博士論文

Subjective Impression Prediction by Complexity Related Features

（複雑さに関連した特徴を用いた主観的印象予測）

氏名　　孫　理天

Abstract

As the amount of multimedia content shared and consumed online is growing exponentially, it is becoming more and more important to develop a system that is capable to understand and interpret these content in a way that is consistent with human beings. In addition, from the perspective that the best way to understand ourselves is to build a machine that could think similar with us, predicting users' impression towards such content is of great meaning in both psychology and computer science. Our aim in this thesis is to predict human impressions towards photographs and videos through analysing both the content and the viewer under the inspiration of psychological works.

Complexity is considered as an important indicator in cognitive process. And a stimulus with a moderate complexity level leads to the most positive cognitive experience. In this work we predict the viewers' impression towards photographs by analyzing the complexity of the content. In addition, the level of cognitive load caught by the complexity of a stimulus could be measured through the viewer's eye movement. Thus, we predict individual impression towards video lectures using gaze information. However, psychological theories concerning complexity are only verified on limited situations, and the relationship between complexity and viewer's experience on extensive scope of application is not yet clear. To these end, we propose a series of complexity related features, verify the relationship between complexity and viewer impression, predict the subjective impression for both photographs and video lectures.

Firstly, we evaluate the role of complexity played in aesthetic assessment and verify the relationship between complexity and aesthetics on large-scale photographs through computational methods. We designed an experiment to collect human ratings on the complexity of various photos. We proposed a set of visual complexity operators taking reference of the factors used in psychological experiments and extract visual complexity properties of the photograph

from the aspects of composition, shape and distribution. We extract a set of visual complexity features using these operators from various perception cues (VCPC). And we applied gradient boost trees regression on these features to set up the complexity model and showed that the complexity level calculated from the proposed features have a near-monotonic relationship with human beings' beauty expectation on thousands of photos. After that we calculated complexity levels for large-scale photo database, and analysed the relationship between public aesthetics ratings and complexity level.

Secondly, we built up a hierarchical framework to extract structures of different size and intensity contrast, and applied the visual complexity operators to extract the visual complexity features from hierarchical structure (VCHA). We then applied the VCHA features to estimate the aesthetic quality for photographs. There is no standard training and testing protocol for the public aesthetics dataset, so we conducted various experiments under different conditions in order to ensure fair comparisons with state-of-the-art methods. The experimental results demonstrated that the proposed visual complexity features could outperform existing manually prepared features and even better than deep features for balanced training samples. In addition, the proposed features can be extracted directly from samples without tedious learning stage required by deep features.

Thirdly, we use features extracted from gaze information to predict individual rating for video talks. We constructed a dataset of eye movements during video lecture watching together with viewer's rating. Then we proposed a set of gaze features, which not only include the conventional distribution features but also include the analysis of the relationship between visual saliency and gaze point in both static and dynamic aspects. By doing so, we set up a baseline for researches in personal rating prediction for video lectures using gaze.

Acknowledgements

With the utmost gratitude, I would like to express my deepest gratitude to Professor Kiyoharu Aizawa, my advisor, for his persistent support, and enduring guidance over the last few years. His enthusiasm in education and research, his experience and knowledge led me to become a full-fledged person. His inspiring encouragement and insightful advice brought me to success. I feel very fortunate to have him as my supervisor, and the precious experiences will be irreplaceable assets in my life.

I would like to thank Dr. Toshihiko Yamasaki for his illuminating instruction and patience to give me lots of advises in my research work. I have learned lots of skills through his guidance and constant encouragement.

Prof. Katsushi Ikeguchi and Mr. Kohei Onoda provided me precious advices and guidance during my internship in Microsoft Japan (MSD Japan), though unfortunately that work is not included in this thesis.

I owe deep thanks to all the other members in the Aizawa-Yamasaki laboratory for being supportive and sharing opinions and knowledge.

I would also like to show my gratitude to Hirose International Scholarship Foundation for their financial supports.

Finally, I would like to give my special thanks to my parents and husband for their support, understanding and love.

Contents

Abstract **i**

Acknowledgements **iii**

1 Introduction **1**

1.1 Impression estimation: what and why 1

1.2 Impression prediction in computer science 4

1.3 Impression formation as a cognitive process 6

1.4 Objectives 7

1.5 Organization of this thesis 8

2 Complexity and its cognitive indicator **11**

2.1 Complexity and cognitive process 11

2.2 Complexity and aesthetic judgement 13

2.2.1 Challenges in computational aesthetics 17

2.3 Cognitive indicators of complexity 18

2.4 Summary 20

3 Visual complexity from perception cues (VCPC) **21**

3.1 Introduction 21

3.2 Visual complexity operators 22

3.2.1 Composition 23

3.2.2 Shape 25

3.2.3 Distribution 26

3.3 VCPC features 27

3.3.1 Photographs clusters 31
3.4 Summary 32

4 Visual complexity from hierarchical abstraction (VCHA) 41
4.1 Introduction 41
4.2 VCHA features 42
4.3 Visual complexity estimation 47
4.3.1 Dataset construction 47
4.3.2 Related works 49
4.3.3 Experimental results 51
4.4 Summary 52

5 Photo aesthetic quality prediction 55
5.1 Introduction 55
5.2 Related works 56
5.2.1 Complexity in aesthetics estimation 56
5.2.2 Photography-rule-based aesthetics features 57
5.2.3 Content description features 58
5.3 Aesthetics estimation dataset 58
5.4 Relationship between visual complexity and aesthetics 64
5.5 Aesthetic quality prediction by visual complexity features 67
5.5.1 Performance on categorized photos 67
Comparison with complexity features 67
Comparison with other features 71
5.5.2 Comparison with state-of-art aesthetics features on generic photos 72
AVA20K5 74
AVA20Km and AVA230Km 74
AVA40Kh 76

AVA11KC 77
Analysis of VCHA and VCPC features 78
5.6 Summary 81

6 Video lectures personal preference prediction by gaze information 89
6.1 Introduction 89
6.2 Gaze features 91
6.2.1 Gaze statistics features 91
6.2.2 Saliency related features 93
6.3 Experiment setup 95
6.4 Analysis 98
6.4.1 Collected data 98
6.4.2 Framework 98
6.4.3 Rating prediction 99
6.5 Summary 101

7 Conclusions 103
7.1 Summary 103
7.2 Future challenges 106

Bibliography 109

List of Publications 123

List of Figures

1.1 Photographs with rich meaning and varieties. Photographs in the first line are from the challenge "Bad", and the second line from challenge "Enthused". The left two columns show example photographs of high rates and the two columns in the right are of low rates. 5

2.1 (A)Illustration of the original Yerkes-Dodson law [110], (B) The modified version of Yerkes-Dodson law by Hebb [32]. 12

2.2 (A)Illustration of primary reward system and aversion system under varying arousal potential, (B) The inverted-U shape, resulting from the summation of the two curves in (A) [7]. 14

2.3 Information processing model [46]. 16

3.1 From left to right in the first row are Original image, contour, horizontal and vertical flipped contour, the second row: rotation for 30, 45, 60, 90 degree. 25

3.2 The original images are shown at the top and the ellipses on the contours at the bottom. 26

3.3 The original image is shown in the top row, the second row shows the corresponding color histogram, and the third row shows the R_2 reference histogram. The difference between the color histogram and the reference histogram is shown at the bottom as d_1 and d_2. 28

3.4 Examples of perceptual cues: the top line shows the original image, contour, texture, and sharpness from left to right; the bottom line shows the line segments and three layers of the line segments separated by different thresholds from left to right. 29
3.5 Computation of relative color along the contour line. 29
3.6 Sample photographs clusters according to object numbers. 33
3.7 Sample photographs clusters according to histogram divergence features. 34
3.8 Sample photographs clusters according to color histogram features. 35
3.9 Sample photographs clusters according to composition features. . 36
3.10 Sample photographs clusters according to ellipse fitness features. 37
3.11 Sample photographs clusters according to curvature features. . . 38

4.1 Hierarchical abstractions of an image. The intensity scale increases from left to right and the spatial scale increases from top to bottom. 44
4.2 Hierarchical abstractions of three images and the corresponding contour and edge maps. The original image and abstractions at different scales are shown from left to right. The abstraction, edge map, and contour map are shown from top to bottom. . . . 45
4.3 Interface of complexity labelling experiment for the "Animal" category. 48
4.4 Example images of 5 complexity levels labelled by participants. The average complexity levels are rounded to integers, and from left to right they are 1(very simple), 2 (simple), 3 (medium), 4 (complex) and 5 (very complex). Images from the same column share the same averaged complexity level. 48
4.5 Distribution of mean and standard deviation of complexity level labelled for 80 images. 49

5.1 Normalized 50-bin histogram of public aesthetic scores and the fitted normal distribution curve. 59
5.2 Human labelling consensus along the aesthetics score gap. 61
5.3 (A)Histogram of public aesthetic scores for consented samples overlapping all the samples. (B) Histogram of public aesthetic scores for consented samples, which is the dark green area in (A). 62
5.4 Sample photographs with consented aesthetics judgements. CP: consented High quality. CL: consented Low quality. 62
5.5 Sample photographs with dissented aesthetics judgements. DP: Dissented High quality. DL: Dissented Low quality. 63
5.6 Example photos from "Cityscape" category of 5 complexity levels calculated by proposed visual complexity model. The two images from the same column share the same complexity level. . 65
5.7 Relationship between aesthetic experience and complexity level. Distribution of beauty experience along complexity level is represented by box plot in the left. And the difference significance is shown in the right. 69
5.8 Performance comparisons on high/low-quality classification task. 82
5.9 Performance of VCHA features when the aesthetic score gap changed 2δ. The sample numbers in the training and testing set are shown by the bars. 83
5.10 Screenshot of the web page showing the prediction results. 83
5.11 Sample photographs predicted correctly by VCHA for AVA40Kh. H: High quality photographs. L: Low quality photographs. 84
5.12 Sample photographs predicted wrongly by VCHA for AVA40Kh. GH: Ground truth is High quality but predicted as of low quality. GL: Ground truth is Low quality but predicted as of high quality. 85

5.13 Sample photographs predicted wronly by VCHA but correctly by VCPC for AVA11KC. GH: Ground truth is High quality. GL: Ground truth is Low quality. 86
5.14 Sample photographs predicted wronly by VCPC but correctly by VCHA for AVA11KC. GH: Ground truth is High quality. GL: Ground truth is Low quality. 87
5.15 Boxplot of VCHA features for color and hierarchical abstraction layers. 88

6.1 Boxplot of shot length. The numbers below are IDs in TED Talks. 93
6.2 Illustration of the saliency hit feature. 94
6.3 Distribution of public preference. 96
6.4 Illustration of the gaze collection experiment environment. 97
6.5 Example of gaze position for different participants. 99
6.6 Classification accuracy for different features. 100

List of Tables

3.1 OVM representation for contours in Figure 3.1 24
3.2 Summary of the VCPC features 30

4.1 Summary of the VCHA features 46
4.2 Average of standard deviation value for different categories. . . . 49
4.3 Comparison of the regression results for visual complexity features. 52

5.1 Comparison of beauty prediction results 70
5.2 Comparison of categorized classification performances 72
5.3 Comparison of the performance of various aesthetic features and classification models using AVA20K5. 74
5.4 Comparison of the classification accuracy with AVA40Kh. 77
5.5 Performance on AVA11KC 77
5.6 Classification accuracy for AVA40Kh using part of VCHA features. 78
5.7 Top 20 important features in VCHA 80
5.8 Top 20 important features in VCPC 80

6.1 Summary of gaze statistics features 92
6.2 Video information. 97
6.3 Viewers' rating summary. 98

Dedicated to my parents and husband.

Chapter 1

Introduction

With the increasing popularity of digital recording devices and social sharing platforms, the amount of multimedia content accessible on the web is growing exponentially. Such large amount of content together with the communication conveyed through them inspire the need to understand how people consume and appreciate them.

In order to measure, predict and further improve user's experience, it is becoming more and more important to develop a system that is capable to understand and interpret these content in a way that is consistent with human beings. In addition, from the perspective that the best way to understand ourselves is to build a machine that could think similar with us, predicting users' impression towards such content is of great meaning in both psychology and computer science.

1.1 Impression estimation: what and why

Previously only photos taken by professional photographers are adopted and published on magazines, TV programs and advertisements, thus widely appreciated, Nowadays photos taken by ordinary users are uploaded to social

media, shared and appreciated world widely. Everyday there are millions photographs uploaded and appreciated on platforms, such as Flickr[1], Instagram[2], and so on. Amateur users are eager to improve their photography technique and are seeking advices that could help making their photos more popular.

Video content is also thriving on the internet, and the way in which knowledge spreads is also changing. Besides the emerging massive online open courses (MOOC) such as edX[3] and conference talks[4], various how-to instruction videos and lecture videos are available on web, for instance TED Talks[5] and Howcast[6]. Lecturers are no longer facing students in the class room, and the viewer in front of computers could be at anywhere on the earth. Thus, lecturers can not adjust their explanation method by observing students reaction. Collecting, measuring, and understanding viewer's individual preference during watching video lectures are becoming more and more important for tasks like improving the quality of video materials, recommending personalized content, and investigating human cognitive processing.

As described in the previous two scenarios, "Will my work be enjoyed by the audience?" and "how can I make it more popular?" are the essential questions every producer wonders. Besides the automatic feedbacks and advices that the producers are looking for, viewers are expecting to a better sorted repositories on the websites in aspect of aesthetics or general impression. Considering that the great number of images displayed on websites such as Flickr, and videos on Youtube, semantics interpreted from tags or titles are no longer the only criterion for image search and organization. Introducing some kind of appeal measure or preference level into the organization of the samples will improve the usability and user experience a lot.

[1]https://www.flickr.com/
[2]https://www.instagram.com/
[3]http://www.edx.org
[4]http://videolectures.net
[5]https://www.youtube.com/user/TEDtalksDirector/featured
[6]http://www.howcast.com/

When appreciating the online multimedia content, such as a photograph or a video, a viewer may experience many thoughts. At the very beginning, the viewer may have some expectancy towards the content based on him or her choice of the material or from the title or keywords. Then during viewing, the viewer explore the content and try to understand the meaning and feeling that the producer want to convey. Finally, a summary impression is reached and the viewer will conclude in a way whether he or she enjoy the work.

Impression referred in this thesis is defined as the summary preference or judgement from the viewer towards the multimedia content. And the key purpose of impression estimation is to understand and improve user experience.

As for one piece of photograph or a shot of video, different individuals may form divergent impressions due to their differences in aspects of knowledge background, experience, etc. In this way, the individual impression is closely related with the viewer. And the viewer's impression is leaked through, expressed unconsciously, their body movement, facial expression, eye movement and so on during his or her viewing and thinking process. Therefore, such indicators could be used to measure and estimate individual impression, without request of the viewer's personal nor historical information such as knowledge background, mental status.

An general assessment of the material could be reached by averaging various impressions collected from a large group of viewers. Such general assessment is closely related with the content and is a reflection of its quality. By analysing the average impression, we expect to find out the most important factors of the material that influence the viewers' experience. And we could further improve users experience based on such estimation.

Our aim in this thesis is to predict the subjective impressions towards photographs and videos through analysing both the content and the viewer. For

photographs we mainly predict the aesthetic preference by analysing the photograph. And for video lectures, we aim to predict the individual viewer's preference using signals collected from the viewer.

