3) orientation; 4) location within the screen; 5) color. Frame magnetism in the forces frame edges place on objects located near them. A good image composition requires proper force-arrangement. Even though unstable spatial force distribution, so-called *labile balance*, can create special effects in extreme cases (a heavily tilted horizon emphasizes tension and velocity in a racing image, for example), when we come back to the problem of advertisement selection, a stabilized web page distribution (*stable balance*) is more appreciable, considering the fact that most of the web page content is textual, which is horizontal and of high visual order. Taking the force distribution theory into consideration, we restate the proposed advertisement selection problem:

Problem Definition(II): Given a web page W and the fixed advertising location w , find the optimal advertisement A from the set of advertisement candidates S , which are contextually related to W , such that the insertion of A at location w can balance the force distribution within W .

To computationally solve the balancing problem, the force-based algorithm

in graph drawing provides an effective way to arrange the graphic forces and offers aesthetically pleasing display results [FR91]. The whole system is simulated as a physical model, and forces are assigned to edges between adjacent vertices. The *minima* of the system gets to a stable and aesthetically pleasing graph configuration. The proposed advertising strategy is inspired by the force-based algorithm. But instead of finding the optimal vertices layout, we work in an inverse way

Problem Definition(III): Given a vertices distribution configuration W and a vacant position w , find an optimal node candidate A , such that the forced-based system comes to a stable state.

According to the definition, a web page is directly treated as a group of nodes which form a graph system. In practice, a given web page can be semantically segmented into different blocks. Each block is assigned with a saliency value and becomes a node in the graph system. The inner force between each node is the aesthetic interaction between two elements. Since the layout of the webpage has been decided by the publisher, the relative position of the nodes in the graph has been decided. The advertising candidate is inserted with the functionality that the whole force system can reach a stable state.

6.4 Our Approach

The flowchart of our proposed force-based aesthetic advertisements selection system is shown in Figure 6.1, and Figure 6.2 shows the results of corresponding important steps. The first image in Figure 6.2 is a snapshot of a typical news web site with a banner as the header, text as the main body, an image

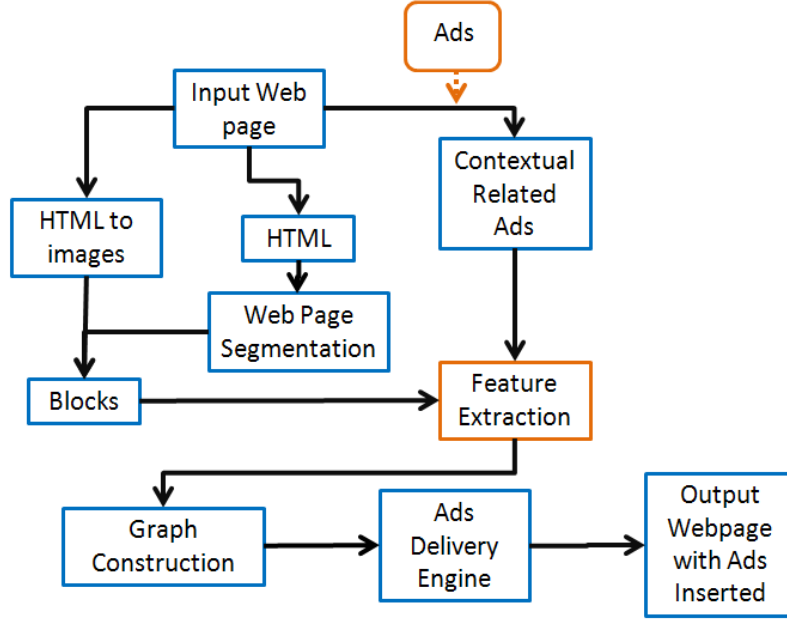


Figure 6.1: The flowchart of the proposed system.

going with the textual content, and navigation links on the edge. The blank rectangular area on the upper-right is the reserved place for an advertisement banner. The second images shows how the web page can be segmented into blocks (pink rectangles). The third image of Figure 6.2 shows the abstracted graph nodes. The sizes of the circles represent the visual importance of the corresponding blocks. The last image shows an example of the generated force system.

6.4.1 Visual Weights of Elements

The input web page can be semantically segmented into blocks using the VIPS algorithm [CYWM03], and the blocks are evaluated independently. Since we are now considering the visual features of the web page, the blocks are taken as images, no matter it is a textual block, an image or the combination of the two. A similar approach is adopted in PageSense [LMNN10], where the

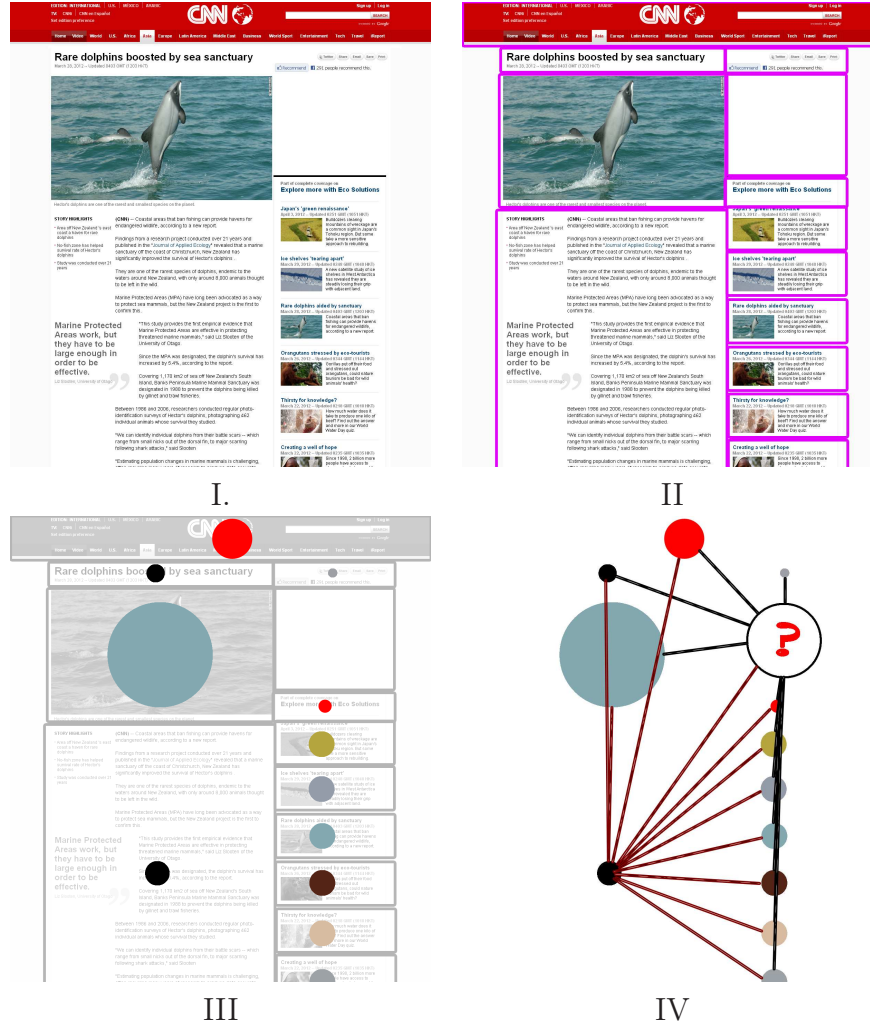


Figure 6.2: The procedures of the proposed system. I. The input web page; II. The input web page is semantically segmented into blocks; III. Blocks are abstracted into vertices in a graph system by feature Area vectors containing the style and saliency information; IV. The graph system is built up by integrating nodes and forces.

received web page is taken as a snapshot as well.

Consider a block B_i . To abstract a node in the graph drawing system from the given block, there are 3 factors that we need to consider: position P_i , visual weight W_i , and color C_i . Position decides the placement of the node in the final graph system. Visual weight is the visual importance of the node, and it will also influence the magnetism between nodes (blocks). Color serves as a style-wise attribute for the visual consistence in the output web page when the ad candidate has been selected and inserted.

We use GBVS [HKP06] to estimate the visual weight W_i . Moreover, since GBVS computes the saliency map of a given image, we define the region of interest to be the area taking up 60% of the salient information. The center of ROI together with the location of B_i within the web page decides the block position P_i .

The color attribute contains three aspects of information: hue C_i^H , saturation C_i^S and brightness C_i^B . C_i^H is the hue of the dominant color, i.e. the one of highest statistical frequency. C_i^S and C_i^B are the corresponding mean value over the block. In practice, the hue layer is quantized into 12 colors, which is the typical number of color patches on a color wheel (Figure 3). Hue ranges from 0 to 1, where red covers the color interval $[0, 1/24] \cup (23/24, 1]$, orange covers the interval $(1/12, 1/6]$, and the remaining 10 colors can be done in the same manner. In the current case, the hue information is derived from the RGB color space, which is also called *additive color wheel*.

The use of complementary color pairs is important in aesthetic color combination, because the contrast between complementary colors makes colors bright and distinguishable. Different color models offer different complementary pairs. In the RGB color model (additive color model), the pairs are given

as: red /cyan, green /magenta and blue /yellow. This is not consistent with artistic complementary pairs: red /green, blue /orange and yellow /violet, as in the red-yellow-blue (RYB) color model.

RYB color model (the Goethe's color wheel) is a traditional set of colors that is more widely used in art and design. Therefore in our proposed system, hue is first approximately converted to the RYB color model by cubic interpolation, where 0 (red in RGB) is converted to 0 (red in RYB), $2\pi/3$ (green in RGB) to π (green in RYB), $4\pi/3$ (blue in RGB) to $4\pi/3$ (blue in RYB) and 2π (red in RGB) to 2π (red in RYB). Let c denote the hue value in the RGB color model, \bar{c} denote the corresponding hue value in the RYB color model, the conversion is given as

$$\bar{c} = 2.25c^3 - 3.72c^2 + 2.5c \quad (6.1)$$

In the following discussions, we still use the notation C_i^H for convenience. But whenever we mention the hue information, it refers to that in the RYB color model. Moreover, we do not perform the conversion directly from RGB to RYB. To make it computational efficient, a backward mapping is performed, i.e. the boundaries of 12 colors in the RYB model are converted to the RGB color space. Because we are only interested in the statistic properties of hue, the segmenting points on the RGB color wheel can differentiate the 12 colors directly. Let ϕ_{c_k} denote the amount of pixels that are of the k -th color (the clockwise looping sequence from red to purple red), the dominant color is defined to be

$$C_i = \operatorname{argmax}_{1 \leq k \leq 12} \phi_{c_k} \quad (6.2)$$

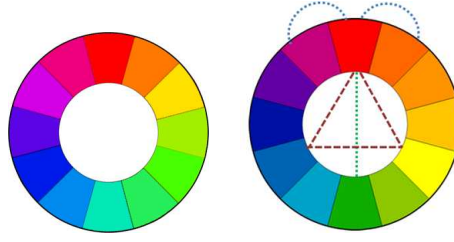


Figure 6.3: The color wheels and color harmony. Left: RGB wheel. Right: RYB wheel. Take red ($c_i = 0$) as an example on the RYB color wheel, the 3 sets of harmonized color patches are: red/red purple & red/orange(the dashed-blue-line), red/green (the dashed-green-line), red/blue,red/yellow(the dashed-red-line).

For the convenience of the following discussions, the interval $[0, 1]$ is then stretched to $[0, 2\pi]$ of the 12 colors.

6.4.2 Force-based System Formulation

Given a block B_i , the corresponding feature vector is denoted as f_i , which is made up of its saliency S_i , position P_i , and color C_i (C_i^H , C_i^S and C_i^H). S_i decides the size of vertices in the graph, P_i and C_i decide the corresponding properties of the nodes. From the aesthetic point of view, the layout design aims to balance the forces within the two-dimensional field. Therefore, at the very beginning we will build up the force model.

As being discussed in Section 3, the graphic weights of objects are decided by several features: size, shape, orientation, location and color. Actually, an object does not need to display all these features to have the graphic weights. In our current case, the web page blocks are almost of the same shape (rectangular), and after object abstraction, the orientation information has been omitted. Since the main body of a web page is often the textual content, which is strictly horizontal, Therefore, only three features are considered in our model:

Factor	Heavy	Light
Size	Large	Small
Location	Corner	Centered
	Right	Left
Color	Hue: Warm	Cold
	Saturation : Strong	Weak
	Brightness: Dark	Light

Table 6.2: Factors influencing graphic weight [Zet99].

- **Graphic Weight.** Every graphic object appearing on the two-dimensional screen carries graphic weight, which is somewhat similar to the weight of an object [Zet99]. Factors influencing graphic weights have been listed in Table 6.2. In the proposed model, it is denoted as a 3-dimensional vector gw_i : the saliency S_i , the ratio between the distance of the node to the furthest frame corner and the nearest one D_i (estimated by P_i), and color information cl_i

$$gw_i = (S_i, D_i, cl_i) \quad (6.3)$$

$$cl_i = a_1(\cos(C_i^H) + \varepsilon) + a_2C_i^S + a_3C_i^H \quad (6.4)$$

where the 3 weight coefficients a_1 , a_2 and a_3 are empirical decided by the importance of the 3 features. In our experiments, we set them to be 0.4, 0.4, and 0.2. The function $\cos(C_i^H) + \varepsilon$ computes the hue contribution based on color warmth, with red ($C_i^H = 0$) the highest and cyan ($C_i^H = \pi$) the lowest (Figure 4). Parameter ε removes the minus values. In our experiments, it is set to be 1.1.

- **Internal Forces.** The internal forces are the interactions between objects, which behaves as attractive forces or repulsive forces. The mass attraction theory states that objects of higher graphic weights attract objects of lower graphic weights, but not vice versa. Moreover, we tend to combine the graphic weights of neighboring objects, especially when they are of similar colors. We do not directly use the Hooke's law to model the inner forces, which is adopted in traditional force-based graph drawing systems [FR91]. In the current case, vertices carry graphic weights themselves, so their interactive forces can be modeled as mass gravity. Given a block B_i , the force introduced by another object B_j is defined as

$$If_{ji} = \begin{cases} k_1 \frac{gw_j}{d^2}, & \cos(cl_i)\cos(cl_j) \leq 0 \\ d^2 \frac{gw_j}{k_1}, & \cos(cl_i)\cos(cl_j) > 0 \end{cases} \quad (6.5)$$

where d is the distance between the two vertices $d = |P_i - P_j|$, k_1 is an empirical parameter which adjusts the contribution of graphic weights. gw_j denote the graphic weight of the block B_j . When the colors of the two blocks are of the same warmth (cold colors or warm colors), the inner force is attractive. Otherwise, if one is cold while the other is warm, it behaves as repulsive force. Notice that the model does not really follow the gravity properties, because the inner forces on the two objects are not equal when they have different graphic weights. The reason has been discussed previously: the mass attraction theory does not assume that the forces are mutual. Larger objects places higher

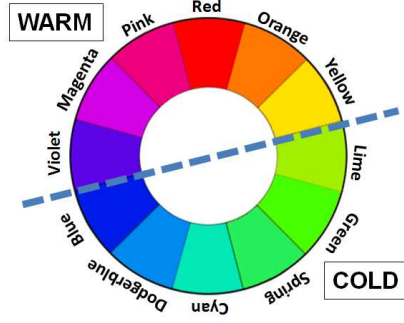


Figure 6.4: The segmentation of cold and warm colors. Warm colors that are further away from the segmentation line have higher graphic weights, and it is the same as the cold colors.

attractive forces on smaller object, but the lighter objects do not.

- **Frame magnetism** is the force that the edges of the screen place on objects. The closer the object is to the edges of the screen, the more powerful the magnetic force will be. Therefore, the force is modeled as

$$fm_i = k_2 ||gw_i||/d_2^2 \quad (6.6)$$

still the parameter k_2 is empirical, an d_2 is the distance from the object to the nearest edges, The direction of the force is from the object to the screen boundary.

One thing to mention that d_1 , d_2 and D_i are not the Euclidean distance, but the ratios between Euclidean distance and the length of the corresponding edges. For example,

$$d_2 = \min\left\{\frac{x_i}{w}, \frac{w - x_i}{w}, \frac{y_i}{h}, \frac{h - y_i}{h}\right\} \quad (6.7)$$

where $P_i = (x_i, y_i)$ represents the position of block B_i . w and h are the width and length of the web page snapshot respectively.