1.2 Impression prediction in computer science

Impression prediction has wide applications in computer intelligence and human computer interaction. For example, in e-learning scenario, analysis of learner's experience and detection of their learning status, whether a learner is bored, frustrated or interested are important for both the learner and the tutor. The lecture could automatically replay the part that the learner did not understood or missed. And tutors could adujst the lecture style and pace according to the feedbacks. Furthermore, guideline for lectures is possible to be extracted from large scale data. Impression prediction could help estimate the viewers' response for advertisement, web pages, architecture and other designs before the products are presented to the real world users. Finding out the elements which interest or frustrate the viewers most is essential to improve users' experience.

Viewers' impressions could be collected through questionnaire or ratings. Collecting all the aspect of subjective impression is very difficult due to the multi-dimension nature of impression. Multiple selections among a set of impressions such as that on the TEDtalk website[7] is one way. Another example is the online photograph challenge[8], in which users could rate the photographs in the range of 1 to 10 according to how much they liked it. The previous one is basic an yes or no selection for each impression. The latter one is a levelled preference rate, which omits the various possible emotions nor the meaning of the content and leads to a summary judgement.

[7] https://www.youtube.com/user/TEDtalksDirector/featured
[8] http://www.dpchallenge.com/

FIGURE 1.1: Photographs with rich meaning and varieties. Photographs in the first line are from the challenge "Bad", and the second line from challenge "Enthused". The left two columns show example photographs of high rates and the two columns in the right are of low rates.

The impression prediction problem could be simplified into predicting the general like or dislike impression of a given material using a classification approach. In the typical strategy, features that represents the properties that influence users' impression are extracted first and machine learning methods are then employed for training based on these features. In such a framework, the bottleneck for accuracy is how well the features can capture the properties that impact on subjective impression of the content.

The difficulty mainly comes from the variety in the materials and the wide semantic range. Thus, viewers' impression are subtle and subjective. Figure 1.1 show some examples selected from online photograph challenges. The first line of Figure 1.1 shows samples from the challenge "Bad", and the second line from challenge "Enthused". The left two columns show example photographs of high rates and the two columns in the right are of low rates.

To answer such question, we need to first find out what factors leads to an impression towards the materials, and psychological works are a good direction to search for help.

1.3 Impression formation as a cognitive process

Impression estimation is a challenging problem because the impression is highly subjective and that its formation is associated with high-level cognitive activities. A final preference impression is the summary of all the perceptual, cognitive, emotional and aesthetic responses to the stimulus.

Complexity is believed to be an important factor influencing user's experience [7, 46]. Simple stimulus leads to boredom, and too complex stimulus that is difficult to understand makes viewers feel confusion and frustrated. Only content of moderate complexity level will lead to the best experience. The general user experience could be computed from the visual complexity property of the stimulus based on the complexity theory.

As for video lecture viewing, more complex cognitive processing is taking place. According to Cognitive Load Theory (CLT) [95], when learning complex tasks, the working memory can only process a limited number of information elements. If the cognitive load associated with a task exceeds the learner's working memory capacity, the learner will have worse performance and in turn the intrinsic aversion system will generate negative feedback about the learning process.

Among many methods to measure learner's cognitive load, gaze information could be collected easily and most non-intrusively. And the measurement of the gaze points are of high accuracy comparing with other indicators such as reaction time which could be influenced by the experiment design and method.

Thus, we analyse viewer's impression from gaze information, following the method in [102].

1.4 Objectives

As illustrated in the previous sections, for photograph, in order to predict its aesthetics value, we aim to apply complexity theory in the public impression prediction. And as for video, we aim to use gaze information as an indicator of viewer's cognitive process to predict individual viewer's impression.

However, psychological theories concerning complexity are only verified on limited situations, and the relationship between complexity and viewer's experience on extensive scope of application is not yet clear. Prediction of individual impression for video lectures is challenging due to the complex cognitive process during viewing. To solve these problems, we explore the cognitive processes during appreciating the photographs and video lectures.

Firstly, we evaluate the role of complexity played in aesthetic assessment and intend to verify the Berlynes inverted-U curve on large-scale photographs through computational methods. We design an experiment to collect human ratings on the complexity of various photos. We proposed a set of visual complexity operators taking reference of the factors used in psychological experiments and extract visual complexity properties of the photograph from the aspects of composition, shape and distribution. We extract a set of visual complexity features using these operators from various perception cues (VCPC). And we applied gradient boost trees regression on these features to set up the complexity model and showed that are consistent with the human perception of complexity the complexity level calculated from the proposed features have a near-monotonic relationship with human beings' beauty expectation on thousands of photos. After that we calculated complexity levels for large-scale

photo database, and analysed the relationship between public aesthetics ratings and complexity level.

Secondly, we built up a hierarchical framework to extract structures of different size and intensity contrast, and applied the visual complexity operators to extract the visual complexity features from hierarchical structure (VCHA). We then applied the VCHA features to estimate the aesthetic quality for photographs. There is no standard training and testing protocol for the public aesthetics dataset, so we conducted various experiments under different conditions in order to ensure fair comparisons with state-of-the-art methods. The experimental results demonstrated that the proposed visual complexity features could outperform existing manually prepared features and even better than deep features for balanced training samples. In addition, the proposed features can be extracted directly from samples without tedious learning stage required by deep features.

Thirdly, we use features extracted from gaze information to predict individual rating for video talks. We constructed a dataset of eye movements during video lecture watching together with viewer's rating. Then we proposed a set of gaze features, which not only include the conventional distribution features but also include the analysis of the relationship between visual saliency and gaze point in both static and dynamic aspects. By doing so, we set up a baseline for researches in personal rating prediction for video lectures using gaze.

1.5 Organization of this thesis

In this thesis, we focus on the issue of subjective impressions prediction.

We first introduce the development of psychology theories and experiments on the role of complexity played in cognitive process. We will introduce the

complexity theory in detail together with its applications and challenges in aesthetics judgement. And the role gaze played as an indicator of the cognitive process will be introduced in Chapter 2.

We will introduce the factors of visual complexity used in psychological experiments together with our visual complexity operators in aspects of composition, shape and distribution in Chapter 3. This chapter also include the proposed visual complexity from perception cues (VCPC) feature derived from the visual complexity operators.

In Chapter 4, we explore the visual complexity concept in another aspect and introduce a hierarchical framework to separate the image into structure with different size the intensity contrast. The proposed visual complexity feature from hierarchical framework (VCHA) will be introduced in this chapter.

We first introduce the aesthetic estimation dataset and the related works in Chapter 5, then we construct a visual complexity dataset to evaluate how the proposed features are capable to model human perception on complexity. Then we compute the visual complexity levels for thousands of photographs belonging to various categories using our visual complexity features and models, and we discuss the relationship between the visual complexity levels and the aesthetic ratings. As for the aesthetic quality estimation, we design various experimental conditions and compare the proposed features with previous works, both the computational aesthetic features and the complexity features.

Chapter 6 is about the prediction of individual viewer impression for video lectures using gaze information. We first carefully selected several video lectures and use them to conduct an experiment to collect eye movements during watching together with the viewer's preference measured as rating values. We develop two categories of features: one is the gaze statistics and the other is visual saliency related features. Gaze statistics features focus on the information that eye movements provides and extract mental status indicators such as

fixation duration, saccades length and etc. On the other hand, saliency related features focus on the difference between the visual saliency and viewer's actual gaze points. Then we applied the two kinds of gaze features to predict individual preference for video lectures. We not only extracted the gaze features from the whole period of viewing but also divided the video lectures into several clips and analysed the features under different clips.

Chapter 7 summarizes the researches conducted in this these and gives the conclusions. The limitations of current complexity features and gaze analysis are discussed, and possible methods to extend the visual complexity features and to improve the gaze analysis are also included.

Chapter 2

Complexity and its cognitive indicator

In this chapter we will review the psychological theories about the influence of arousal, which is mostly evoked by the complexity of the material or difficulty level of the cognitive tasks, in cognitive process and aesthetic judgement. We will also introduce measurements of complexity including both subjective and objective methods.

2.1 Complexity and cognitive process

In the early ages, psychologists have already noticed the influence of the arousal level on performance of cognitive tasks.

The Yerkes-Dodson law [110] proposed at the beginning of twentieth century is an empirical relationship between arousal and performance, which claims that tasks of different difficulty level require corresponding level of arousal in order to achieve optimal performance. For example, difficult task that could not be well done without high concentration and persistance would require higher levels of arousal. Figure 2.1 shows an illustration of such relationship.

Considering the variety and differences between the tasks, the shape of the

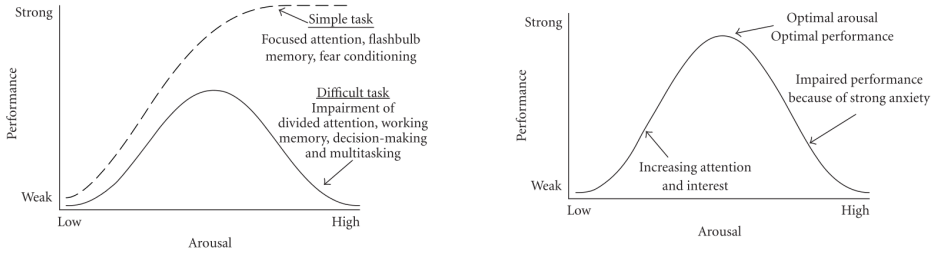


FIGURE 2.1: (A)Illustration of the original Yerkes-Dodson law [110], (B) The modified version of Yerkes-Dodson law by Hebb [32].

curve may be quite different [17]. Linear relationship is often observed in simple or well-learned tasks, while in complex or unfamiliar tasks arousal increase may damage the performance.

A recent review [56] summarized the literature on the effect of stress hormones (glucocorticoids) on human cognition performance. Their work supports Yerkes-Dodson law that memory performance is a inverted-U shaped function of the glucocorticoids level, which means that mildly elevated glucocorticoids levels have the optimal influence in the process of forming long-term memories. In addition, the review reported that subjects exposed to situations that are novel, unpredictable, uncontrollable by the subject, or social evaluative threat would experience higher levels of stress [42, 62].

The latter Cognitive Load Theory (CLT) [95] proposed has the similar idea with the Yerkes-Dodson law, but focus on problem solving process together with the mental effort required by the task rather than measure the complexity or difficulty of tasks in term of stress. CLT suggests that human beings' working memory has limited capacity. And tasks that require a relative large amount of cognitive process capacity may be difficult to form long-term memory or skill, which is called schema construction. CLT further pointed out that alternative instructional materials which do not involve problem solving are helpful in reducing cognitive load. Examples of alternative instructional materials are

worked-examples and goal-free problems.

Besides the subjective measurement self-reporting questionnaire, pupillary responses are widely used as indicator for mental workload [74], which is more convenient and could be measured real-time compared with the stress hormones.

2.2 Complexity and aesthetic judgement

As aesthetics judgement belongs to the cognitive process, psychologists also developed complexity related theories to explain how people achieve an aesthetic judgement [9, 7, 6, 46]. In early empirical aesthetics, several predictive formulae were proposed using complexity and order [9, 20]. According to Birkhoff [9], "aesthetic measures" could be computed from complexity of an object and associated order or symmetry. And complexity was the amount of effort in processing the stimulus, and such an effort provokes the experience of aesthetic reward. He proposed to measure complexity by counting the edges and vertices of polygon. Birkhoff's idea is purely empirical and lacks convincing results in applications. However his work started a new direction in aesthetics research and inspired further researches concerning complexity.

Berlyne first provided a psycho-biological explanation for the role of complexity in preference. In his model of the relationship between complexity and aesthetic preference [7, 6], Berlyne suggested that the aesthetic appeal of a stimulus is related to the viewer's arousal potential, which is stimulated by three types of variables: psychophysical (e.g., brightness, saturation, and the predominant wavelength), ecological (elements associated with biological events, e.g., innate or learned signal values and meaningfulness), and collative (complexity, novelty, uncertainty, conflict, and unfamiliarity). Berlyne claimed that collative variables are the most important and that the preference for an image

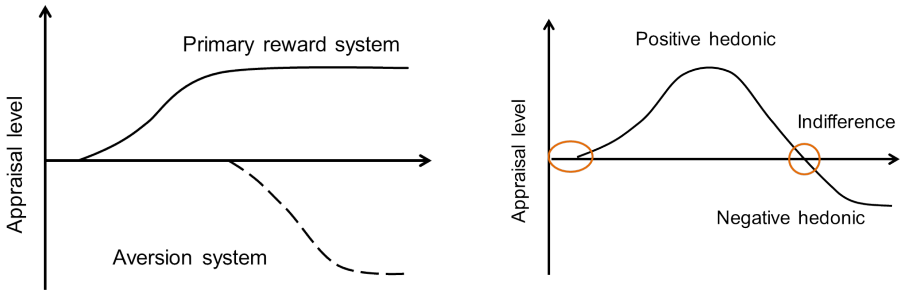


FIGURE 2.2: (A)Illustration of primary reward system and aversion system under varying arousal potential, (B) The inverted-U shape, resulting from the summation of the two curves in (A) [7].

depends primarily on the apparent complexity to the viewer. The perceived complexity was defined according to aspects such as the regularity of shape and arrangements, amount of elements, heterogeneity of elements, and incongruity (typicality).

An increase in arousal potential first activates the primary reward system, and when it reaches a certain level, the aversion system will be triggered and increase along with the arousal potential. When arousal potential is further increased, the aversion level will finally exceed the reward level and generate a negative hedonic value, as shown in the left of Figure 2.2. The right of Figure 2.2 shows the summation curve of the reward and aversion system. In this way, the aesthetic appeal increases with complexity until an optimal level of arousal is reached, and further increases in complexity after this point will elicit a decline in the viewer's level of appreciation. The curve in the right of Figure 2.2 is also called as inverted-U curve.

Berlyne's theory has been applied in many fields, such as architecture [68, 1, 71], music [5], marketing environments [106, 4], poetry [90], and webpage design [87, 100].

Further aesthetic theories have been proposed based on Berlyne's hypothesis. In more recent Leder's information processing model of aesthetic experience [46], to achieve an aesthetic judgement, information is processed repeatedly through several stages: perceptual analysis, implicit memory integration,

explicit classification, cognitive mastering and evaluation, as well as a continuously ongoing emotional evaluation, as shown in Figure 2.3. And complexity is one of the most influential factors in the first stage of perceptual analyses. Besides, the familiarity and prototypicality in the second stage are believed to be related with complexity and will influence human perception on complexity [59, 97]. Patterns that are more consistent with the prototype in our mind will be recognized more easily, thus considered as more simple. According to Leder's model [46], the perception of complexity is correlated with the amount of groupings a user performs unconsciously, connectedness, symmetry, and other factors.

Leder's model took reference of Berlyne's theory about arousal potential and pointed out that complexity plays an important role in the very beginning of the aesthetic experience. Furthermore, the model integrated both cognitive and affective processes involved in aesthetic judgement formation and accommodate a large body of findings on the cognitive foundation of aesthetic judgement. Different from previous theories that emphasized on single factor that determine aesthetic experience, Leder's model put the focus on the interaction among cognitive and affective processes and showed that the variety of aesthetic experience is rooted in how differently the various information can be associated, combined and absorbed.