The functionality of colors has always been an important topic in computational aesthetics. Human perception is influenced by colors but the effects themselves still remain an elusive subject. Even though there are several plausible color theories, no single one can explain all the perceptual phenomena of colors. Fortunately, even though we lack of a precise scientific color theory, there is enough knowledge to predict how colors could go harmonized well with each other. It has been widely accepted that colors go together when they are [Zet99]

1. next to each other on the hue circle;
2. on the opposite sides of the hue circle (complementary colors);
3. on the tips of an equilateral triangle superimposed on the hue circle;

The hue circle used here is the red-yellow-blue(RYB) hue circle, which is widely used in the artistic domain. Figure 6.3 shows an example of the harmonic color patches. Since we do not want the inserted advertisement intrusive to the customers, the most desirable color patch is achieved when they are neighbors. The distance between two colors c_i and c_j is defined as

$$D_i(j) = \begin{cases} \cos(|c_i - c_j|), & |c_i - c_j| \leq \pi \\ \cos(2\pi - |c_i - c_j|), & \text{else} \end{cases} \quad (6.8)$$

6.4.3 An Optimization-based Solution

In this part, we will discuss the system stabilization in details. Figure 6.5 shows 3 standard cases of graphic mass stabilization. The first one is screen-centered, and it maximizes the stability because only frame magnetism exists in the force field, and the forces in all direction cancel out each other. The

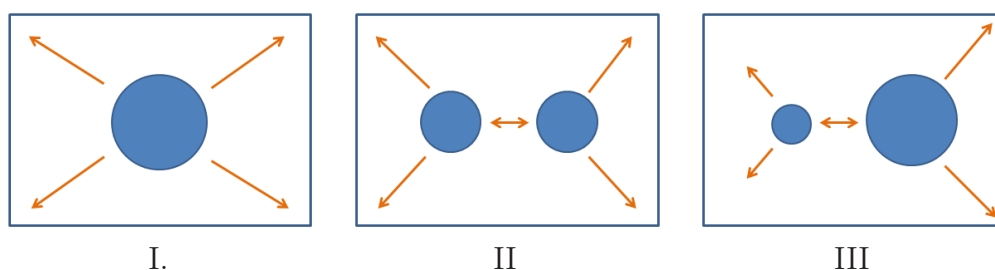


Figure 6.5: Graphic mass and screen position. I. Screen-centered position provides the maximum stability; II. Object-counterweighting can also be balanced if the objects have similar graphic weights; III. The larger and heavier graphic mass on the right surpasses the one on the left, and the system becomes unstable.

second one takes object-counterweighting into consideration. The two objects are of the same graphic weight, so that the system is still balanced. The third one shows an unstable case, where the graphic weight on the right is much heavier than that on the left.

In practice, the first case is relatively rare. But the second one casts light onto our advertisement selection problem. To aesthetically arrange the layout of the screen elements, it is reasonable to assume a force-balanced state. When the force field, including the inner forces and frame magnetism, is balanced, the layout is of high aesthetics.

Since the forces are denoted as vectors, the most straight-forward way to balance all the forces is to add them up. If the sum of the direct forces equals to 0, the system is balanced. However, the case is slightly different in our current case. We want to balance the forces within a predefined web page by inserting an advertisement. Vertices are abstracted from blocks, but blocks are actually informative. Not all the blocks in the system are of the same informative importance. Therefore, adding up all the forces within the system without considering the sources is not reasonable.

To build the web page force balancing strategy, we put up two principles:

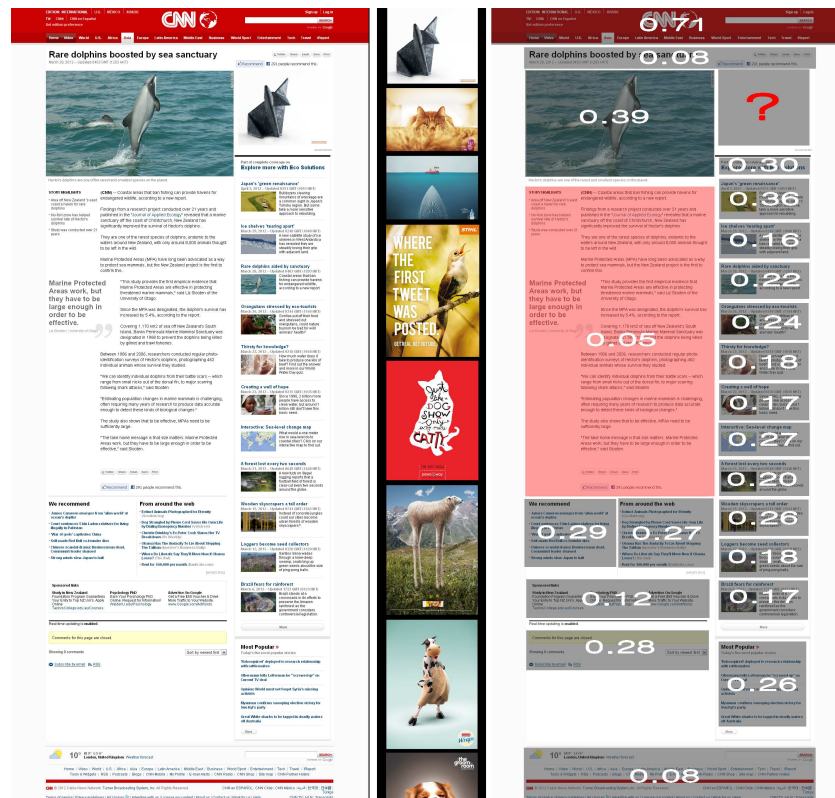


Figure 6.6: Left: Experimental Result 1. A snap shot of CNN news with inserted advertisement. Some of advertisement candidates are listed on the right. Right: The estimated graphic weights of Experimental Result 1

1. The main body of the web page is of the highest importance.
2. To enhance the efficiency of advertising, the selected advertisement should be noticeable to the users but not intrusive.

A web page is made up of the header, the footer, side columns and the main body. The main body contains information that users are most interested in, and it is the reason that a user opens this web page. For example, the main body of a news web page is the textual content, and that of a picture library is the displaying images. Therefore, the block that represents the main body of the web page is selected as the *critical* one in our model. All the group of forces acting on the block is denoted as f_M .

Meanwhile, according to the second criterion, we want the selected advertisement to be eye-catching, so this position is set as another *critical* node, whose corresponding force set is denoted as f_A . Now we can define the *critical* node from the graph point of view. A critical node is a vertices that connects to all the rest vertices in the graph. The edges in f_M and f_S are directed and weight differently if they are reverted. For the two critical node, only the forces on them are considered. And the rest forced are omitted so as to highlight the two elements.

The problem is given as

$$A = \operatorname{argmin}_{\alpha \in S} |f_w(f(\alpha), f(M)) + C_{dist}(\alpha)| \quad (6.9)$$

where

$$\begin{aligned}
f_w(f(\alpha), f(M)) &= \sum_{i \leq N, j \in \{\alpha, M\}} cl_i + If_{ij} + fm_i \\
C_{dist}(\alpha) &= \sum_{i \leq N} D_\alpha(i)
\end{aligned}$$

where S is the set of candidate advertisements that are contextually related to the web page, N is the total number of segmented blocks, function $f_w(\cdot, \cdot)$ is the sum of graphic forces in the graph, function $f(\cdot)$ is the force set of critical nodes, and function $C_{dist}(\cdot)$ is the chroma distance between the advertisement candidate to the web page (Equation 6.8).

The function can be solved by iteratively testing all the advertisement candidates in S . But it is computational expensive when the number of candidates are large. Instead, we solve gw_α directly from Equation 6.9 by mapping forces to x and y directions in the two-dimensional coordinate. And the optimal advertisement A is defined to be the one in the candidate set S that minimizes $|gw_A - gw_\alpha|$.

6.5 Experimental Results

To test the proposed force-based advertising system, we perform 10 sets of experiments. In each experiment, we make use of a real web page which has inserted advertisements. The original advertisements are manually removed and the size and location information serves as constraints in our experiment. Then for each web page, we download 25-30 real advertisements which are contextually related to the page. We run the proposed system to obtain the optimal one that introduces the highest force stability.

Item	Range	Explanation
Effectiveness Evaluation		
Eye-Catching	Y/N	Do you notice the advertisement?
Aesthetics Evaluation		
Intrusiveness	1-5	Is the inserted ad intrusive for your browsing?
Aesthetics	1-5	How do you rate the visual pleasure of the composition of the webpage and advertisement?
Contribution	1-5	How do you evaluate the contribution that the inserted ad makes to the global visual pleasure?

Table 6.3: Evaluation criteria for subjective user study.

The experimental results are shown in Figure 6.6. The web page snapshot is on the left with the selected advertisement inserted. Some of the advertisement candidates are shown on the right column for comparison. Figure 6.6 shows the estimated graphic weights of each block. The block representing the main body is marked in red, and the advertising location is marked by a red question mark. The two critical blocks serve as the visual centers of the web page. To balance the whole force system, the computed optimal graphic weight of the potential advertisement is 0.17, a block of low graphic weight. Then we go back to search the set of advertisement candidates for the one with the closest properties. Visually speaking, the result is intuitive. The main body block is purely textual and has the lowest graphic weight all over the page. Meanwhile, the blocks near it have relatively high graphic weights. The surrounding situation is similar for the advertising candidate. To balance the whole system, the inserted advertisement ought to have low graphic weight to strike towards the system equilibrium.

To evaluate the results of advertisement selection, we invite users to give

scores to the output web page with inserted advertisement. We do not use the traditional classic objective evaluation approaches such as CTR, because we want to rule out other influences, such as webpage content, and focus on the visual effects only. Admittedly, CTR is an important attribute in evaluating the successfulness of advertising schemes, and our experimental designs are based on the assumption that the visual appearance is positively correlated with CTR. This assumption has been justified by the work of Fern’s team [JA12], i.e. visually appealing webpages often have higher user response propensity. The detailed experiments designs are given below:

- Give 10 webpage screen shots. The proposed system generates 10 outputs, and for each page, 4 advertisements are randomly selected from the candidates. Therefore, we have 10 sets of experiments, and each contains 5 samples. The screen shots have been resized to fit the screen by width, and users do not need to draw the horizontal scroll. The height of the webpages, on the other hand, influences the visual importance of advertisement placement, therefore the vertical scroll is left untouched.
- For each set of the experiments, web page screen shots with advertisements inserted are shown to the users one-by-one randomly. Users will score the samples according to the questions (Table 6.3). Noticing that the answer to the first question may be biased if users are asked when they have been shown all the 5 samples in the group. Therefore, our so-called “random” sequence of samples are not really “random”. We label the samples in the group from 1 to 5, where sample 1 is the estimated result while the rest 4 are the random advertisements. We segment users into 10 groups. Advertisements are shown to users in Group 1 by the

sequence of 1 to 10, in Group 2 by the sequence 2,1,3-10, and the like.

The question is asked only at the first sample of each set.

- In order to rule out the erroneous inputs given by the users during the experimentss, transitivity satisfaction rate (TSR) and trust thresholding are considered in the data post-processing.. We assume that users' preference on color and image density has the transitive relation. Therefore, the TSR value is computed as the number of triplets satisfying the transitivity property divided by the number of triplets that the transitivity rule may apply to [CWCL09].

In our experiment, we obtain 30 user responses, and the average scores are given in Table 6.4. According to the study, most of the inserted advertisements are noticeable to the users. Therefore, our scheme of selecting neighboring colors as the optimal color patches is acceptable. Noticing that some users do not notice several randomly inserted advertisements at the first glance (The percentage is 97% for Set 2,5 and 8) because the textures of the inserted advertisement are very similar to that of the target webpages.)

According to the user evaluation, the inserted advertisements given by the proposed system perform relatively better than that those randomly inserted. However, the scores still vary between different experimental sets. For example, the ad in Set 9 gets the highest score of visual pleasure (4.40) while that in Set 2 only gets 3.88. Considering the designs of the advertisements themselves, we assume that advertisements of better artistic design are more welcome to the users, and more users believe that these kind of advertisement could make contribution to the visual pleasure to the webpages (4.24 for Set 9).

		E.C	In.	V.P.	Cnt.			E.C	In.	V.P.	Cnt.
Set 1	P.M	1	4.17	3.92	4.00	Set 6	P.M	1	3.93	4.04	3.96
	R.N	1	3.98	3.44	3.54		R.N	1	3.6	3.4	3.82
Set 2	P.M	1	4.00	3.88	3.92	Set 7	P.M	1	4.04	3.96	4.12
	R.N	0.97	3.42	3.12	3.54		R.N	1	3.44	3.92	3.68
Set 3	P.M	1	4.2	4.08	3.96	Set 8	P.M	1	3.88	3.92	4.04
	R.N	1	3.48	3.98	2.58		R.N	0.97	3.98	3.98	3.66
Set 4	P.M	1	3.92	3.92	4.04	Set 9	P.M	1	4.00	4.40	4.24
	R.N	1	3.12	3.14	3.76		R.N	1	3.52	3.48	3.48
Set 5	P.M	1	4.04	4.12	4.16	Set 10	P.M	1	4.28	4.04	4.12
	R.N	0.97	3.43	3.02	3.72		R.N	1	3.76	3.88	3.68

Table 6.4: User Evaluation. E.C: Eye Catching. In.: Intrusiveness. V.P: Visual pleasure. Cnt: Contribution. P.M: proposed method; R.D: random results.

6.6 Conclusion

In this chapter, we demonstrate an innovative web page advertisement selection strategy based on the force model. It refines the results of contextual advertising by introducing aesthetic criteria. The web page is segmented into semantic blocks, and each block is an element on the two-dimensional screen. Aesthetic theories on the screen balancing are adopted in the proposed system. We compute the graphic weights of blocks and treat them as vertices in a graph. Weighted graph edges are the forces between the elements. The aesthetically optimal advertisement is the one that balances the force system. User study shows visual improvements comparing with randomly selected advertisements from the contextually related candidates.

In practice, contextual relevance and bidding price are two important selec-

tion criteria for advertising. Since our current work is based on the assumption that the advertisement candidates are contextually relate to the publishing website, it is actually a refinement of the relevance selection results. Price bidding, on the other hand, could be considered in our current optimization work 6.9.

Future work includes an objective evaluation model that can be used to evaluate the success of the advertisement selection. Moreover, in the current work, the advertisement candidates are assumed to be contextually related to the given web page beforehand, and the proposed system refines the results of contextual advertising by introducing the visual layout requirements for aesthetics. To make the system self-contained, the two could be integrated for better performance and higher computational efficiency.

Conclusion

In this chapter, we summarize the conclusions of the aesthetic-related media processing framework. In addition, a few potential areas for improvement of these research results will be presented.

7.1 Summary of The Dissertation

This dissertation proposes several media processing approaches based on media aesthetics. The aim is to properly utilize basic aesthetic elements to reveal and improve the visual impact of media on human beings. We argue that computational media aesthetic theories can improve the efficiency of traditional media processing techniques and enhance the output appeal. Single image dehazing based on aesthetics shows how adding aesthetic constraints could simplify the traditional ill-posed problem. The applications of media aesthetic theories raise several research issues such as the relationship between subjective human feelings and aesthetic factors, the interpretation of abstract aesthetic criteria for computational models, and the applications in media processing. The dissertation has tried to address these issues based on the application cases. In the following discussion, we will summarize the specific contributions and findings of our works.

7.1.1 Aesthetics for Single Image

The dissertation begins with the application of the aesthetic theories on a classic image processing problem, single image dehazing. We demonstrate how properly applying aesthetic criteria can dramatically improve the output quality of the under-constrained problem in Chapter 3. The inherent problem of solving the dehazing equation is to place constraints on the properties of the underlying haze-free images. By developing an aesthetics-based visual constraint, we can solve the once ill-posed equations, get the depth map, and obtain a visually pleasant haze-free image with vivid colors. The notion of a “vivid color” is an aesthetic concept that is used to depict the visual effect of an image. The experimental results show that the proposed approach outperforms the dark channel prior [HST09] in dealing with the color quality especially at dark regions.

7.1.2 Aesthetics for Multiple Images

The application of media aesthetic theories on multiple images is discussed after that on single images in Chapter 4. We present an automatic image slideshow authoring system. Traditional image slideshow generation approaches often attach higher importance to the visual information, i.e. the image properties. However, since a slideshow is the integration of image slides and background music, proper incorporation of the two can intrigue the viewers’ visual and auditory feelings and therefore enhance the impact. In our work, we propose a synaesthetic approach for image slideshow generation. The basic aesthetic elements are extracted from both the image set and the input audio clip. We build up a mapping model that integrate visual and

audio features based on aesthetic energy, arrange their sequence, and design the displaying patterns to compose the output slideshow. Experiments show that our proposed system produces better results than the existing methods.

7.1.3 Aesthetics for Videos

In Chapter 5, we present a post-editing scheme for home produced videos. It fuses retargeting and re-projection based on the sequence shot editing techniques, aiming to enhance the aesthetic interest. Various approaches have been proposed in the past for video enhancement, and most works look into the visual quality issues, such as noise, blur, hand-shaking etc. In our work, we consider video enhancement from an aesthetic point of view. We create special effects by introducing artificial camera work and by swiftly adjusting the project velocity. A single shot video is taken as a long take. Media elements related to the long take characteristics are extracted, mainly from the temporal domain. The long take editing techniques are applied to the computational model of the post-processing system. We also propose a perception-based model to evaluate frame interest and a video segmentation algorithm based on frame interestingness.

7.1.4 Aesthetics for Online Advertising

We finally demonstrate an innovative web page advertisement selection strategy based on the force model in Chapter 6. Webpages are not a traditional medium which delivers aesthetics, therefore, we exhibit how the traditional media aesthetic theories can aid the efficiency and quality of modern media experiences. Our proposed system refines the results of contextual advertising by introducing aesthetic criteria. The web page is segmented into semantic

blocks, we compute their visual weights and take these blocks as elements on the two-dimensional screen. Aesthetic theories related-to the screen balancing are adopted in the proposed system. A graph drawing problem is formulated by letting blocks be vertices and forces between edges be the weighted edges. The aesthetically optimal advertisement candidate is the one that balances the force system. User study shows visual improvements comparing with randomly selected advertisements.

7.2 Conclusions

The thesis has developed four novel applications of computational media aesthetics. Here we can draw some conclusions which we have not discussed in details in the first chapter. A summary of the algorithms discussed in the thesis is given in Table 7.1.