The role that complexity plays in aesthetic preference prediction is also emphasized in the processing fluency theory [79, 78] which goes further to explore the reason behind the relationship. It suggests that aesthetic experience is a function of the perceiver's processing dynamics: the more fluently the perceiver can process an image, the more positive is their aesthetic response. Fluency theory works well in predicting aesthetic effects due to many low-level features such as preferences for larger and more highly contrastive displays. However, fluency theory does not square well with the Berlyne's inverted-U

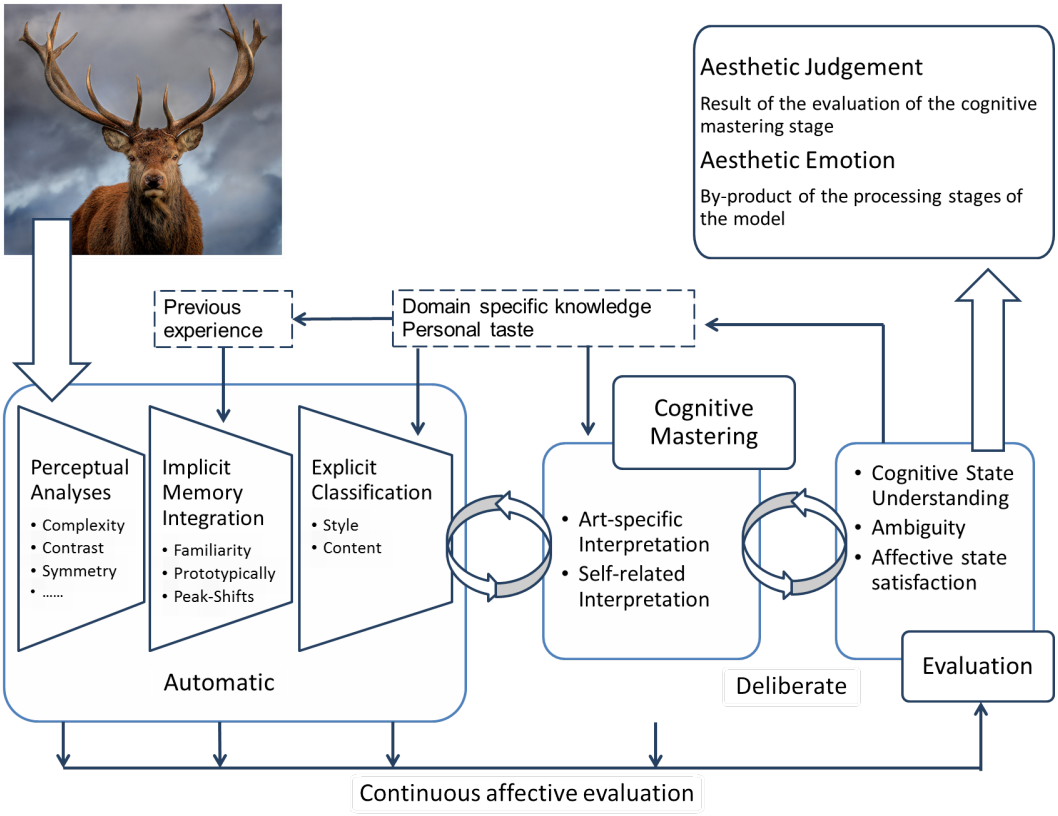


FIGURE 2.3: Information processing model [46].

results, in that it indicates a monotonic decrease in preference as a function of complexity.

Despite the various theories that interpret the mechanism of subjective impression using complexity related factors and the experimental efforts to verify these theories, the relationship between arousal (complexity) and valence (aesthetic pleasure) are still debatable, due to the unveiling nature and underlying architecture of affect system. In some models of affect valence and arousal are defined as independent to each otehr and used as the fundamental axes of more subtle emotions such as depression, excitement, contentment and distress [85, 81]. Such independence relationship is widely assumed in experiments that collect users experience through questionnaires.

In addition, the applicable range of the complexity related theories is ambiguous, considering the wide range of contents. For example artworks belonging to different genres have totally different appearance and evoke completely different feelings in human minds. And daily photographs could be quite different from still life photographs, not to mention the gap between photographs and other forms of paints, such as cartoons, sketch, oil painting. It is even intuitive that the relationship between complexity or arousal level with aesthetic pleasure may vary according to different type of content or style. Thus it is necessary to verify the role of complexity played in aesthetic judgement in a specific range.

2.2.1 Challenges in computational aesthetics

Although complexity is regarded as an important indicator for aesthetic assessment, in computational aesthetics field, complexity theory has not yet been paid enough attentions due to the following challenges:

- What factors contribute to the visual complexity has been a difficult question in order to apply the complexity theory in aesthetic estimation. Basically, visual complexity consist of three factors: the number of elements included in the scene, the differences between the elements, and the arrangement of the patterns. Besides, the application scope of psychology theories is not clear, as complexity may vary greatly for intra and extra-category images.

- The relationship between complexity and aesthetic appeal is still debatable and further verification is necessary. Psychologists have conducted a lot of experiments to verify or evaluate Berlyne's theory. Many have successfully observed an inverted-U function between complexity and

aesthetic experience concerning architecture [68, 1], while some only observes the ascending part of the curve [31], and some shows no support for an inverted U relation between preference and entropy [91].

- The main difficulty in psychological experiments is the limitation of sample size, leading to the problem of insufficient complexity range. Empirical experiments would become time-costing for participants when the sample size numbered in thousands. Thus empirical experiments could not yield a general guideline for aesthetic assessment.

2.3 Cognitive indicators of complexity

There are several methods to measure the multidimensional construct of cognitive load. Subjective methods are self-reporting ways, for example, interviewing the learner for their feeling [95], collecting the difficulty ratings [103]. Objective methods measure the learner's effort either by assistant task or by collecting physical indicators. A questionnaire can evaluate how much of the content the learner's understood. And the reaction time in a parallel secondary task can reveal the spare working memory [24, 95]. Pupillary response also varies along the task difficulty [102].

EEG [111] and fMRI [84] are widely used to monitor and predict internal brain dynamics status. Comparing with EEG and fMRI, eye movement is believed to be a cognitive indicator that could be non-invasively collected with low cost and compact devices. Gaze information is used to infer visualization task and user cognitive abilities in [92], and gaze pattern would differ in decision-making task and a search task as reported in [25].

Recently, gaze is applied as an indicator in more subtle cognitive processes, such as personal preference profiling and artwork appreciation. Gaze aversion is found to be related with high cognitive load [18] and used in lie detection

[104] and other cognitive process exploration [60]. The work in [108] uses eye tracking results to build a recommendation system for online multimedia materials. As suggested in [11], viewers' eye movements during the process of appreciating artworks are closely related with the attractiveness of the artworks. And gaze features are proved to be accurate to predict viewer's preference for images under controlled viewing environment by [94]. In [77], eye tracking method is applied to explore how the viewers' previous training is related to their aesthetic viewing in various interactions. Gaze is also used to infer user's latent interest [76] and language expertise [43]. All of these works suggest that gaze information could serve as an efficient indicator of cognitive process, including viewer's knowledge background, previous experience, mental status, emotion, and etc. In this work we step further to deal with individual rating for on-line video lectures by using gaze.

Conventional gaze features refer to distribution of the two event types of gaze, fixation and saccades. Features such as fixation count and duration, saccades velocity and angle, and etc., are widely used to characterize viewer's cognitive process [92, 107, 45]. Considering the lack of the analysis of the original content in the distribution features, more works use pre-defined areas of interest (AOIs) derived from the content, and convert gaze points to the sequence of AOI hits. For example, in [11] image is divided into two or three AOIs, and the transition entropy is calculated from the gaze point shifts between AOIs. The dynamic video stimuli make the AOI identification much more challenging. One method would be to determine the AOIs by clustering gaze points as described in [44].

2.4 Summary

In this chapter, we introduced psychological works on the important role of complexity and arousal potential in cognitive process and in aesthetic judgement. The Yerkes-Dodson law states the inverted-U shape in the relationship between arousal level and cognitive performance. And cognitive load theory show that instructional design could help reduce mental workload which is could be measured through the task-evoked pupillary responses. In the field of aesthetics experience, early theory proposed by Berlyne is similar to the Yerkes-Dodson law which shows that middle-level arousal potential leads to the best aesthetic experience. Complexity is also viewed as a important factor in more recent Leder's model.

We also introduced measurements of complexity and arousal level including both subjective and objective methods used in psychological works. Subjective measurements could be self-reporting questionnaires. Objective measurements include hormones analysis, EGG, gaze and etc.

In addition, the challenges in applying complexity theories in impression prediction are discussed.

Chapter 3

Visual complexity from perception cues (VCPC)

3.1 Introduction

Pioneers in computational aesthetics as D. E. Berlyne [7, 6] suggested that the aesthetic appeal of a pattern seems to depend on the arousing and de-arousing influence of its collative or structural properties, and that arousing quality is a direct linear function of complexity, or the amount of information, whereas pleasantness is generally related to these determinants in an inverted-U manner. Specifically, aesthetic appeals increase with complexity until an optimal level of arousal is reached, and after this point, further increase in complexity would elicit a drop in preference level.

Visual complexity can be defined as the difficulty in describe or reproduce the photograph [73, 30, 28]. This definition indicates that the amount and variety of the elements in the scene contribute to the complexity level.

Many experiments have been conducted to test complexity theory by employing diverse visual stimuli. Studies using artificially generated patterns aim to control the visual complexity via factors such as the number of turns [12] and density of texture [35]. The visual complexity of artworks [66], architecture [68], and portraits [86] can be determined by the normalized scores collected from

participants. Based on these experiments, various factors were identified as relevant to the human perception of complexity, including the following three basic types.

- **Composition** - how the elements are organized. Structural variables such as the spatial organization of objects and contours, connectedness, and symmetry have been identified as influential in the perception of complexity in a scene [73, 34, 67].

- **Shape** - the shapes of elements and heterogeneity in the appearance of the elements. Quantitative variables such as the number of elements, the number of turns, the amount of contours, and the concentrations of elements were used to control the complexity level in artificially generated patterns by [12, 64, 67]. Other indicators such as the compactness, outer contour length, and randomness have also been shown to be important [37, 22]. Specific patterns that depict real objects or people are recognized as more complex than abstract ones [23], and curvature is preferred over angularity [75]. The regularity (entropy), roughness, and density of textures are relevant to complexity [27].

- **Distribution** - variations according to different aspects of information theory. Variations in color [73, 33] and directionality [27, 33] are related to the perception of complexity, and they can be measured using methods from information theory, such as entropy and histograms.

3.2 Visual complexity operators

We extract visual complexity in terms of the composition, shape, and distribution aspects as indicators verified by psychological experiments, as illustrated

in Section 3.1. We propose two frameworks for summarizing the visual complexity of features as descriptors of an input. The first extracts VCPC and the other extracts VCHA. In VCPC, shape complexity features are extracted from the contours and textures. Composition and distribution complexity features are applied to various basic human perceptual cues, including the intensity contrast, color, and sharpness. In VCHA, visual complexity features are extracted from the hierarchical structures in the image. We divide the image into abstractions at different scales and we then extract the complexity features separately from these abstractions. At the most abstract scale, only the most important objects are preserved, whereas smaller objects and more details are included at the least abstract scale. To generate the abstraction of an image, we first use a rolling guidance filter [114] to blur the image, which can also preserve the edges to some extent, before applying Sobel and Canny filters to extract the edges and contours. In this manner, the objects in the input image are separated according to their size and intensity contrast relative to the background. Color, intensity contrast, and contour are used considering that texture information is included in the less abstract layer and that sharpness is included in the difference between abstractions.

3.2.1 Composition

The composition is calculated using the orthogonal variant moments (OVM) method proposed by [61], which is designed to be sensitive to specific perturbations such as transformation, as well as being tolerant to a certain amount of unexpected disturbance. For a specific perception $P(x, y)$, such as an edge, derived from an image $I(x, y)$, OVM generates a five-dimensional vector: $f_{ovm} = (A, L_x, L_y, D_x, D_y)$, where A is the spatially accumulated value of the input, or surface area, (L_x) and (L_y) are orthogonal components of the surface area, and

TABLE 3.1: OVM representation for contours in Figure 3.1

Input	A	L_x	L_y	D_x	D_y
Original contour	17.09	18.53	21.86	1069.32	2278.31
Horizontally flipped	17.09	18.53	21.86	1810.68	2278.31
Vertically flipped	17.09	18.53	21.86	1069.32	1665.69
30^o rotated	8.98	11.42	12.23	1585.09	3559.43
45^o rotated	8.42	11.04	11.25	1677.05	3834.09
60^o rotated	8.99	12.03	11.66	1663.32	3879.86
90^o rotated	17.09	21.86	18.53	1281.55	3218.98

(D_x) and (D_y) represent the position of the object in the image. The detailed calculation process is as follows.

$$\eta = \frac{1}{height \times width} \qquad A = \eta \int\int P(x,y)dxdy \tag{3.1a}$$

$$L_x = \eta \int\int \sqrt{1 + \left(\frac{\partial P}{\partial x}\right)^2} dxdy \qquad L_y = \eta \int\int \sqrt{1 + \left(\frac{\partial P}{\partial y}\right)^2} dxdy \tag{3.1b}$$

$$D_x = \eta \int\int (x + \frac{\partial P}{\partial x}) P(x,y)dxdy \qquad D_y = \eta \int\int (y + \frac{\partial P}{\partial y}) P(x,y)dxdy \tag{3.1c}$$

The contour of an image is shown in Figure 3.1, together with several transformations. The OVM representations of these contours are listed in Table 3.1. The surface area A are all the same for the original contour, horizontally and vertically flipped and 90^o rotated contours. L_x and L_y are influenced by the rotation. The value of L_x and L_y are almost the same for 45^o rotation and are swapped when rotated by 90^o. D_x and D_y reflects the position, thus the original contour shares the same D_y with its horizontally flipped counterpart, the same D_x with its vertically flipped counterpart.

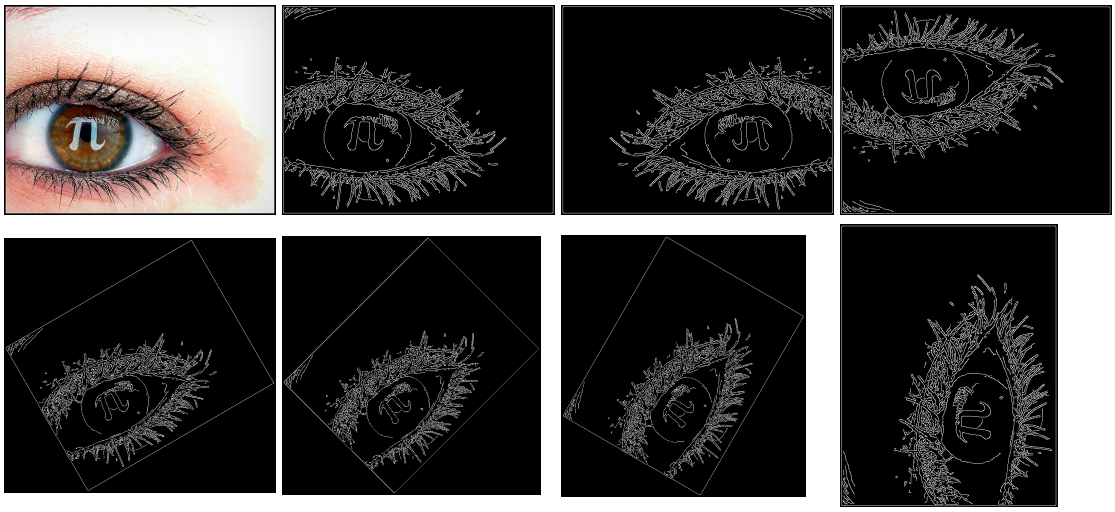


FIGURE 3.1: From left to right in the first row are Original image, contour, horizontal and vertical flipped contour, the second row: rotation for 30, 45, 60, 90 degree.