- Media aesthetics tries to find out how the basic elements can help to enhance impact. Ideally speaking, if the proposed aesthetic models are correct, they will embed the output results with corresponding aesthetic characteristics. Therefore, the underlying critical issue is how to correctly interpret the aesthetic elements and the corresponding abstract criteria. For example, in Chapter 2 we propose a single image dehazing algorithm based on photographic theories. Color quality can be influenced by their saturation level, and vivid colors are thought to be more visually pleasant. We extract the saturation information as the basic aesthetic element, and propose the computational model to depict the characteristics of “vivid” colors, which is stated as the full saturation assumption. Here the basic element we make use of is color. The aes-

-	Media Elements	Aesthetic Criteria	Computational Model
Dehazing	Saturation, Hue	Vivid colors	Full Saturation Assumption
Slideshow	Audio/Visual Features	The balance between visual and audio information	Mapping based on aesthetic energy
Video Editing	Motion	Proper camera work and projecting speed for long takes	Equalize interestingness spatially and temporally
Advertising	Color, Texture	Visual balance of webpages	Equalize the forces between visual nodes

Table 7.1: A summary of the proposed media aesthetic applications.

thetic theory lies a constraint on the color property, i.e. vivid. We build a computational model to interpret the element, which in return enhance the output impact.

- Computational media aesthetics can simplify the traditional media processing problems. The single image dehazing algorithm based on the full saturation assumption shows how aesthetic theories can be applied to obtain the solution of under-constrained problems, which ensure the visual quality of output results. In other words, computational media aesthetics helps to find a solution which is most visually pleasant under the artistic guidance. And it coincides with the requirements of ordinary users. The advantage of such aesthetic-related constraints lies in the fact that in certain cases (as in our proposed study), it can dramatically improve the computational efficiency (comparing with the

dehazing approach based on FSA and those solving an under-constraint equation) and provide acceptable results as well. Admittedly speaking, these results are not mathematically optimal because it assumes that the underlying degradation-free images are aesthetically pleasant, which is not always true. Therefore, such approaches are more like enhancement algorithms than restoration.

- Computational media aesthetics ensures the visual quality of outputs. The models are based on the studies of basic aesthetic elements. Media aesthetics studies the impact of these elements on human beings, provides guidelines with which we can increase the effectiveness of media aesthetic products, and optimally decide the structure of basic aesthetic elements. Our proposed slideshow authoring, video post-processing and advertisement recommendation systems show the competence of the media aesthetic theories, which can be justified by the positive feedback from user studies on the perceived visual pleasure.
- Computational media aesthetics can optimize the results of traditional algorithms, such as image ranking, retrieval and online advertising. In our advertising scheme, the advertisement candidates are assumed to be contextually related to the targeting webpages, therefore intrinsically speaking they can be taken as the output of traditional content-related advertising schemes. Our system further optimizes the content-related candidate pool by introducing the aesthetic criteria. Research efforts on image ranking and retrieval have also introduced aesthetic constraints to improve the visual quality of outputs. Therefore, based on the traditional media quality prerequisites such as contextual correlation and non-degradation, computational media aesthetics can provide an

additional criterion that further improves the results.

7.2.1 Future Direction

The limitations and potential extensions in computational media aesthetics have been stated in the previous discussions, and we are going to summarize them here.

7.2.1.1 Feature Extraction and Interpretation

One of the most difficult problems of media aesthetics is how to overcome the gap between the human semantic requirements of media products and the information current computers can extract. We have proposed several models to interpret the aesthetic effects of basis media elements, and discussions of related models have been given in the literature survey chapter. An obvious limitation of these aesthetic models is that they are all based on low level features, such as color, luminance, contrast, motion vector, sound frequency, pitch etc. Admittedly, media aesthetics starts from the analysis of media elements, but aesthetic-related elements also include information of higher levels. Take spatial composition as an example. At present, almost all the aesthetic criteria related to spatial compositional evaluation considers the placement rules of foreground objects, such as Rule of Thirds and the Golden ratio. Based on the estimated salient regions - human faces for example - the positions are evaluated to see if the Rule of Thirds is satisfied. However, aesthetic criteria that require object understanding of higher levels are seldom considered. Even if the placement of the detected human face satisfies the Rule of Thirds, the face orientation and foreground interaction can also influence the aesthetic quality [Zet99]. Therefore, in order to achieve better

aesthetic models, the system should be able to utilize higher levels of media features.

7.2.1.2 Dependency on Other Techniques

Various aesthetic models have been proposed to evaluate the aesthetic level of media pieces. However, applying these criteria to enhance the aesthetic quality of a given media piece is still a difficult problem, because successful implementation of these algorithms relies heavily on the accuracy of other techniques such as those related to computer vision or graphics. Take the video editing part in the thesis as an example. There are limitations in the feature extraction stage even for these low-level ones: correct motion vector extraction, foreground detection, and salient region detection. These problems have been studied by specialists in the corresponding computer vision areas for a long time, but are still open problems without a solution ensuring the output accuracy. Camera work is another important feature in differentiating professional video from amateur ones. Motion vectors, optical flow, and the tracking of shift-invariant features have their own advantages but are not universally correct. There is still a long way to go in the analysis of camera work unless a reliable motion estimation algorithm could be proposed. Sometimes it is also acceptable to make the framework semi-automatic and rectify the results of automatic algorithms manually. Anyway, compromises are often made in aesthetic modeling because the availability of corresponding techniques.

7.2.1.3 Objective Evaluation

For both aesthetic assessment and applications, experiments are almost always designed to be based on subjective user evaluation. There is no objective eval-

uation system that could be used to assess the reliability of outputs. The situation might be understandable for aesthetic media assessment. Researchers are still making efforts to build an objective evaluation system for aesthetics. There cannot be an existing assessment system to evaluate the media processing results while researchers working on aesthetic assessment are still struggling to build up such systems. We can not directly apply the systems offered by current automatic aesthetic assessment algorithms to evaluate the processing results, because they are based on the same aesthetic assumptions. For example, when we take the Rule of Thirds as a compositional criterion, images whose foreground arrangement follows the rule are thought to be of higher aesthetic value. Otherwise images that fail to obey the rules will be rectified. By rearranging the object placement, the output image inherently satisfies the requirements of high aesthetic images. We should try to solve the dilemma before we can obtain a reliable and objective evaluation system that makes the studies of computational media aesthetics less subjective.

Bibliography

- [Abb88] Adriano Abbado. Perceptual correspondences of abstract animation and synthetic sound. *Leonardo*, 1:Supplemental Issues, 3–5, 1988. (Cited on page 92.)
- [AdS10] Google AdSense. <http://www.google.com/adsense>, 2010. (Cited on page 154.)
- [AdW10] Google AdWord. <http://www.adwords.google.com>, 2010. (Cited on page 154.)
- [Alt02] Stanley R. Alten. *Audio in Media, sixth edition*. Wadsworth, Ca, 2002. (Cited on pages 88, 89 and 102.)
- [AV03] Brett Adams and Svetha Venkatesh. Weaving stories in digital media: when spielberg makes home movies. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 207 – 210, 2003. (Cited on page 127.)
- [AV05] Brett Adams and Svetha Venkatesh. Injection, detection and repair of aesthetics in home movies. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2005. (Cited on page 56.)
- [AYK06] Radhakrishna S.V. Achanta, Wei-Qi Yan, and Mohan S Kankanhalli. Modeling intent for home video repurposing. *IEEE Multimedia*, 13(1):46–55, 2006. (Cited on pages 21, 28, 58 and 128.)
- [BKMY08] Mathieu Barthet, Recharad Kronland-Martinet, and Sölvi Ystad. *Improving Musical Expressiveness by Time-Varying Brightness*

- Shaping*. Springer-Verlag Berlin, Heidelberg, 2008. (Cited on page 92.)
- [Blo] Dr. Brian Blood. Music theory online : Tempo. Available at <http://www.dolmetsch.com/musictheory5.htm>. (Cited on page 102.)
- [Boy01] Cailin Boyle. *Color Harmony for the Web*. Edition Olms, 2001. (Cited on page 95.)
- [BSS10] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 271–280. ACM, 2010. (Cited on pages 11, 40 and 55.)
- [BSS11] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. A holistic approach to aesthetic enhancement of photographs. *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP)*, pages 21:1–21:21, November 2011. (Cited on page 56.)
- [CH09] Peter Carr and Richard Hartley. Improved single image dehazing using geometry. In *Proceedings of the Digital Image Computing: Techniques and Applications (DICTA)*, pages 103–110, 2009. (Cited on page 65.)
- [CKS10] Peter Ciuha, Bojan Klemenc, and Franc Solina. Visualization of concurrent tones in music with colors. In *Proceedings of the*

- ACM international conference on Multimedia (ACM MM)*, pages 1677–1680, 2010. (Cited on page 92.)
- [CL09] C.D. Cerosaletti and A.C. Loui. Measuring the perceived aesthetic quality of photographic images. In *International Workshop on Quality of Multimedia Experience (QoMEx)*, pages 47–52, july 2009. (Cited on page 53.)
- [CNSF10] Teresa Chambel, Sérgio Neves, Celso Sousa, and Rafael Francisco. Synesthetic video: Hearing colors, seeing sounds. In *ACM MindTrek*, 2010. (Cited on page 92.)
- [CS4] Adobe Premiere Pro CS4. <http://www.adobe.com/products/premiere/>. (Cited on page 128.)
- [Cui02] Thomas Cuifo. Real-time sound/image manipulation and mapping in a performance setting. In *MAXIS Symposium Proceedings*, 2002. (Cited on page 92.)
- [CWCL09] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowdsourcable qoe evaluation framework for multimedia content. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 491–500, 2009. (Cited on page 175.)
- [CYWM03] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Vips: a vision-based page segmentation algorithm. *Microsoft Technical Report*, (MSR-TR-2003-79), 2003. (Cited on page 160.)
- [CZM⁺11] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region de-

- tection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 409–416, june 2011. (Cited on page 134.)
- [Dav10] Bordwell David. *Film art : an introduction 9th ed.* McGraw-Hill Higher Education, New York, 2010. (Cited on pages 28, 124 and 125.)
- [DDN08] Thomas Deselaers, Philippe Dreuw, and Hermann Ney. Pan, zoom, scan - time-coherent, trained automatic video cropping. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. (Cited on page 129.)
- [DeW87] Tom DeWitt. Visual music: Searching for an aesthetic. *Leonardo*, 20(2), 1987. (Cited on page 90.)
- [Dir00] F. Dirfaux. Key frame selection to represent a video. In *International Conference on Image Processing (ICIP)*, pages 275 – 278, 2000. (Cited on page 129.)
- [Div07] Ajay Divakaran. A video-browsing-enhanced personal video recorder. In *International Conference on Image Analysis and Processing Workshops (ICIAPW)*, 2007. (Cited on pages 27 and 29.)
- [DJLW06] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European conference on Computer Vision (ECCV) - Volume Part III*, pages 288–301, 2006. (Cited on pages 1, 35, 37, 38, 39 and 40.)

- [DLW08] R. Datta, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *International Conference on Image Processing (ICIP)*, 2008. (Cited on pages 40, 54 and 155.)
- [DSR89] George Dickie, Richard Scalfani, and Ronald Rabbin. *Aesthetics - A Critical Anthology*. Boston and New York: Bedford/ St. Martin's 2nd Ed, 1989. (Cited on page 3.)
- [DV01] Chitra Dorai and Svetha Venkatesh. Computational media aesthetics: Finding meaning beautiful. *IEEE Multimedia*, 8(4):10–12, 2001. (Cited on pages 4, 6, 10, 12 and 155.)
- [DV02] Chitra Dorai and Svetha Venkatesh. *Media Computing: Computational Media Aesthetics*. Kluwer Academic Publishers, 2002. (Cited on pages 4 and 7.)
- [DVKG⁺00] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2000. (Cited on page 34.)
- [DW10] Ritendra Datta and James Z. Wang. Acquine: aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, pages 421–424, 2010. (Cited on pages xi, 11, 20, 38, 40 and 41.)
- [Ead] P. Eades. A heuristic for graph drawing. *Congressus Nutnerantiunt*, pages 149 – 160. (Cited on page 156.)

- [Eff] The Ken Burns Effect. https://en.wikipedia.org/wiki/Ken_Burns_effect.
(Cited on page 93.)
- [Eva05] Brian Evans. Foundations of a visual music. *Computer Music Journal*, 29(4), 2005. (Cited on page 89.)
- [Fat08] Raanan Fattal. Single image dehazing. In *ACM Special Interest Group on GRAPHics and Interactive Techniques*, pages 72:1–72:9, 2008. (Cited on pages xii, 62, 63, 65, 75, 80 and 81.)
- [FCG02] Jonathan Foote, Matthew Cooper, and Andreas Girgensohn. Creating music videos using automatic media analysis. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 553 – 560, 2002. (Cited on page 30.)
- [FNH08] E. Fedorovskaya, C. Neustaedter, and Wei Hao. Image harmony for consumer images. In *International Conference on Image Processing (ICIP)*, pages 121 –124, 2008. (Cited on page 155.)
- [FR91] Thomas M. J. Fruchterman and Edward M. Reingold. Graph Drawing by Force-directed Placement. *Software - Practice & Experience*, 21, 1991. (Cited on pages 156, 159 and 166.)
- [Gal12] Philip Galanter. Computational aesthetic evaluation: steps towards machine creativity. In *ACM Special Interest Group on GRAPHics and Interactive Techniques*, pages 14:1–14:162, 2012. (Cited on page 35.)
- [GBC⁺00] Andreas Girgensohn, John Boreczky, Patrick Chiu, John Doherty, Jonathan Foote, Gene Golovchinsky, Shingo Uchihashi, ,

- and Lynn Wilcox. A semi-automatic approach to home video editing. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, pages 81 – 89, 2000. (Cited on page [128](#).)
- [GH05] G. R. Greenfield and D. H. House. A palette-driven approach to image color transfer. In *Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 91–99, 2005. (Cited on page [53](#).)
- [GMZ08] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. (Cited on page [129](#).)
- [GW07] Rafael C. Gonzalez and Richar D. Woods. *Digital Imageing Processing (3rd Edition)*. Prentice Hall, 2007. (Cited on page [67](#).)
- [HKP06] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *The Neural Information Processing Systems (NIPS)*, pages 545–552, 2006. (Cited on page [162](#).)
- [HLZ03a] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Ave - automated home video editing. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 490–497, 2003. (Cited on page [128](#).)
- [HLZ03b] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Content based photograph slide show with incidental music. In *IEEE Inter-*

- national Symposium on Circuits and Systems (ISCAS)*, 2003.
(Cited on page 91.)
- [HLZ03c] Xiansheng HUA, Lie Lu, and Hongjiang Zhang. Photo2video. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 92–593, 2003. (Cited on page 91.)
- [HLZ04a] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Optimization-based automated home video editing system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):572 – 583, 2004. (Cited on pages 31 and 128.)
- [HLZ04b] Xiansheng Hua, Lie Lu, and Hongjiang Zhang. Automatically converting photographic series into video. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 708–715, 2004. (Cited on pages 91 and 108.)
- [HLZG03] Junwei Had, Mingjing Li, Hongjiang Zhang, and Lei Guo. Automatic attention object extraction from images. In *International Conference on Image Processing (ICIP)*, pages 403–406, 2003.
(Cited on page 27.)
- [HMH10] Xian-Sheng Hua, Tao Mei, and Alan Hanjalic. *Online Multimedia Advertising: Techniques and Technologies*. IGI Global, 1st edition, 2010. (Cited on page 150.)
- [HST09] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages

- 1956–1963, 2009. (Cited on pages [xii](#), [63](#), [64](#), [65](#), [66](#), [69](#), [79](#), [80](#), [81](#), [98](#) and [180](#).)
- [HZ07] Xiaodi Hou and Liming Zhang. Saliency detection: A spectral residual approach. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. (Cited on page [110](#).)
- [JA12] Yang Zhou Vidhya Navalpakkam Jianchang Mao Xiaoli Fern Javad Azimi, Ruofei Zhang. The impact of visual appearance on user response in online display advertising. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 457–458, 2012. (Cited on pages [157](#) and [174](#).)
- [JI] Wilfried Jentzsch and Hiromi Ishii. Multimedia visual music: Analogical processes between images and sounds. (Cited on pages [89](#) and [93](#).)
- [JLC10] Wei Jiang, A.C. Loui, and C.D. Cerosaletti. Automatic aesthetic value assessment in photographic images. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages 920–925, 2010. (Cited on pages [45](#) and [155](#).)
- [JRW97] D.J. Jobson, Z. Rahman, and G.A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997. (Cited on page [69](#).)
- [KAA⁺02] Masahito Kumano, Yasuo Ariki, Miki Amano, Kuniaki Uehara, Kenji Shunto, and Kiyoshi Tsukada. Video editing support sys-

- tem based on video grammar and content analysis. In *International Conference on Pattern Recognition (ICPR)*, 2002. (Cited on pages 58 and 127.)
- [KCKK00] Jae-Gon Kim, Hyun Sung Chang, Jinwoong Kim, and Hyung-Myung Kim. Efficient camera motion characterization for mpeg video indexing. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages 1171–1173, 2000. (Cited on pages 23 and 24.)
- [KCLU07] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. In *ACM Special Interest Group on GRAPHics and Interactive Techniques*, 2007. (Cited on page 75.)
- [KH00] Changick Kim and Jenq-Neng Hwang. An integrated scheme for object-based video abstraction. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 303–311, 2000. (Cited on page 129.)
- [KKK09] Jin-Hwan Kim, Jun-Seong Kim, and Chang-Su Kim. Image and video retargeting using adaptive scaling function. In *17th European Signal Processing Conference (EUSIPCO)*, 2009. (Cited on page 129.)
- [KN09] L. Kratz and K. Nishino. Factorizing scene albedo and depth from a single foggy image. In *International Conference on Computer Vision (ICCV)*, pages 1701–1708, 2009. (Cited on page 63.)