3.2.2 Shape

We use contour information to represent the shapes of the elements in the input. We count the number of separate contours to approximate the elements and we measure the variety of the elements by using the mean and standard deviation of certain indicators extracted from a contour. First, we fit each separate contour line to an ellipse, as shown in the second row of Figure 3.2. Five indicators are calculated for each contour line: direction, circularity, curve degree, area, and solidity. The direction of the element is measured by the angle of the fitted ellipse. The circularity is represented by the relative ratio of the minor and major axes of the ellipses. The curve degree is measured as the ratio of the contour length relative to the perimeter of its minimum enclosing rectangle. The area of the region surrounded by the contour is calculated and divided by the image area. The solidity is the relative ratio between the contour area and the area of its convex hull. The parameters of the contour lines in the image are then summarized as mean and standard deviation values.

We also extract an eight-dimensional curvature feature from the contour using the method proposed by [29].

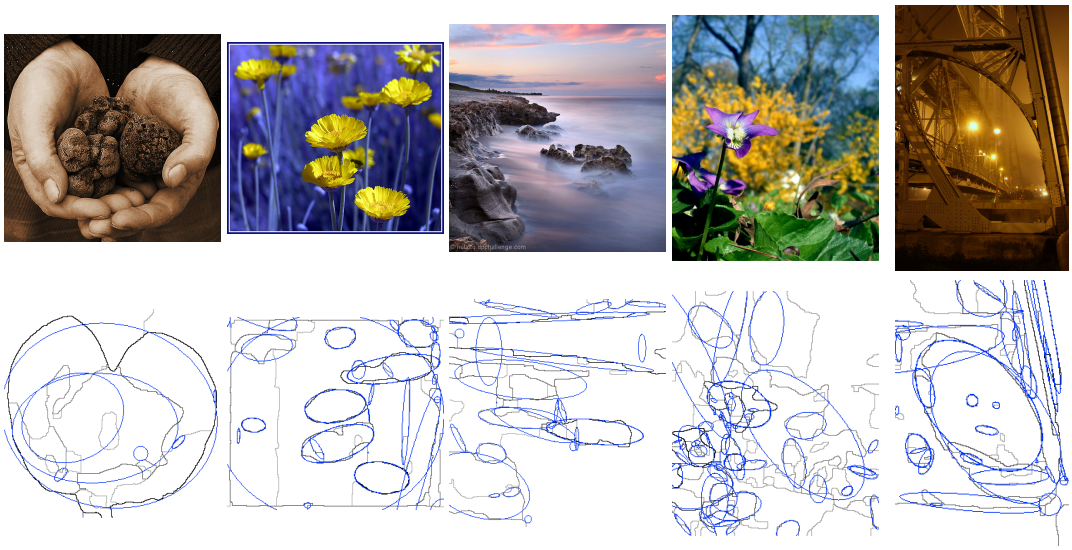


FIGURE 3.2: The original images are shown at the top and the ellipses on the contours at the bottom.

The granularity and regularity of textures are measured using the area statistics and entropy.

3.2.3 Distribution

We calculate the distribution information as the combination of the histogram H and its differences D from histograms of the templates $f_{db} = [H, D]$.

For a specific perception P, such as color, the values of all the pixels in P are accumulated and normalized into a histogram with n bins, $H = [h_1, h_2, ..., h_n]$. The differences between the histogram and those of the reference perception, R, are measured using chi-squared divergence. We select two reference histograms: one has the lowest entropy with an averaged distribution in the histogram and the other has the highest entropy, where only one bin has a value of 1 and all the other bins have values of 0. The differences between the two reference histograms characterize the irregular or regular degree of the distribution of perception P. The detailed calculation is shown by Equation 3.2:

$$D = [d_1, d_2], \quad d_i = \sum_{j=1}^{n} (\frac{h_j}{r_{i,j}} - 1)^2 h_j \tag{3.2a}$$

$$R_1 = [r_{1,1}, r_{1,2}, ..., r_{1,n}], \quad r_{1,i} = \frac{1}{n} \tag{3.2b}$$

$$R_2 = [r_{2,1}, r_{2,2}, ..., r_{2,n}], \quad r_{2,i} = \begin{cases} 0, \text{ if } i \neq m \\ 1, \text{ if } i = m \end{cases}, \text{where } m = [\frac{\sum_{j=1}^{n} h_j * j}{\sum_{j=1}^{n} h_j}]. \tag{3.2c}$$

Figure 3.3 shows some example images and their color histogram. The color of the image is first transformed to CIECAM02 space, and the histogram is calculated from the hue composition, which ranges from 0 to 400. The reference histograms R_1 are all the same for any input, so we only show the R_2 reference histograms. The difference $D = [d_1, d_2]$ between the color histogram and the reference histogram, R_1 and R_2, is listed at the bottom. As the complexity of the color distribution increases from left to right, d_1 decreases, thereby indicating that the color of the input approaches a random distribution. A smaller d_2 indicates that the colors in the input are simpler compared with reference R_2 in the third row.

3.3 VCPC features

Shape complexity features are extracted from contours and textures. The composition and distribution of the complexity features are then applied to various basic human perceptual cues, including intensity contrast, color, and sharpness.

We compute the line segments, contours, and textures using the method proposed by [2], which segments the image hierarchically and sets the parameters by supervised learning. We calculate the sharpness using the method described by [105], which considers both the spectral and spatial properties of the image. An example of the preprocessed results is shown in Figure 3.4. The

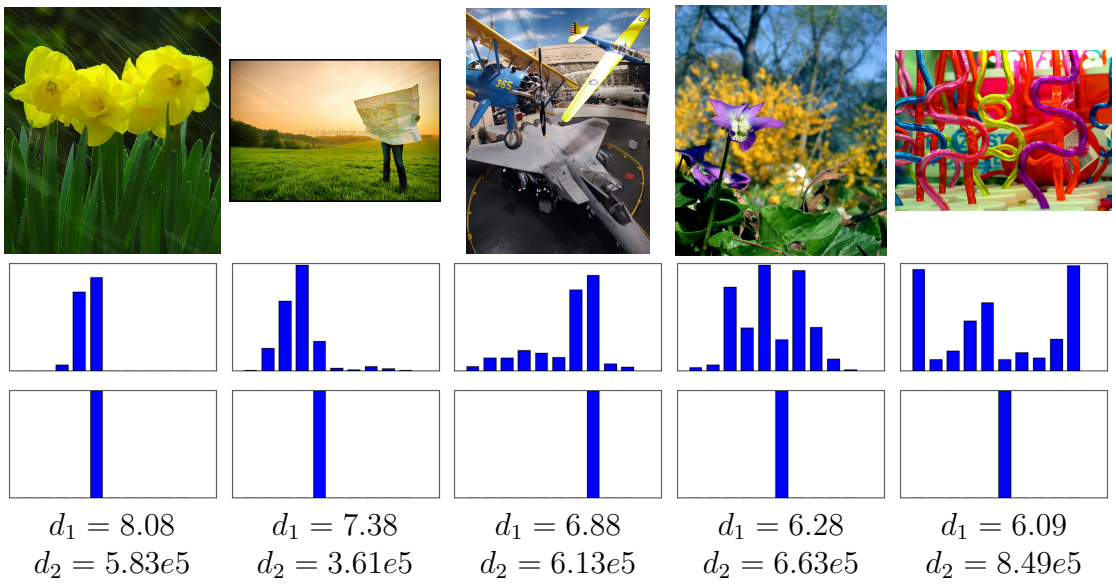


FIGURE 3.3: The original image is shown in the top row, the second row shows the corresponding color histogram, and the third row shows the R_2 reference histogram. The difference between the color histogram and the reference histogram is shown at the bottom as d_1 and d_2.

color information for the original image is transformed into the CIECAM02 color space and then divided into hue (including the hue angle, hue eccentricity, and hue composition), chroma, and lightness.

To reflect the combination of colors, we also prepare relative colors in the regions around the contours. For each circular region with a center point on the contour lines, the main relative color is calculated as the difference between the most dominant and second most dominant colors. The hue and chrome are also extracted from the relative color. The contour lines are downsampled to improve the computational efficiency. Figure 3.5 shows such computation process. The first and second dominating color are decided as in the left of the figure and in the right we show the downsampling process along the contour line.

We employ the composition complexity features of the line segments, color, sharpness, and relative color information. To distinguish the objects with different importance in the image, we split the edge map generated according

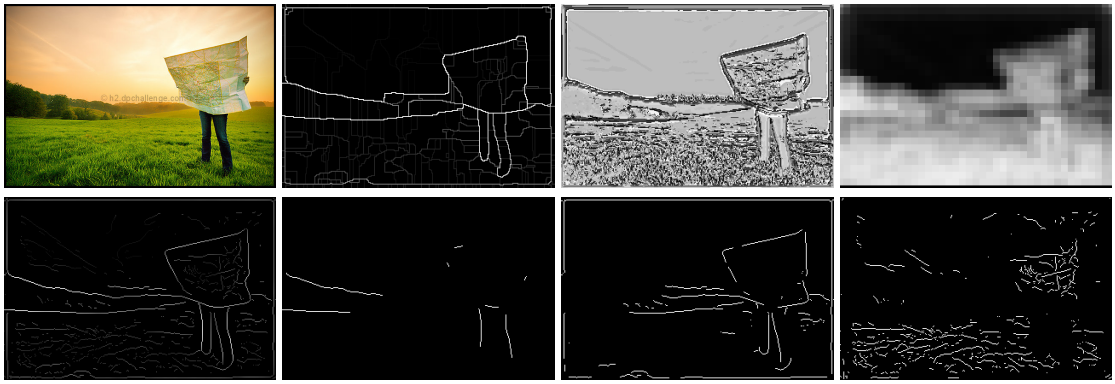


FIGURE 3.4: Examples of perceptual cues: the top line shows the original image, contour, texture, and sharpness from left to right; the bottom line shows the line segments and three layers of the line segments separated by different thresholds from left to right.

FIGURE 3.5: Computation of relative color along the contour line.

TABLE 3.2: Summary of the VCPC features

Category	Perception	Details	Dimension
Composition	Line segments	Line segments and three layers	20
	Color	Hue (angle, eccentricity, and composition), chroma, and lightness	25
	Sharpness		5
	Relative color	Relative hue and relative chroma	10
Shape	Contour	Ellipse fitness	10
		Object number	1
		Curvature	8
	Texton	Texture entropy	1
		Texture area	2
Distribution	Line segments	Orientation	8+2
	Texture	Orientation	$8+2$
	Color		10+2
Total			114

to [2] into three layers using different thresholds, as shown at the bottom of Figure 3.4.

To calculate the distribution complexity, we divide the orientations of the line segments and textures into eight bins, and the color into 10 bins. Thus, by adding the two dimensions for the difference between the histogram of the input perception and the reference histogram, we can obtain the visual complexity in terms of the distribution. A VCPC feature with a total of 114 dimensions is summarized in Table 3.2.

3.3.1 Photographs clusters

In the following, we show what kind of images are similar to each other when defined by our features. We use k-means clustering algorithm to divide a set of 20K photographs into 10 clusters. We show five groups of photographs. In each group, we select the 10 photographs that are closest to the cluster center. In order to put the images into columns and rows, We resized the images to 4:3 ratio, and only show the horizontal photographs although the composition features are more distingtive between horizontal and vertical photographs.

Figure 3.6 and 3.7 show photograph clusters generated from object number and histogram divergence separatly. These two features are quite intuitive in complexity property extraction. As object number is one-dimension feature, we set some thresholds. Group A in Figrue 3.6 has photographs with less than 5 objects. And the other thresholds are 20,50,100,300. Photographs in group E have more than 100 but less than 300 objects. Photographs with larger number of object than 300 are not shown. From such grouping we can observe a clear increase in the complexity as the increase of object number. The histogram divergence results shown in Figure 3.7 gives another interpretation of complexity. Group A has photographs less complex than the ones in groups B. Group D contain samples with repeated elements. Group E is simple but in a different way comparing with the simple samples shown in Figure 3.6.

Figure 3.8 shows example clusters using color histogram features. Group A are photographs with large area filled by blue color, group B is mostly black and white, and group C has photographs colored in yellow. Group D and E are photographs with more various colors.

Clear grouping could also be observed in clusters generated from composition features. As shown in Figure 3.9, photographs in group A have the main object placed in the left of the scene. Group B and C have more objects and most of them are placed at the center. Group D and E have the objects places in

either right of left.

Figure 3.10 shows the photographs cluster generated by ellipse fitness, which is a 10-dimension vector with mean and standard deviation. Photographs in group C have objects of similar sizes. Group D is consist of one main object, and group E has photographs with thin lines.

As for Figure 3.11, the similarity of photographs in the same group is not as obvious as in previous figures.

Although photographs in the same cluster shown in these figures may not share perfect similarity due to the variety of photograph content, such clustering analysis help us to understand what kind of properties are extracted by the proposed features. The object number and histogram divergence features reflect the complexity intuitively. Clusters generated using color histogram and composition features are very easy to understand. Photograph groups extracted using shape features such as ellipse fitness and curvature have more varieties. More carefully adjusting the clustering parameters may lead to more interpretable clusters.

3.4 Summary

In this chapter, we explored the visual complexity concept " description difficulty", which emphasize the amount and variety of the elements.

We introduced various factors of visual complexity used in psychological experiments and summarized them into three categories: composition, shape and distribution. We designed a set of visual complexity operators in the corresponding aspects and implemented the factors in psychological experiments. We proposed visual complexity from perception cues (VCPC) feature by applying the visual complexity operators to well-extracted perception cues such as contour, texture, sharpness and etc.

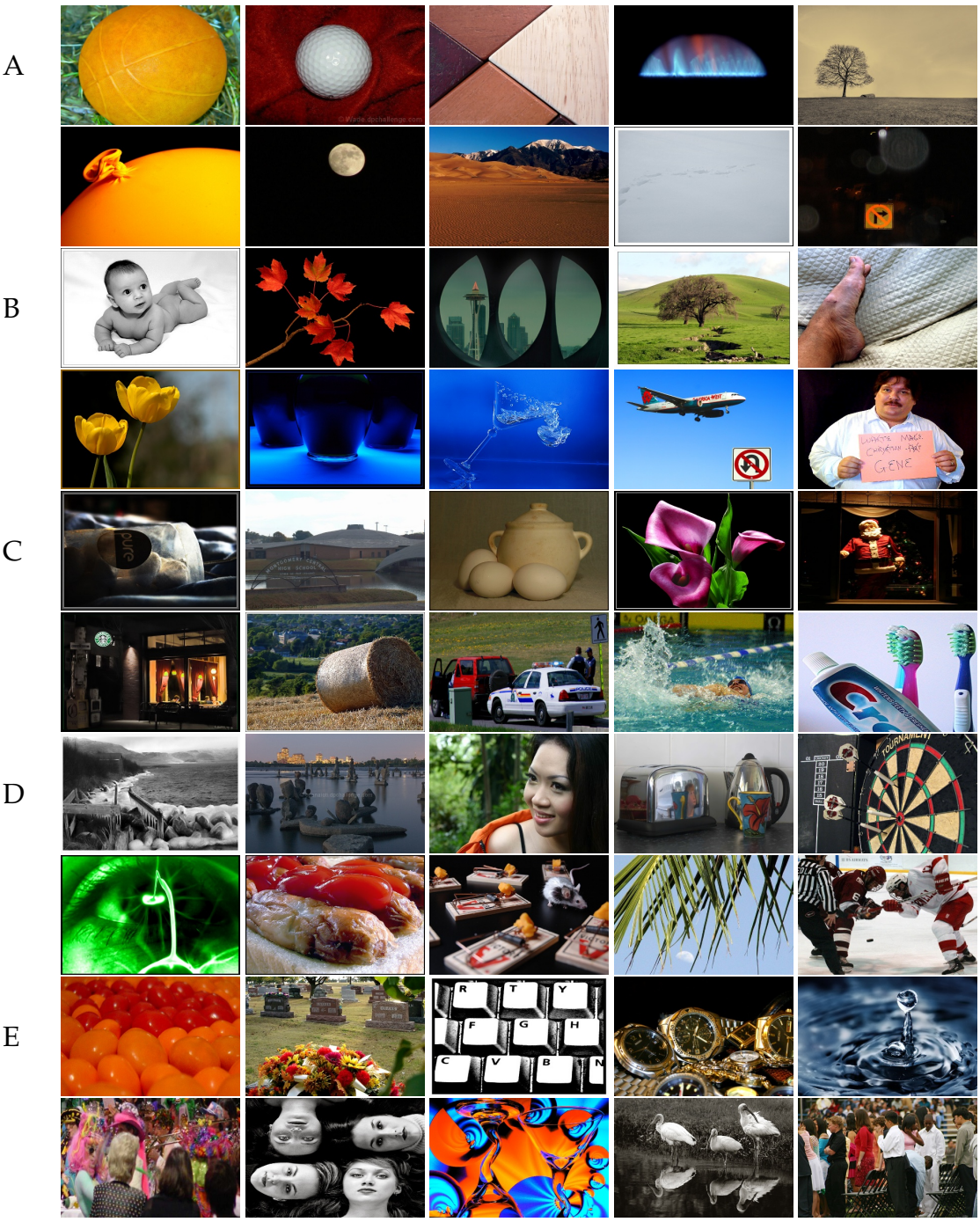


FIGURE 3.6: Sample photographs clusters according to object numbers.

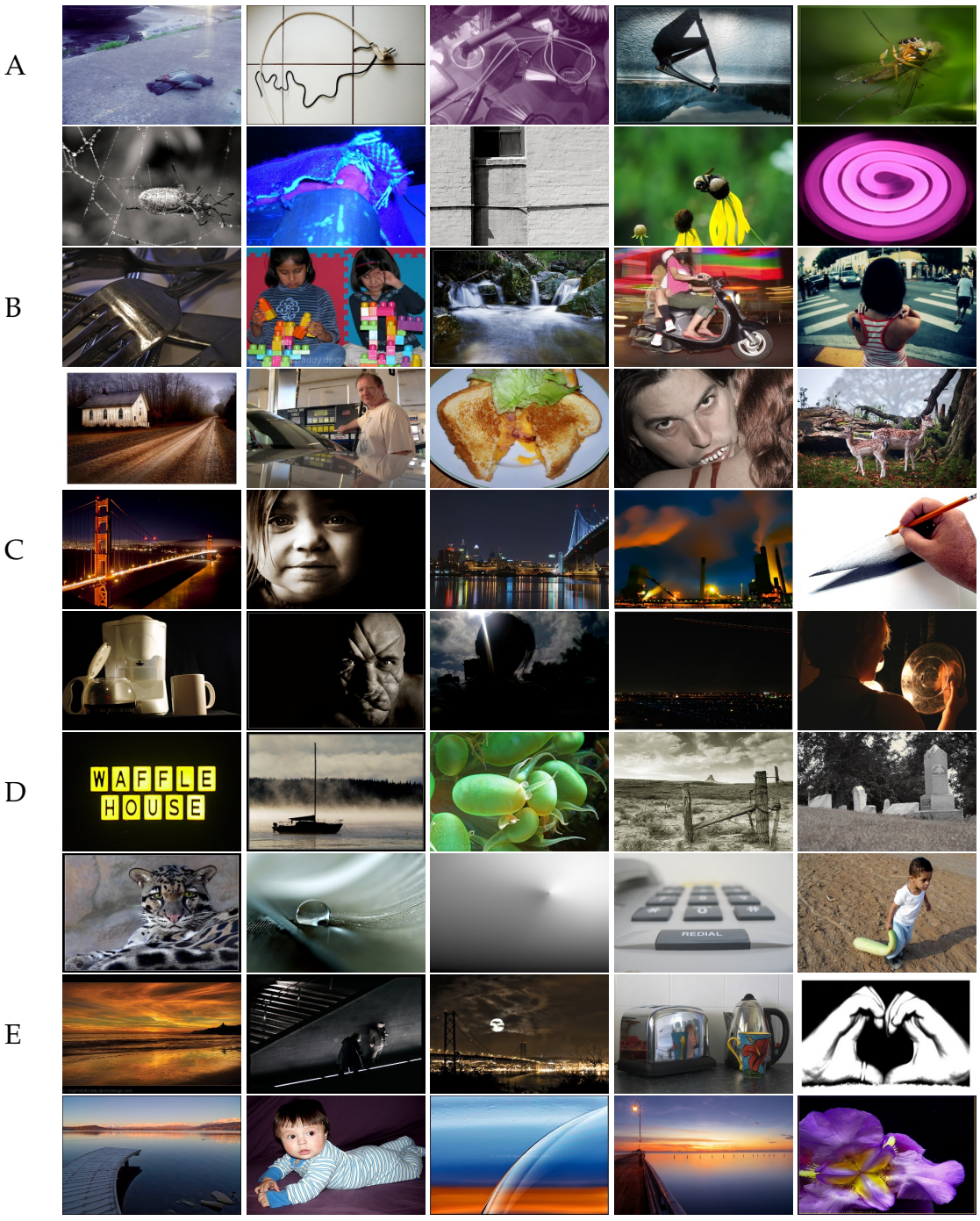


FIGURE 3.7: Sample photographs clusters according to histogram divergence features.

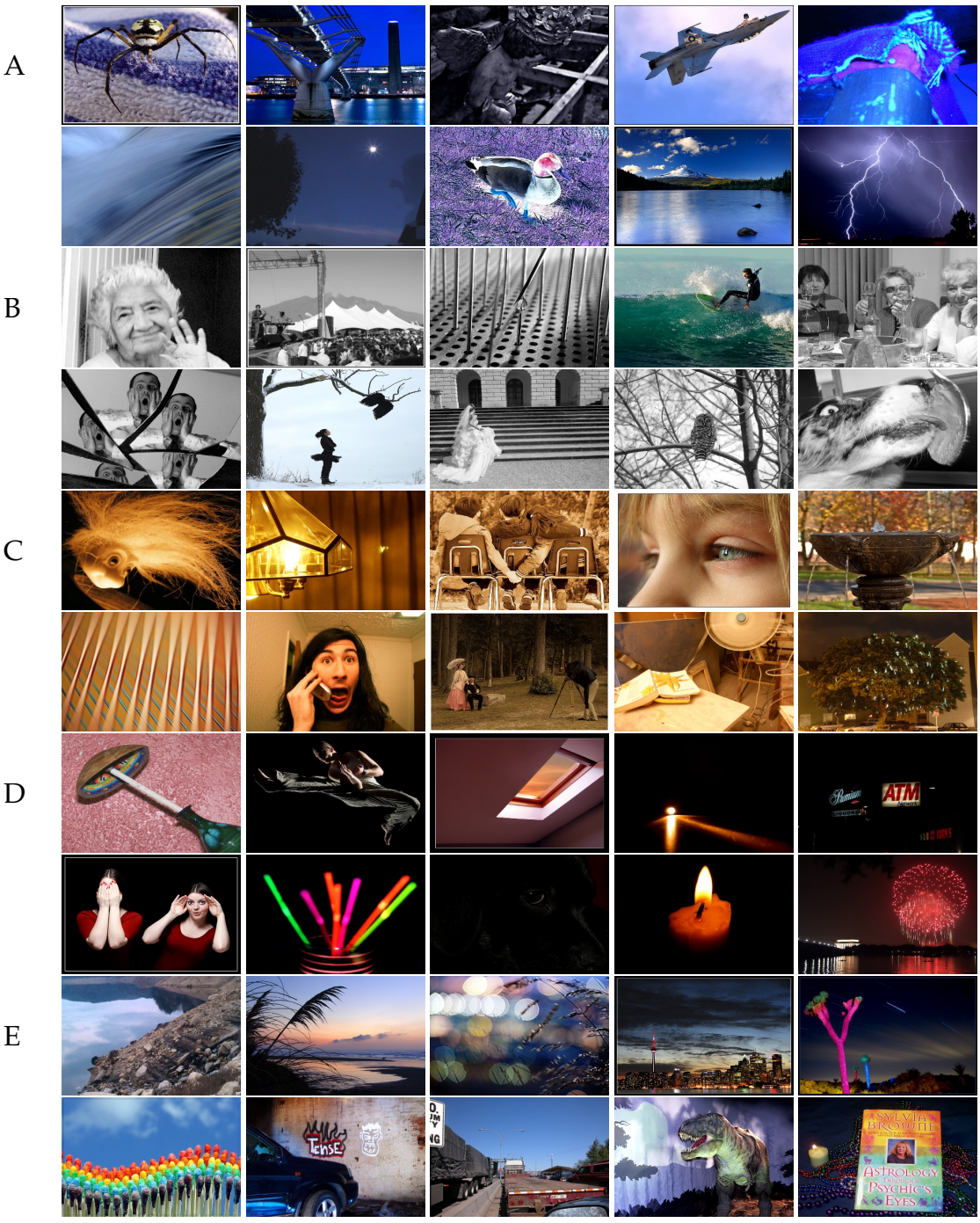


FIGURE 3.8: Sample photographs clusters according to color histogram features.

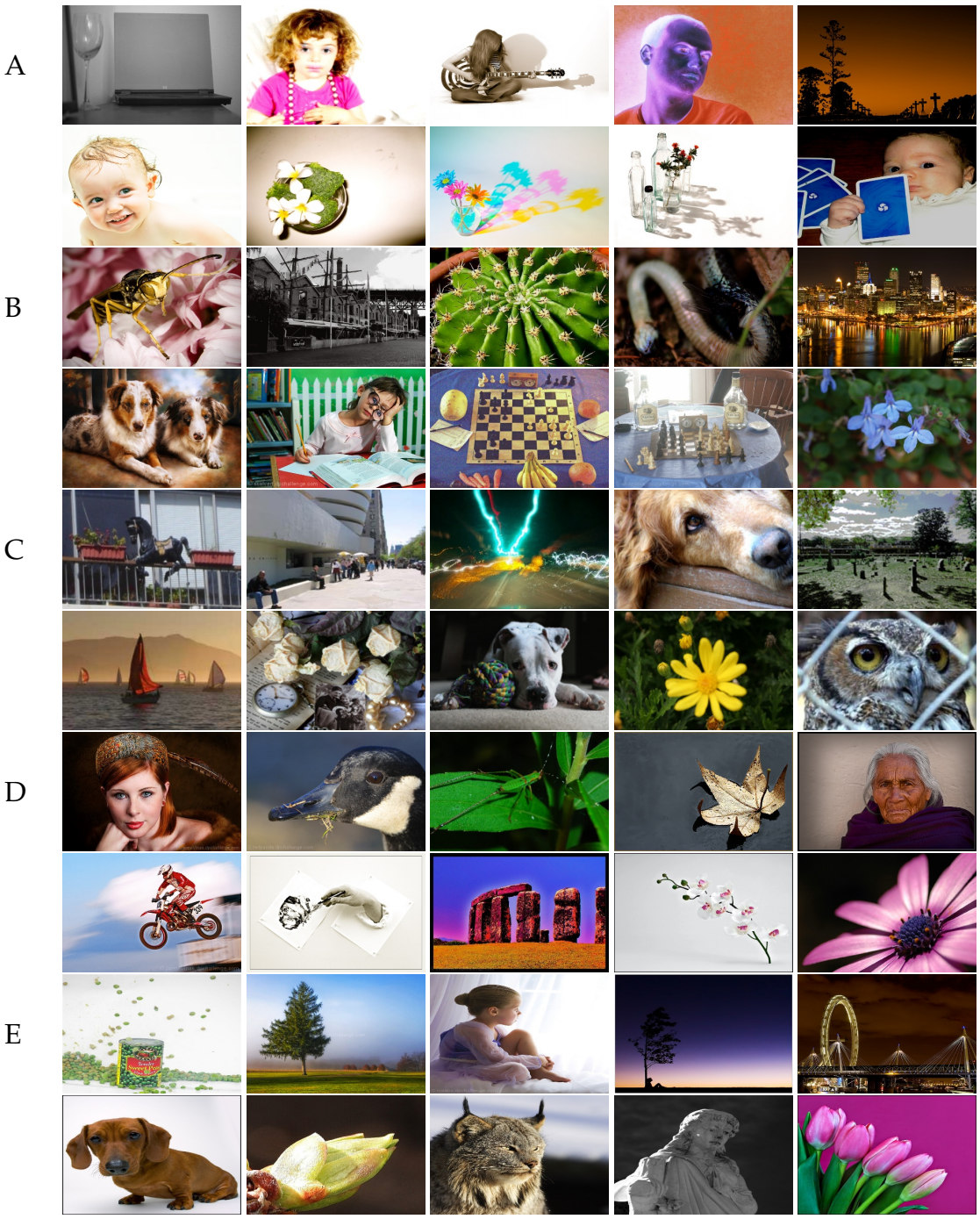


FIGURE 3.9: Sample photographs clusters according to composition features.

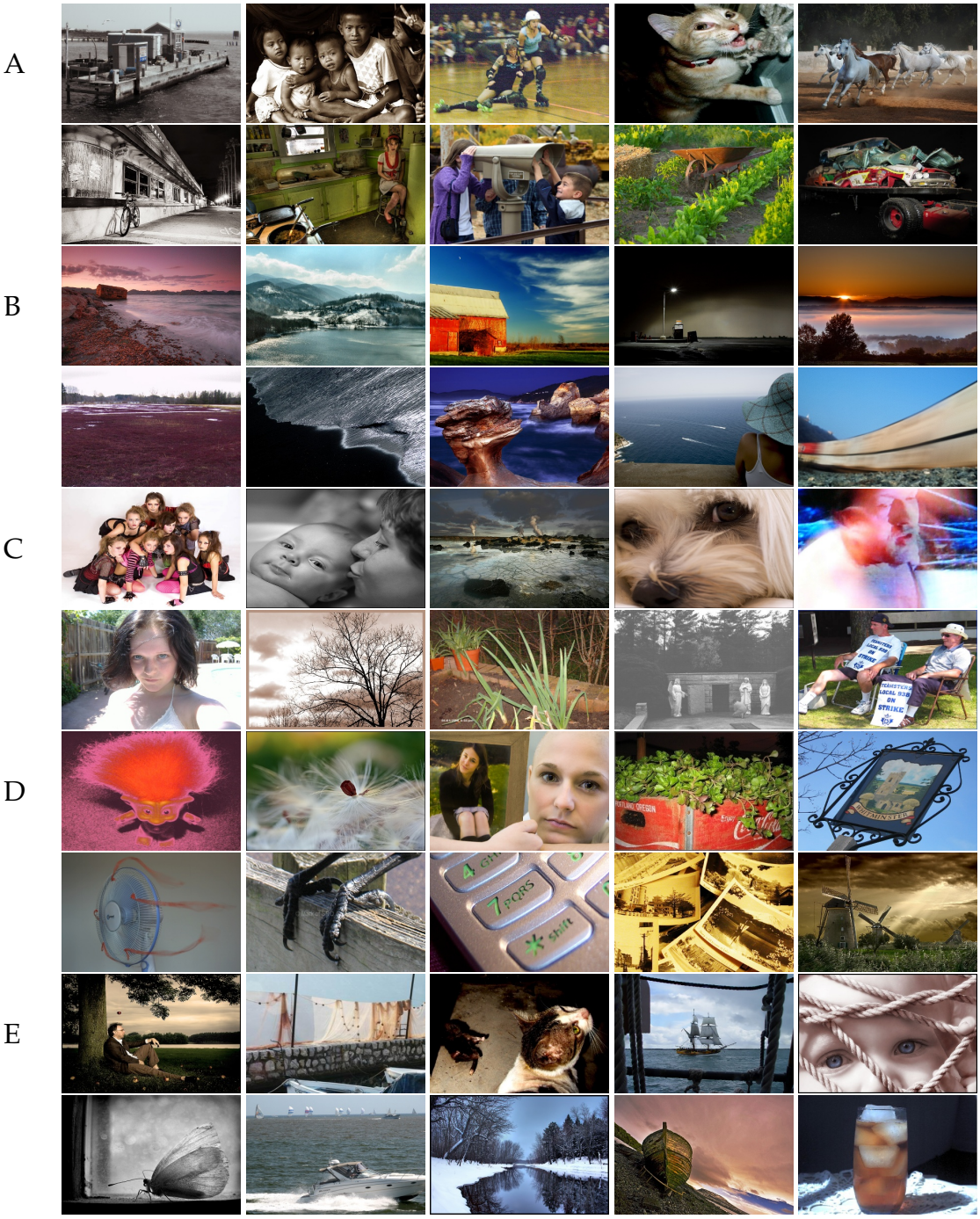


FIGURE 3.10: Sample photographs clusters according to ellipse fitness features.