- [KTJ06a] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 419 – 426, 2006. (Cited on pages 1, 35 and 37.)
- [KTJ06b] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 419–426, 2006. (Cited on pages 1, 37, 38, 54 and 155.)
- [KV12] Shehroz S. Khan and Daniel Vogel. Evaluating visual aesthetics in photographic portraiture. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe)*, pages 55–62, 2012. (Cited on pages 1, 41, 42 and 54.)
- [LBY04] Zune Lee, Jonathan Berger, and Woon Seung Yeo. Mapping sound to image in interactive multimedia art. *Available at <https://ccrma.stanford.edu/zune/sources/papers/papers.files/ccrma2004.pdf>*, 2004. (Cited on page 92.)
- [LC09] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236 –252, 2009. (Cited on pages xi, 37, 42 and 43.)
- [LGLC10] Congcong Li, A. Gallagher, A.C. Loui, and Tsuhan Chen. Aesthetic quality assessment of consumer photos with faces. In *In-*

- ternational Conference on Image Processing (ICIP)*, pages 3221–3224, 2010. (Cited on pages [1](#), [40](#) and [41](#).)
- [LMNN10] Lusong Li, Tao Mei, Xiang Niu, and Chong-Wah Ngo. Page-sense: style-wise web page advertising. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1273–1276, 2010. (Cited on pages [151](#), [153](#), [154](#) and [160](#).)
- [LS07] Cheng-Te Li and Man-Kwan Shan. Emotion-based impressionism slideshow with automatic music accompaniment. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 839–842, 2007. (Cited on pages [91](#) and [92](#).)
- [LT08] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the 10th European Conference on Computer Vision (ECCV): Part III*, pages 386–399, 2008. (Cited on pages [1](#), [44](#), [45](#), [47](#) and [54](#).)
- [LTE08] Olivier Lartillot, Petri Toivainen, and Tuomas Eerola. *A Matlab Toolbox for Music Information Retrieval*. in C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), *Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, 2008. (Cited on page [101](#).)
- [LTY⁺10] Yuanning Li, Yonghong Tian, Jingjing Yang, Ling-Yu Duan, and Wen Gao. Video retargeting with multi-scale trajectory optimization. In *International Conference on Multimedia Information Retrieval*, 2010. (Cited on page [129](#).)

- [LWT11] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *International Conference on Computer Vision (ICCV)*, pages 2206–2213, 2011. (Cited on pages 1, 36, 37, 44, 45, 46 and 54.)
- [Mak] Windows Movie Maker. <http://www.microsoft.com/windowsxp/using/moviemaker/default.mspx>. (Cited on page 128.)
- [McD07] Maura McDonnell. Visual music essay. In *the programme catalogue for the Visual Music Marathon Event*, 2007. (Cited on pages 89 and 92.)
- [MH10] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 83–92, 2010. (Cited on page 108.)
- [MHB08] Eleni Michailidou, Simon Harper, and Sean Bechhofer. Visual complexity and aesthetic perception of web pages. In *Proceedings of the 26th annual ACM international conference on Design of communication (SIGDOC)*, pages 215–224, 2008. (Cited on pages 49 and 156.)
- [MHL08] Tao Mei, Xian-Sheng Hua, and Shipeng Li. Contextual in-image advertising. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 439–448, 2008. (Cited on page 153.)
- [MHYL07] Tao Mei, Xian-Sheng Hua, Linjun Yang, and Shipeng Li. Videosense: towards effective online video advertising. In *Pro-*

- ceedings of the ACM international conference on Multimedia (ACM MM)*, pages 1075–1084, 2007. (Cited on pages 153 and 154.)
- [Mic06] CHRISTEL Michael. Evaluation and user studies with respect to video summarization and browsing. In *Proceedings of the International Society for Optical Engineering (SPIE)*, 2006. (Cited on page 20.)
- [MKYH03] Philippe Mulhem, Mohan S. Kankanhalli, Ji Yi, and Hadi Hassan. Pivot vector space approach for audio-video mixing. *IEEE Multimedia*, 10(2), 2003. (Cited on pages 100, 105 and 122.)
- [MLHL12] Tao Mei, Lusong Li, Xian-Sheng Hua, and Shipeng Li. Image-sense: Towards contextual image advertising. *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP)*, 8(1):6:1–6:18, 2012. (Cited on page 154.)
- [MLZL02] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 533 – 542, 2002. (Cited on pages 20, 24, 25, 26, 29 and 130.)
- [MMK04] Chitra L. Madhwacharyula, Philippe Mulhem, and Mohan S. Kankanhalli. Content based editing of semantic video metadata. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages 33–36, 2004. (Cited on page 127.)
- [MMP12] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *International Con-*

- ference on Computer Vision and Pattern Recognition (CVPR)*, pages 2408 –2415, 2012. (Cited on pages 1, 36 and 37.)
- [MOO10] Anush K. Moorthy, Pere Obrador, and Nuria Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. In *Proceedings of the 11th European conference on Computer vision (ECCV): Part V*, pages 1–14, 2010. (Cited on page 48.)
- [MSVV07] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM*, 54(5), October 2007. (Cited on page 154.)
- [muv09] Video editing software for home movie making: Muvee reveal. <http://www.muvee.com.sg>, 2009. (Cited on pages 88 and 118.)
- [MZZH05] Tao Mei, Cai-Zhi Zhu, He-Qin Zhou, and Xian-Sheng Hua. Spatio-temporal quality assessment for home videos. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 439 – 442, 2005. (Cited on pages 1 and 31.)
- [NL12] Yuzhen Niu and Feng Liu. What makes a professional video? a computational aesthetics approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(7):1037 –1049, 2012. (Cited on pages 1, 47, 48 and 54.)
- [NN00] S.G. Narasimhan and S.K. Nayar. Chromatic framework for vision in bad weather. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 598 –605 vol.1, 2000. (Cited on page 64.)

- [NN03] Srinivasa G. Narasimhan and Shree K. Nayar. Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 2003. (Cited on pages 57 and 64.)
- [NN05] Laszlo Neumann and Attila Neumann. Color style transfer techniques using hue, lightness and saturation histogram matching. In *Computational Aesthetics in Graphics, Visualization and Imaging*, 2005. (Cited on page 57.)
- [OSSO12] Pere Obrador, Michele A. Saad, Poonam Suryanarayan, and Nuria Oliver. Towards category-based aesthetic models of photographs. In *Proceedings of the 18th international conference on Advances in Multimedia Modeling (MMM)*, pages 63–76, 2012. (Cited on page 42.)
- [Pet03] Bryan F. Peterson. *Learning to See Creatively: Design, Color and Composition in Photography*. Amphoto Press, 2003. (Cited on page 18.)
- [PKW08] Tim Pohle, Peter Knees, and Gerhard Widmer. Sound/tracks: Real-time synaesthetic sonification and visualisation of passing landscapes. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 599–608, 2008. (Cited on page 92.)
- [PM03] Dionysios Politis and Dimitrios Margounakis. Determine chromatic index of music. In *Proceedings of the third International*

- Conference on Web Delivering of Music (WEDELMUSIC)*, 2003.
(Cited on page 92.)
- [PMtH07] Dave Payling, Stella Mills, and tim Howle. Hue music: Creating timbral soundscapes from coloured pictures. In *Proceedings of the 13th International Conference on Auditory Display*, 2007. (Cited on page 92.)
- [RAGS01] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34 – 41, 2001. (Cited on page 57.)
- [RSA03] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Multi-operator media retargeting. *Multimedia Systems*, 9(4):353 – 364, 2003. (Cited on page 110.)
- [RSA09] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Multi-operator media retargeting. In *SIGGRAPH*, pages 23:1 – 23:11, 2009. (Cited on page 129.)
- [RSB10] Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. Automatic creation of photo books from stroies in social media. In *Proceedings of second ACM SIGMM workshop on Social media (WSM)*, 2010. (Cited on page 91.)
- [SA07] Y.Y. Schechner and Y. Averbuch. Regularized image recovery in scattering media. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1655 – 1660, 2007. (Cited on pages 63 and 64.)

- [Sax10] Sushil Kumar Saxena. *Aesthetics: Approaches, Concepts and Problems*. Sangeet Natak Akademi and D.K. Printworld Ltd, 2010. (Cited on page 3.)
- [SB10a] K. Seshadrinathan and A.C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, 2010. (Cited on page 34.)
- [SB10b] Nahar Singh and Samit Bhattacharya. A ga-based approach to improve web page aesthetics. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia (IITM)*, pages 29–32, 2010. (Cited on pages 51 and 52.)
- [SBC05] H.R. Sheikh, A.C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005. (Cited on page 35.)
- [SCK⁺11] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H. Hsu, and Shao-Yi Chien. Scenic photo quality assessment with bag of aesthetics-preserving features. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 1213–1216, 2011. (Cited on pages 1, 43 and 44.)
- [SH04] Huang-Chia Shih and Chung-Lin Huang. Detection of the highlights in baseball video program. In *IEEE International Confer-*

- ence on Multimedia & Expo (ICME)*, volume 1, pages 595–598 Vol.1, 2004. (Cited on page 129.)
- [Sha03] Gaurav Sharma. *Digital Color Imaging Handbook*. CRC Press, 2003. (Cited on pages 62 and 69.)
- [SJ00] Bo Schenkman and Fredrik Jonsson. Aesthetics and preferences of web pages. *Behaviour Information Technology*, 19(5):367–377, 2000. (Cited on page 156.)
- [SK00] Xinding Sun and Mohan S. Kankanhalli. Video summarization using r-sequences. *Real-Time Imaging*, 6:449–459, December 2000. (Cited on page 129.)
- [SN00] Bo N. Schenkman and Fredrik U. Jonsson. Aesthetics and preferences of web pages. *Behaviour & Information Technology*, 19:367–377, 2000. (Cited on page 151.)
- [SNS06] S. Shwartz, E. Namer, and Y.Y. Schechner. Blind haze separation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1984 – 1991, 2006. (Cited on pages 62, 63, 64 and 75.)
- [SPJ09] Pinaki Sinha, Hamed Pirsiavash, and Remesh Jain. Personal photo album summarization. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 1131–1132, 2009. (Cited on page 91.)
- [SREB10] Philipp Sandhaus, Mohammad Rabbath, Ilja Erbis, and Susanne Boll. Blog2book: transforming blogs into photo books employing

- aesthetic principles. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 1555–1556, 2010. (Cited on page 58.)
- [Ste94] William J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994. (Cited on page 113.)
- [SXM⁺06] Xi Shao, Changsheng Xu, Namunu C. Maddage, Qi Tian, Mohan S. Kankanhalli, and Jesse S. Jin. Automatic summarization of music videos. *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP)*, 2:127–148, 2006. (Cited on page 129.)
- [SZL09] Lixin Shi, Junxing Zhang, and Min Li. Note recognition of polyphonic music based on timbre model. In *Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 2009. (Cited on page 103.)
- [Tan08] R.T. Tan. Visibility in bad weather from a single image. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. (Cited on pages 63 and 65.)
- [TCKS06] Noam Tractinsky, Avivit Cokhavi, Moti Kirschenbaum, and Tal Sharfi. Evaluating the consistency of immediate aesthetic perceptions of web pages. *Int. J. Hum.-Comput. Stud.*, 64(11):1071–1083, November 2006. (Cited on page 156.)

- [TDG01] P. Tarasewich, H. Z. Daniel, and H. E. Griffin. Aesthetics and web site design. *Quarterly Journal of Electronic Commerce*, 2(1):67 – 81, 2001. (Cited on page 156.)
- [TH93] Annie H. Takeuchi and Stewart H. Hulse. Absolute pitch. *Psychological Bulletin*, 113:No. 2. 345–361, 1993. (Cited on page 101.)
- [Tho95] Cripps Thomas. Historical truth: An interview with ken burns. *American Historical Review*, 100, 1995. (Cited on page 88.)
- [TV07] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP)*, 3, 2007. (Cited on page 129.)
- [Veg] Sony Vegas. <http://www.sonycreativesoftware.com/vegassoftware>. (Cited on page 128.)
- [Vid] Corel VideoStudio. <http://www.corel.com/servlet/SateIite/us/en/Product/1175714228541>. (Cited on page 128.)
- [WBT10] Yaowen Wu, C. Bauckhage, and C. Thureau. The good, the bad, and the ugly: Predicting aesthetic image labels. In *International Conference on Pattern Recognition (ICPR)*, pages 1586 –1589, 2010. (Cited on page 40.)
- [WCLH10] Ou Wu, Yunfei Chen, Bing Li, and Weiming Hu. Learning to evaluate the visual quality of web pages. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1205–1206, 2010. (Cited on pages 1, 50, 51, 54 and 156.)

- [WGCO07] Lior Wolf, Moshe Guttman, and Daniel Cohen-Or. Non-homogeneous content-driven video-retargeting. In *International Conference on Computer Vision (ICCV)*, 2007. (Cited on page 129.)
- [WH06] Yang Wang and Masahito Hirakawa. Video editing based on object movement and camera motion. In *Proceedings of the working conference on Advanced visual interfaces*, pages 108 – 111, 2006. (Cited on pages 27 and 128.)
- [WL09] Lai-Kuan Wong and Kok-Lim Low. Saliency-enhanced image aesthetics class prediction. In *International Conference on Image Processing (ICIP)*, pages 997 –1000, 2009. (Cited on page 45.)
- [Wol96] Wayne Wolf. Key frame selection by motion analysis. In *Proceedings of the Acoustics, Speech, and Signal Processing*, pages 1228–1231, 1996. (Cited on pages 26 and 130.)
- [WRL⁺04] Jun Wang, Marcel J.T. Reinders, Reginald L. Lagendijk, Jasper Lindenberg, and Mohan S. Kankanhalli. Video content representation on tiny devices. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2004. (Cited on page 129.)
- [WWS⁺06] Zhou Wang, Guixing Wu, H.R. Sheikh, E.P. Simoncelli, En-Hui Yang, and A.C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680 –1689, 2006. (Cited on page 34.)
- [XK10a] Yang Yang Xiang and Mohan S. Kankanhalli. Automated aesthetic enhancement of videos. In *Proceedings of the ACM inter-*

- national conference on Multimedia (ACM MM)*, pages 281–290, 2010. (Cited on pages 130 and 138.)
- [XK10b] Yang-Yang Xiang and Mohan S. Kankanhalli. Video retargeting for aesthetic enhancement. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 919–922, 2010. (Cited on pages 112, 130 and 138.)
- [XLY⁺07] Chengkun Xue, Liquun Li, Feng Yang, Patricia Wang, Tao Wang, Yimin Zhang, and Yankui Sun. Automated home video editing: a multi-core solution. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 453–454, 2007. (Cited on page 128.)
- [YK12] Xiang Yangyang and Mohan Kankanhalli. A synaesthetic approach for image slideshow generation. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2012. (Cited on pages 105 and 120.)
- [YKM03] J.C.S. Yu, M.S. Kankanhalli, and P. Mulhen. Semantic video summarization in compressed domain mpeg video. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages III – 329–32 vol.3, 2003. (Cited on page 129.)
- [YLSL07] Junyong You, Guizhong Liu, Li Sun, and Hongliang Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):273 – 285, March 2007. (Cited on pages 18, 20, 21, 23, 24, 26, 27, 28, 30 and 130.)

- [YYC11] Chun-Yu Yang, Hsin-Ho Yeh, and Chu-Song Chen. Video aesthetic quality assessment by combining semantically independent and dependent features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1165–1168, 2011. (Cited on pages [1](#), [47](#), [48](#) and [54](#).)
- [ZCC11] Xiaoyan Zhang, M. Constable, and Kap Luk Chan. Aesthetic enhancement of landscape photographs as informed by paintings across depth layers. In *International Conference on Image Processing (ICIP)*, pages 1113–1116, 2011. (Cited on page [56](#).)
- [ZCJ⁺06] Lei Zhang, Le Chen, Feng Jing, Kefeng Deng, and Wei-Ying Ma. Enjoyphoto: a vertical image search engine for enjoying high-quality photos. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 367–376. ACM, 2006. (Cited on page [37](#).)
- [ZCLR09] Xianjun Sam Zheng, Ishani Chakraborty, James Jeng-Weei Lin, and Robert Rauschenberger. Correlating low-level image statistics with users - rapid aesthetic and affective judgments of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–10, 2009. (Cited on pages [1](#) and [52](#).)
- [Zet99] Herbert Zettl. *Sight, Sound, Motion: Applied Media Aesthetics 3rd*. Wadsworth Publishing Company, 1999. (Cited on pages [1](#), [4](#), [5](#), [7](#), [10](#), [13](#), [21](#), [35](#), [88](#), [90](#), [93](#), [98](#), [101](#), [104](#), [111](#), [116](#), [122](#), [124](#), [125](#), [126](#), [151](#), [158](#), [165](#), [168](#) and [185](#).)

- [ZLY⁺10] Jiawan Zhang, Liang Li, Guoqiang Yang, Yi Zhang, and Jizhou Sun. Local albedo-insensitive single image dehazing. *Visual Computing*, 26(6-8):761–768, 2010. (Cited on pages [xii](#), [63](#), [66](#), [79](#) and [81](#).)
- [ZS06] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 815–824, 2006. (Cited on pages [126](#), [132](#) and [133](#).)
- [ZW07] Jia Zhu and Ye Wang. Pop music beat detection in the huffman coded domain. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2007. (Cited on page [102](#).)
- [ZZXZ09] Kun Zeng, Mingtian Zhao, Caiming Xiong, and Song-Chun Zhu. From image parsing to painterly rendering. *ACM Transactions on Graphics*, 29(1):2:1–2:11, December 2009. (Cited on page [53](#).)