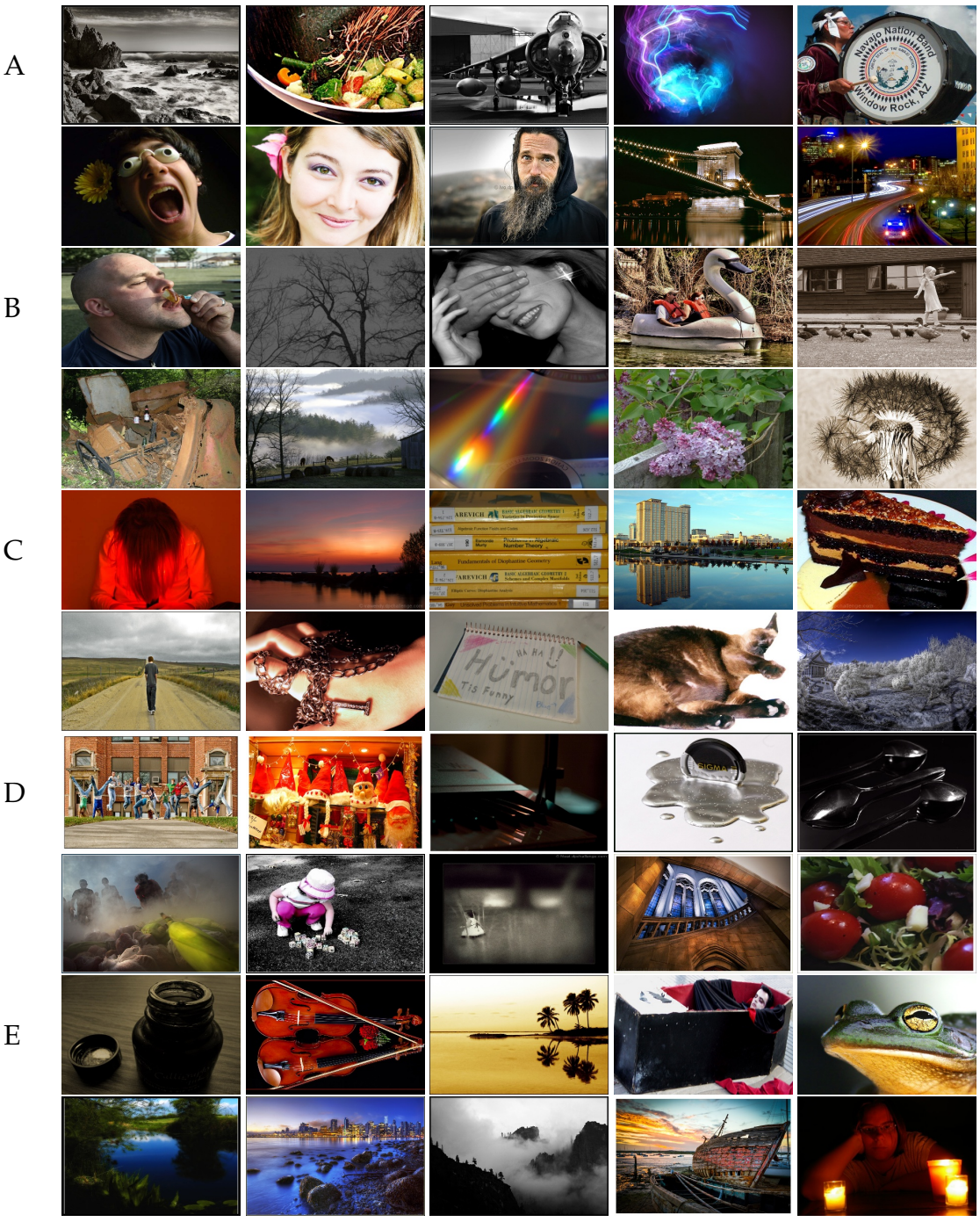


FIGURE 3.11: Sample photographs clusters according to curvature features.

In addition, we conducted clustering experiment on a large set of photographs using the proposed features, and to show what kind of properties are extracted.

Chapter 4

Visual complexity from hierarchical abstraction (VCHA)

4.1 Introduction

In the previous chapter we proposed a set of visual complexity operator that could extract complexity characteristics from photograph. These complexity operators together with the VCPC features mainly analyse the complexity in the aspect of how many elements and how these elements look like and arranged.

In [73], participants are required to classify the images into different complexity groups and take the structure of the scene into consideration by conducting hierarchical grouping task. Such instruction is used to guide the participants as for the concept of complexity: "Simplicity is related to how you see that objects and regions are going well together. Complexity is related to how difficult it is to make sense of the structure of the scene". They found that the control group which merely explored the number and variety of the objects and the structure group have consistent grouping results. Thus they concluded that the hierarchical structure of the scene is one of the factors in human complexity perception.

However, the relationship between the elements, and especially the intrinsic organization of the photograph is absent in VCPC features. To this end, we

propose another set of visual complexity features that could extract the complexity properties in a hierarchical way, which includes the organization of the elements in the photographs.

4.2 VCHA features

First, we identify the different level structures in a hierarchical set of abstractions derived from the original image, and then we extract the edges and contours from these abstractions.

We apply a rolling guidance filter [114] to the original image to smooth out the details and preserve the structures. Rolling guidance filter is proposed for nonlinear image decomposition based on bilateral filter [99]. It can effectively removes smaller structures like texture and noise while preserving larger structures.

The rolling guidance filter has two steps. Firstly, it applies a gaussian filter with spatial scale σ_s to the original image I. Structures smaller than σ_s is completely removed. This output is denoted as J^1. Then a joint bilateral filter is applied to recover the edge iteratively. The output of the tth iteration J^{t+1} is written as

$$J^{t+1}(p) = \frac{1}{S}\sum_{q\in N_p} exp(-\frac{\|p-q\|^2}{2\sigma_s^2} - \frac{\|J^t(p)-J^t(q)\|^2}{2\sigma_r^2})I(q) \tag{4.1}$$

where $S = \sum_{q\in N_p} exp(-\frac{\|p-q\|^2}{2\sigma_s^2} - \frac{\|J^t(p)-J^t(q)\|^2}{2\sigma_r^2})$. p and q are the pixels, N_p refers to the set of pixels in the neighbourhood of p, controlled by spatial scale σ_s. Contrast scale σ_r controls the recovered edge strength.

The algorithm of RGF is summarized as in the following. The two main steps are combined into one by using a constant valued image as the guidance in the initial step.

Algorithm 1: Rolling Guidance Filter

Input: I, σ_s, σ_r and n^{iter}
Initialize J^0 as a constant image;
for $i = 0, i < n^{iter}, i++$ **do**
    $J^t = JointBilateral(I, J^{t-1}, \sigma_s, \sigma_r)$ using 4.1
end
Output: The output image $G = J^{n^{iter}}$

For a given image I, by adjusting the spatial scale σ_s and contrast scale σ_r of the filter, a hierarchical set of abstractions is generated, $\{A_i | i \in [0, 6], A_i = f_{\sigma_{si},\sigma_{ri}}(I)\}$. The basic structures with different sizes and contrast are preserved, whereas details defined by the spatial and contrast scale are smoothed.

We assign the spatial scales as $\sigma_s \in [3, 6, 9, 12]$ and the contrast scale as $\sigma_r \in [50, 150, 300, 500]$. Abstractions with different spatial and contrast scales are shown in Figure 4.1. When the spatial scale is increased with a constant contrast scale, larger patterns are smoothed gradually. If the spatial scale is fixed, a larger contrast scale yields more blurred abstractions. The differences between two abstractions with the same spatial or contrast scale are not significant, so we select the diagonal four abstractions ($(\sigma_s, \sigma_r) \in [(3, 50), (6, 150), (9, 300), (12, 500)]$) together with another two ((3,500) and (12,50)) from the 16 scales, as shown in Figure 4.2. Contour and edge maps extracted using the Canny algorithm and Sobel filter from the six selected abstractions with different spatial and contrast scales are also shown in Figure 4.2. As shown in the contours, at the most abstract scale, the abstraction $(12, 500)$ generated with the filter parameters $\sigma_s = 12$ and $\sigma_r = 500$ only extracts the most important objects, whereas smaller elements and textures are also included at the less abstract scale. The edge maps exhibit similar trends and they also preserve the differences in intensity contrast.

The color and relative color are almost the same in the abstractions and the

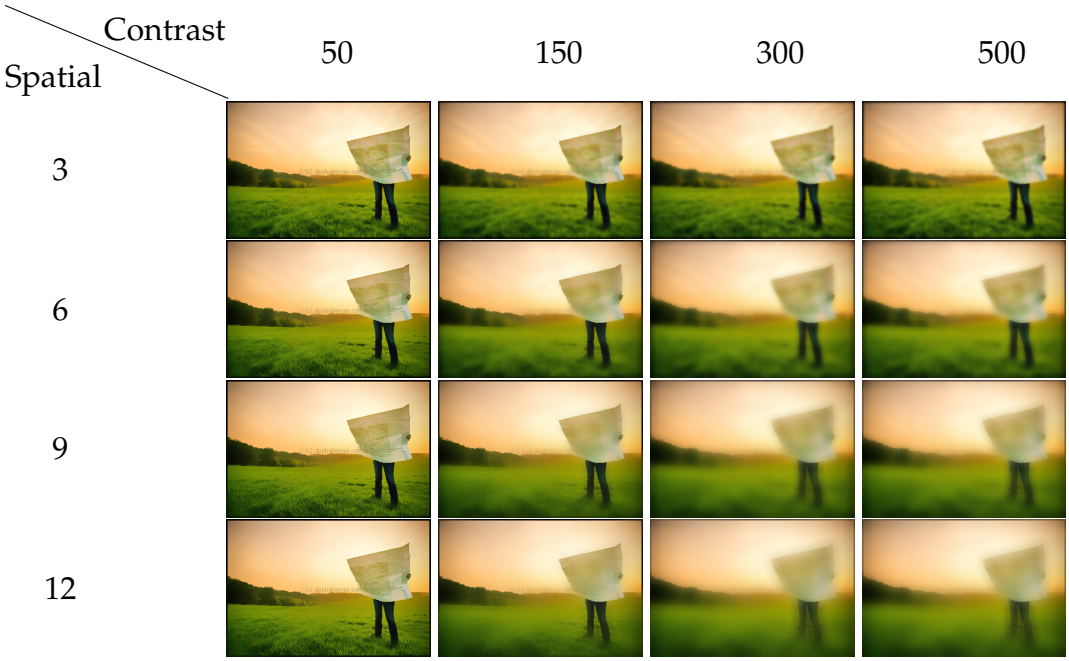


FIGURE 4.1: Hierarchical abstractions of an image. The intensity scale increases from left to right and the spatial scale increases from top to bottom.

original image, so we only calculate the composition and distribution complexity for the color information in the original image. In addition, the composition, statistics, and distribution complexity are calculated for the edge and contour information for all of the abstractions.

The detail steps to calculate the VCHA features are illustrated as in Algorithm 2. We first prepare the abstractions by applying the rolling guidance filter(RGF) with different spatial and contrast scales to the original image and obtain a list of hierarchical abstractions $A = \{A_i | i \in [0, 6], A_i = RGF_{\sigma_{si},\sigma_{ri}}(I)\}$. Considering that the color perception of the original image and the abstractions are almost the same and that we are only interested in the relative color along the most important area, we use the contour extracted from the most blurred abstraction with parameters as $(12, 500)$ as guidance contour for relative color analysis. The color distribution is still calculated from the original image I.

A VCHA feature with a total of 341 dimensions is summarized in Table 4.1.

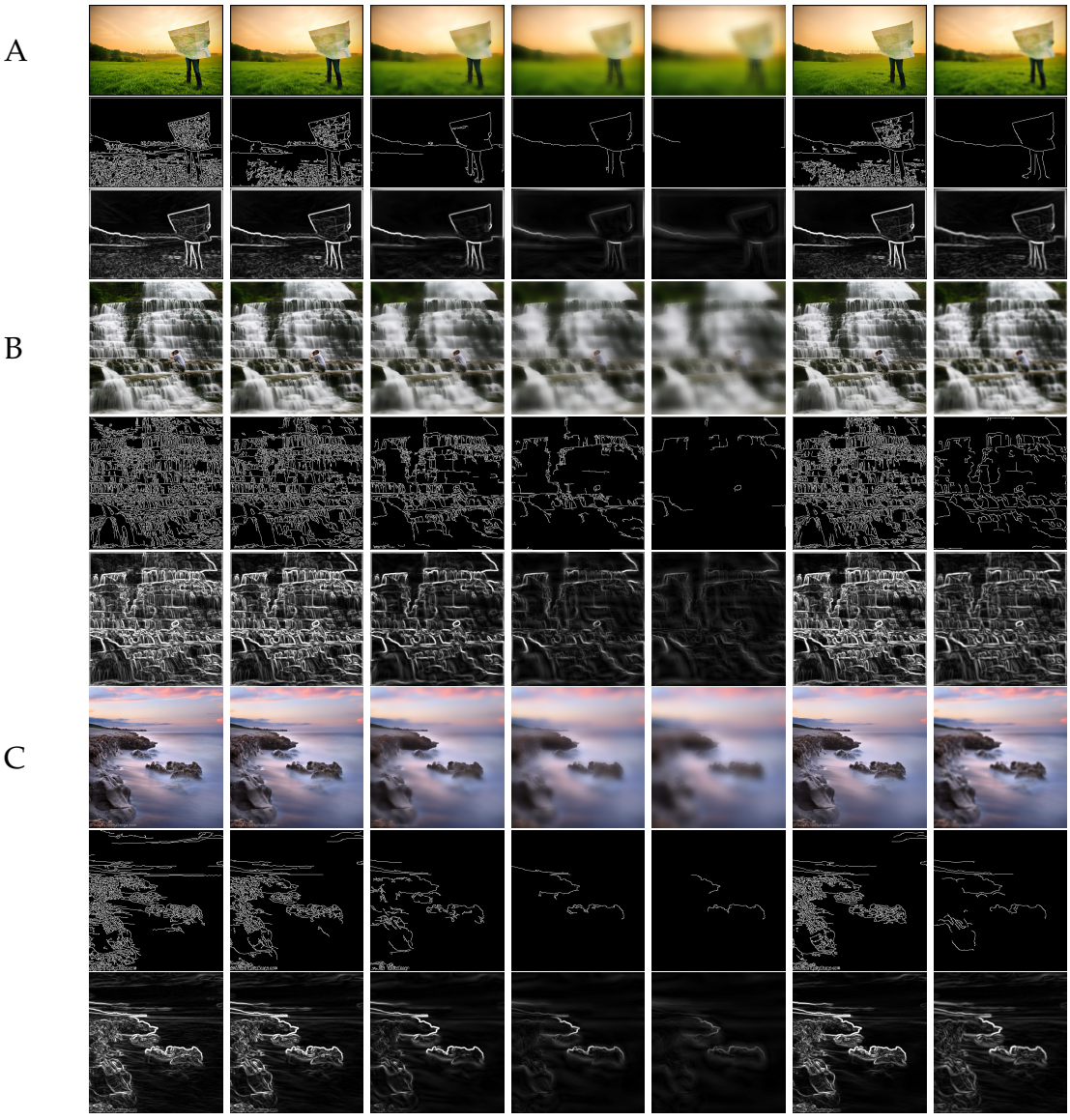


FIGURE 4.2: Hierarchical abstractions of three images and the corresponding contour and edge maps. The original image and abstractions at different scales are shown from left to right. The abstraction, edge map, and contour map are shown from top to bottom.

Algorithm 2: VCHA features calculation

Input: Original image I
parameter set
$\Sigma = [(3, 50), (6, 150), (9, 300), (12, 500), (3, 500), (12, 50)]$
Output: VCHA features f_{vcha}

// Calculate abstractions
Abstraction list $A = []$;
for $i = 0, i < 6, i++$ **do**
    Abstraction $A_i = RGF_{\sigma_{si},\sigma_{ri}}(I)$;
    Append A_i to A;
end
Initialize VCHA features $f_{vcha} = []$;
// Extract featrues from abstractions
for $i = 5, i >= 0, i--$ **do**
    f_{ed} =CalEdgeFeatures((Sobel(A_i)));
    f_{ct} =CalContourFeatures((Canny(A_i)));
    $f_{vcha} = [f_{vcha}, f_{ln}, f_{ct}$;
end
f_{co} =CalColorFeatures((Canny(A_3), I));
$f_{vcha} = [f_{co}, f_{vcha}]$;

TABLE 4.1: Summary of the VCHA features

Category	Perception	Details	Dimensions	Scale
Composition	Edge	Edge and three layers	20	A
	Color	Hue (angle, eccentricity, and composition), chroma, and lightness	25	O
	Relative color	Relative hue and relative chroma within a circular region along contour	10	O
Shape	Contour	Ellipse fitness	10	A
		Object number	1	
		Curvature	8	
Distribution	Edge	Orientation	10	A
	Color		12	O
Total			341	

Note: In the scale column, O refers to the original image and A refers to the abstractions.

4.3 Visual complexity estimation

4.3.1 Dataset construction

We selected 10 photos from 2500 training samples of each category from AVA dataset, making it 80 photos in all. The images were selected as evenly distributed along the aesthetic ratings ranges. Specifically, although in online photograph challenges photos could be aesthetically rated from 1 to 10, the average beauty scores of the 2500 photos in the training set of "Animal" category vary from 2.62 to 8.25. So we sampled photos with the beauty score interval as $(8.25 - 2.62)/10 \approx 0.56$. In this way we managed to collect photos of different aesthetic ratings from various categories.

Five subjects (two females and three males, who were aged from 23 to 28 years) participated in this study, who were all graduate students with normal or corrected to normal vision.

As depicted in Figure 4.3, 10 images from the same category were shown at one time. And the participants were asked to choose a complexity level for these images from 5 options: 1 (very simple), 2 (simple), 3 (medium), 4 (complex) and 5 (very complex). Photos were shown by category in an alphabetic order: "Animal", "Architecture", "Cityscape", "Floral", "Fooddrink", "Portrait" and "Stilllife". Photos were arranged randomly to eliminate any possible pattern between complexity and aesthetic score.

For each image, we calculated the mean and standard deviation of complexity level provided by the five participants. The complexity levels of the 80 images averaged over 5 participants ranged from 1.4 to 5.0, and the standard deviation ranged from 0 to 1.33. As for the image that participants had most different ratings, at least two persons agreed with the same complexity level. This indicates that complexity is measurable for human beings. Figure 4.4 shows example images labelled with different complexity levels.

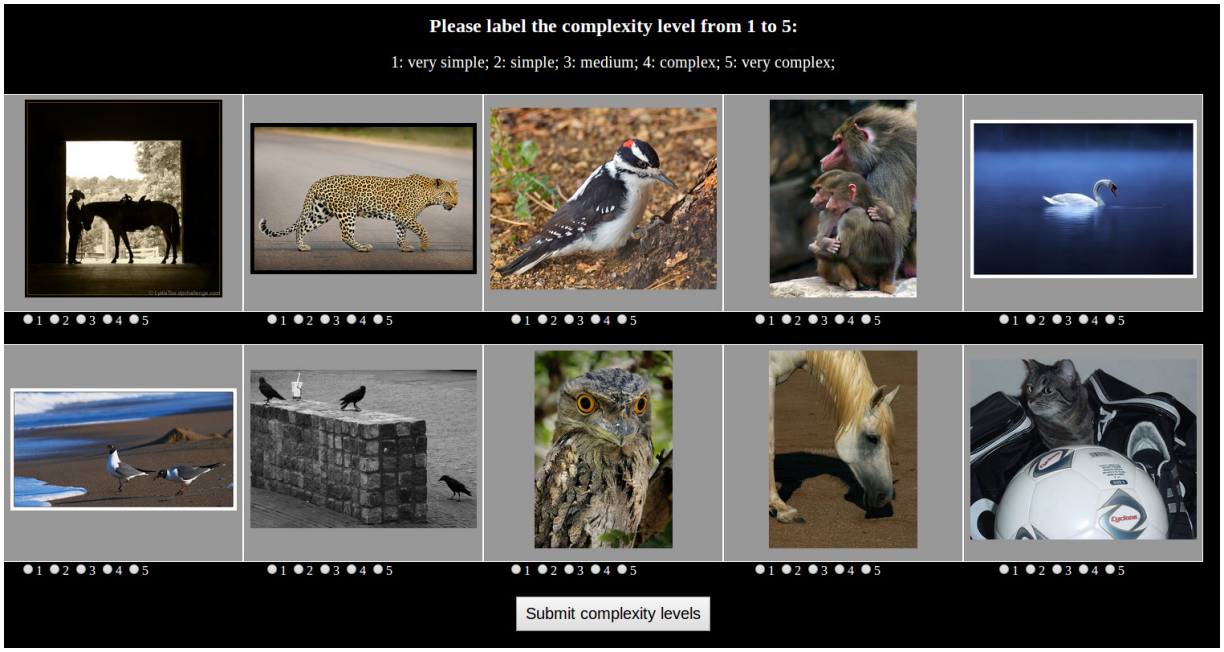


FIGURE 4.3: Interface of complexity labelling experiment for the "Animal" category.

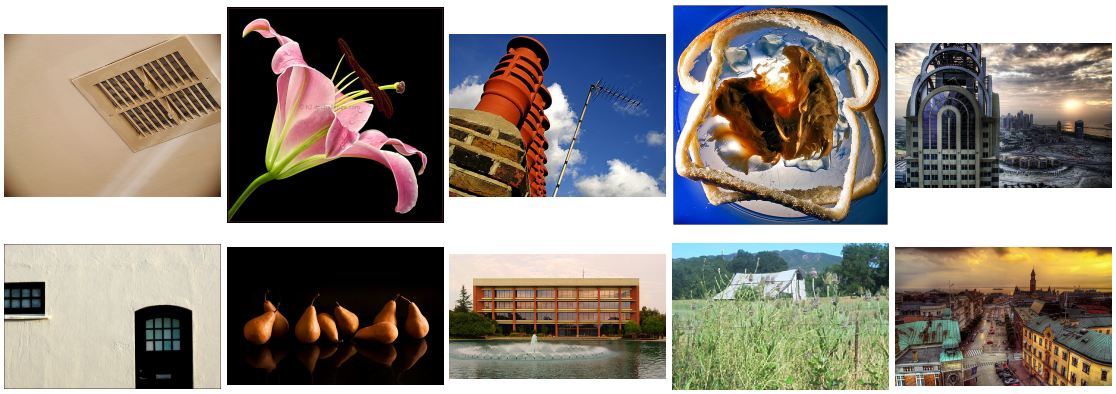


FIGURE 4.4: Example images of 5 complexity levels labelled by participants. The average complexity levels are rounded to integers, and from left to right they are 1(very simple), 2 (simple), 3 (medium), 4 (complex) and 5 (very complex). Images from the same column share the same averaged complexity level.

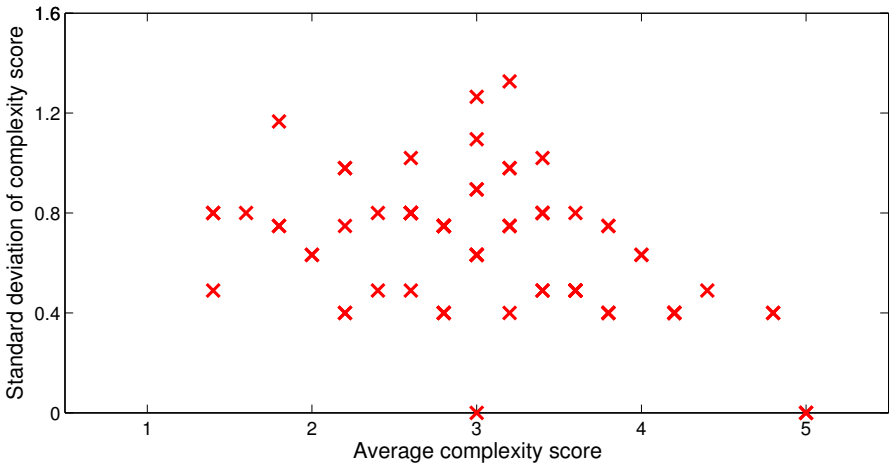


FIGURE 4.5: Distribution of mean and standard deviation of complexity level labelled for 80 images.

TABLE 4.2: Average of standard deviation value for different categories.

Animal	Architecture	Cityscape	Floral	Fooddrink	Landscape	Portrait	Stilllife
0.7302	0.6931	0.4102	0.8260	0.6914	0.5871	0.5690	0.7694

To better understand how participants disagree on complexity levels, we show the distribution of standard deviation along the average complexity score in Figure 4.5. Participants tend to agree with extreme complexity levels. The standard deviation is low for very simple or very complex images, while high for medium images.

Table 4.2 shows the average degree of disagreement of participants concerning different categories. People tend to agree with the complexity level of images from "Cityscape", "Landscape" and "Portrait" categories, while disagreement falls onto categories such as "Floral" and "Animal".

4.3.2 Related works

The idea that complexity contributes to aesthetics judgement has been applied in some existing works [83, 21, 82].

An intuitive measurement of complexity is the mean gradient value [82, 57, 80]. In [80], an image is first split into Lab color space, then the gradient map G is calculated as the maximum value among the three channels,

$$G(x,y) = max(\|\bigtriangledown I_L(x,y)\|, \|\bigtriangledown I_a(x,y)\|, \|\bigtriangledown I_b(x,y)\|) \tag{4.2}$$

$\|\bigtriangledown I_L(x,y)\|$, $\|\bigtriangledown I_a(x,y)\|$, and $\|\bigtriangledown I_b(x,y)\|$ are the gradients at pixel (x,y) for the three channels. Complexity of the image is measured as the averaged gradient values of all the pixels in the image, as shown in the following equation. The W and H are the width and height of the image.

$$C(I) = \frac{1}{WH}\sum_{(x,y)} G(x,y) \tag{4.3}$$

Following a similar idea, features related to the file size of compressed image have been found to be a good approximation of judgements of visual complexity and efficient in aesthetic classification task [83, 19], because compression algorithms such as JPEG and fractal compression generate good abstraction of lines, colors, repetition information of images. For example in [83], complexity is measured by the differences between the compressed image with the original image.

$$C(I) = \sqrt{(I - I_C)^2} \times \frac{S(I_C))}{S(I))} \tag{4.4}$$

where, I is the original image, and I_C is the compressed image obtained by JPEG or fractal compression.

Nevertheless, previous complexity measurements missed the multi-dimensional characteristic of visual complexity and did not take the other factors that are involved in human sensation on complexity, such as curvature, object number, object size, pattern regularity, and pattern compositions.

4.3.3 Experimental results

Complexity is a continuous variable, so regression rather than classification is a better choice for training a complexity model based on previously illustrated features. Thus, we employed gradient boosted trees for regression. Parameters such as the count of boosting iterations and the maximal depth of each decision tree in the ensemble were optimized by 5-fold cross-validation. The accuracy of the regressions was measured using the root mean squared error (RMSE).

To perform tests and comparisons with other complexity features, we randomly selected 10 of the 80 photos labelled with complexity levels in the experiment for testing and the remaining 70 photos were used for training. We conducted the training and testing procedures five times. The performance was measured by averaging the RMSEs.

We compared the proposed visual complexity features with the compression file size-related features proposed by [83] and the summed gradient features used by [80]. The average RMSE for the proposed VCHA feature in the random 5-fold cross-validation test was 0.19/0.56 (training/testing) and 0.35/0.83 for the VCPC feature, whereas the average RMSE values were 0.47/1.05 by [83] and 0.45/0.89 by [80]. Thus, the proposed features outperformed the comparison methods in random 5-fold cross-validation tests.

In order to model the perceived complexity using the labelled complexity levels as accurately as possible, we divided the 80 photos according to the standard deviations of the complexity scores, where lower standard deviation values indicated that the average complexity score was a better approximation of the actual visual complexity. Thus, we only use photos with low standard deviation for training and we expected that the predicted complexity level would be within the variance range for the test photos. We used 70 photos with standard deviations of less than 0.90 for training and the remaining 10 photos with standard deviations varying from 0.90 to 1.33 for testing. The prediction accuracy

TABLE 4.3: Comparison of the regression results for visual complexity features.

Feature	Training		Testing	
	RMSE	Max err	RMSE	Max err
Summed gradient [80]	**0.29**	0.80	0.69	1.39
Compression file size-related features [83]	0.60	1.58	0.73	1.75
Proposed VCPC features	0.31	0.82	0.68	1.26
Proposed VCHA features	0.31	**0.68**	**0.44**	**0.84**
Human perception	0.63	0.89	1.09	1.33

The complexity levels are in the range of [1,5].

and maximum absolute error are listed in Table 4.3.

The worst complexity predictions obtained using the proposed VCPC and VCHA features in tests had absolute errors of 1.26 and 0.84, respectively, which were less than the maximum standard deviation (1.33) among the complexity levels labelled by participants. For the training set, the absolute error of the worst predictions was also lower than the maximum standard deviation (0.89). Thus, the proposed visual complexity model can predict the complexity perceived by humans very well.

4.4 Summary

In this chapter, we explored the concept of complexity in aspect of hierarchical structure.

We introduced the proposed visual complexity features from hierarchical structure (VCHA), which first separates the image into structure with different size the intensity contrast, and then extract the visual complexity features using the operators illustrated in Chapter 3.

We constructed a visual complexity dataset using categorized samples from AVA dataset, and tested the proposed VCPC and VCHA features together with

other complexity features. The experimental results show that the proposed features could predict the average human complexity perception very well.

Chapter 5

Photo aesthetic quality prediction

5.1 Introduction

The image processing and computer vision community has made great efforts to explore computational methods to make aesthetic decisions similar to human beings.

Prediction of photograph aesthetic quality is an undoubtedly challenging and far from solved problem, considering that aesthetic appreciation is highly subjective due to the rich semantics associated with it and the mechanisms of the cognitive process behind it is still not completely understood. Instead of predicting all the subtle aspects of aesthetics, which may include personal experience, moods, and even the trend of art taste and culture, current attempts simplify the problem as predicting the general aesthetic value for an given image in a good-or-bad manner which could be solved by classification, or further quantize the aesthetic ratings to discrete levels, thus transform the problem into a regression task.

The common framework is to first extract aesthetic features to describe the visual appeal property of the content and then adopt machine learning methods to train on these features. Under such framework, the bottleneck of accuracy lies in how well the features could capture the aesthetic properties of the content. Various features have been proposed [55, 38, 14], including low-level

features such as edges distribution and color historgram, and high-level features such as composition, depth of focus, objects and so on. However, most of these features are designed by mimicking photographic rules and practices, which are summarized from limited observations thus lack of generalization. Although there are some other works that adopt generic descriptor and deep features learned by neural network, the design principle of all these features lack the guidance of interpretation for the mechanism of how human process the content in the perspective of aesthetics.

In this chapter, we evaluate the role of complexity played in aesthetic assessment and intend to verify the Berlyne's inverted-U curve on thousands of photos through computational methods.

5.2 Related works

5.2.1 Complexity in aesthetics estimation

Despite the lack of large-scale verification and compelling evidence in psychological theories, complexity has already been widely used for aesthetic classification for photo [83], art [21, 82], and web-page design [101]. Mean gradient value is considered as measurement of complexity in some works [82, 57, 80]. Following a similar idea, features related to the file size of compressed image have been found to be a good approximation of judgements of visual complexity and efficient in aesthetic classification task [83, 19], because compression algorithms such as JPEG and fractal compression generate good abstraction of lines, colors, repetition information of images. Nevertheless, previous complexity measurements did not take the other factors that may influence human sensation on complexity, such as curvature, object number, object size, pattern regularity, and pattern compositions.

5.2.2 Photography-rule-based aesthetics features

Various features have been proposed adhering to photographic rules, common intuition or observed trends on widely appreciated photos. Datta et al [14] proposed a rich set of features, most of which are low-level features, including exposure of light, color, texture, low depth of field, shape convexity, familiarity and so on, and further use SVM for feature selection. Ke et al [38] proposed a set of high-level features which are the characteristics might be used by a human when describe an image, such as edge spatial distribution, color distribution, hue count, blur and some low-level features. Subsequent researchers tend to categorize the features into low-level, mainly referring to color, lightness, texture, and high-level features including composition and content. In many works, low-level and high-level features are jointly used.

[3] analysed five aesthetics attributes: sharpness, colorfulness, tone, clarity, depth and further provided improvements for photographs in these aspects. [54] first detected areas based on clarity ,layout, and human, and then defined the local feature by composition, contrast, sharpness and colourfulness together with global hue and scene composition features.

Among low-level features, color is commonly agreed to be the most important. Comparing with the simple color histogram [14, 55] there are several works concentrating on the aesthetic combination of colors [70, 72, 50].

Composition and content related high-level features have attracted wide attention in aesthetic features design. The basic rule-of-thirds photographic rule is adopted [55, 38, 8]. Further attempt in [93] enumerate more sophisticated composition templates to describe different photograph patterns. Detail spatial relationship between subjects are modelled by graph in [113]. Saliency and sharpness are also considered in [89, 48].

5.2.3 Content description features

Comparing with the "bottom-up" composition features, content related features are more "top-down", which focus on the subject of photographs and evaluate the proper form of subjects in different scenes [55, 16]. Content attributes such as human face, animal region, sky illumination are extracted and used together with other low-level features in [16]. Similarly, generic image descriptor which has been proven to be efficient in recognition tasks are adopted in [58] for aesthetics quality estimation. Following the same idea, several works apply deep learning method to aesthetics estimation [10, 14, 47, 52, 96], considering the recent success of deep learning in various image processing problems.

5.3 Aesthetics estimation dataset

During the past decade, several data sets are designed to serve aesthetics estimation for photos. Most of these data sets consist of samples selected from online photo contests websites, photo.net and Dpchallenge.com. The early ones include the CUHK [38] and PN [14]. And the most recent one is AVA (Aesthetic Visual Analysis) [65] dataset.

CUHK dataset [38] was built in 2006 with 12,000 photographs selected from Dpchallenge.com website. The samples are the top or bottom 10% out of 60,000 photographs, and only binary labels are provided. It was further extended to CUHKPQ dataset [54] by adding 5690 photographs provided by photographic community and university students. CUHK and CUHKPQ datasets are not available recently. PN dataset [14] was set up by Datta et. al. in 2006 with 3,581 samples selected from Photo.net. 30% samples in this dataset are framed and each samples are rated with more than 2 persons. It was extended to a larger dataset PNe [13] latter which has two parts, one with 20,278 samples from Photo.net and the other with 16,509 samples from Dpchallenge.com. All

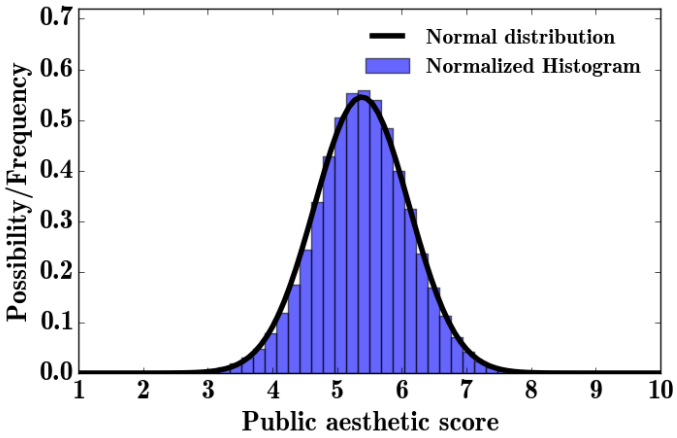


FIGURE 5.1: Normalized 50-bin histogram of public aesthetic scores and the fitted normal distribution curve.

the samples in PNe have more than 10 ratings. AVA dataset [65] is the most recent one and the largest.

We choose the public database AVA [1], which is derived from online photograph challenges, with a rich variety of content. Photographs are selected from more than 1400 topic challenges covering 66 semantic labels. As the aesthetic preference of each image is voted 200 times averagely, the difference between individuals is greatly alleviated.

AVA dataset contains more than 250K photos for generic aesthetic estimation. The photos are rated on a scale from 1 to 10, where 1 refers to the lowest aesthetic value and 10 is the highest. The average score for a photo is used as its public aesthetic value. To avoid confusion, in the following, we use the public aesthetic score to represent the average score for a photo collected from a group of voters. And Figure 5.1 shows the normalized 50-bin histogram for the public aesthetic scores in the dataset, which fit well to a normal distribution with a mean value of 5.38 and standard deviation 0.73.

As introduced in section 5.1, the aesthetics prediction task is implemented by a two-classes-classification method, in which an image is considered as of

[1] http://www.lucamarchesotti.com/ava/

high quality if its public aesthetics rate is over a certain threshold. In such situation, it is necessary to check the variation among users' aesthetics judgements to see how reliably the public aesthetics rate could reflect general users' judgements. We annotate the rates for a certain photograph as $R^{p_i} = \{r_1, r2, ...r_n\}$, where p_i refers to the ith sample in the dataset. We first assume that the distribution of aesthetics rates towards each photograph, follows a normal distribution. And we calculate the standard deviation of the aesthetics rates for each photograph. We consider the accuracy of human labelling as the ratio of the number of photographs, towards which most people agree with the aesthetics quality, in the dataset.

We set the consensus range as $R^{p_i}_{avg} \pm \sigma^{p_i}$, where $R^{p_i}_{avg}$ is the public aesthetics rate, and σ^{p_i} is the standard deviation for the sample. The two standard deviation gap leads to absolute majority (68%) of aesthetics quality judgements. If we divide the samples into two balanced classes, which means we take the average of the public rates for all the photographs in the dataset as the threshold. The threshold is defined as $thres = avg(R^P_{avg})$, where $P = pi|i \in (0, 255330)$ is the whole dataset. Photographs with the lower bondary of consensus range $R^{p_i}_{avg} - \sigma^{p_i}$ larger than the threshold, and the photographs with the higher boundary less than the threshold are considered to have converged aesthetics judgements.

In this way we find out that there are only 11198 photographs towards which absolute majority of the participants give a consensus aesthetics judgement, and the gap between the high and low quality of their public aesthetics rates is 1.91, with 6.47 for the lowest rate for high quality and 4.56 for the highest rate for low quality. Based on such analysis, we could safely claim that it is difficult to divide the photographs with the public rates in the range of [4.56,6.47] into high or low quality classes. In addition, if we discard the samples with public rates close to the threshold, the ratio of the consensus rated samples and

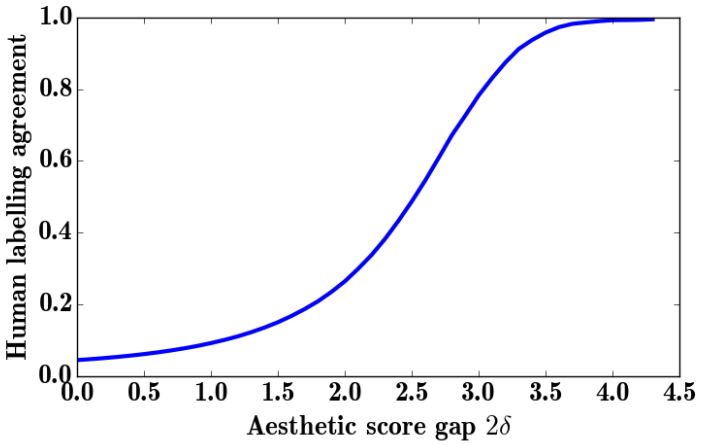


FIGURE 5.2: Human labelling consensus along the aesthetics score gap.

the number of photographs with public rates larger than 6.47 or less than 4.56 is only 23%. We show the ratio of consented samples in the along the aesthetics score gap between high or low quality in the following figure. When the gap is 3.0, human labelling consensus is 78.3%, and it would be 83.1% with the gap as 3.1. Human labelling consensus is over 99% when the aesthetics score gap is larger than 3.8. The ratio of consented samples and the dataset discarding the ambiguous samples within the aesthetics score gap is shown as in Figure 5.2.

The distribution of the aesthetic scores of the consented sample is shown as in Figure 5.3. In the left the ratio between the consented sample with the all the samples with a similar aesthetics level is shown. And in the right, the histogram of the public aesthetics rates of the consented samples are shown.

Figure 5.4 and 5.5 show examples from the two groups: photographs with consented aesthetics judgement and photographs with dissented aesthetics judgements. Samples with high quality are selected from the public aesthetic scores range [5.5, 7.2], and samples with low quality are selected from score range [3.8, 5.5]. Such aesthetics score ranges are the regions that quality are most difficult to distinguish between high or low quality as shown in the left of Figure 5.3.

AVA also defined categorized photographs. For each of the eight categories,

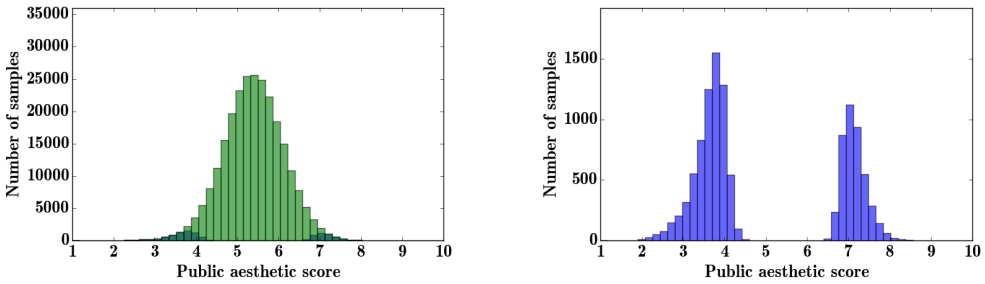


FIGURE 5.3: (A)Histogram of public aesthetic scores for consented samples overlapping all the samples. (B) Histogram of public aesthetic scores for consented samples, which is the dark green area in (A).

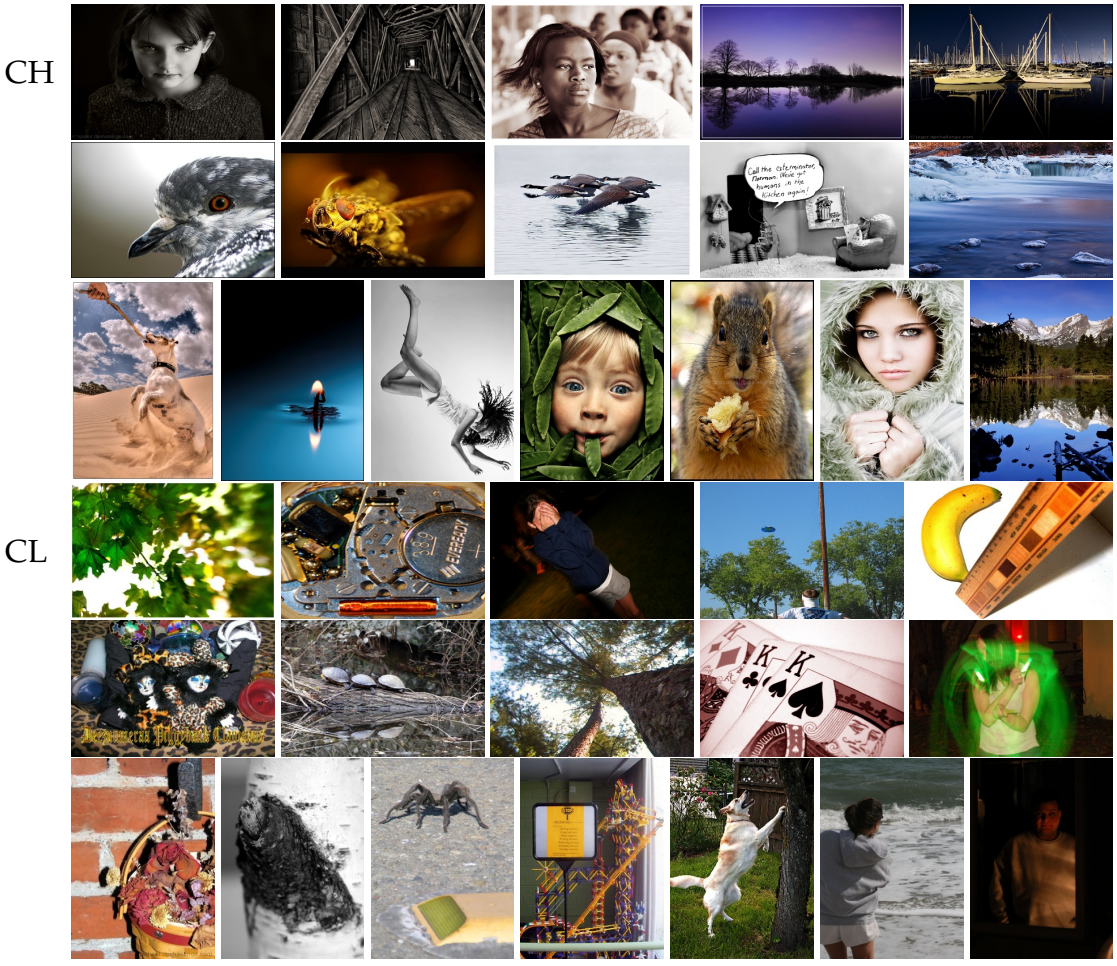


FIGURE 5.4: Sample photographs with consented aesthetics judgements. CP: consented High quality. CL: consented Low quality.

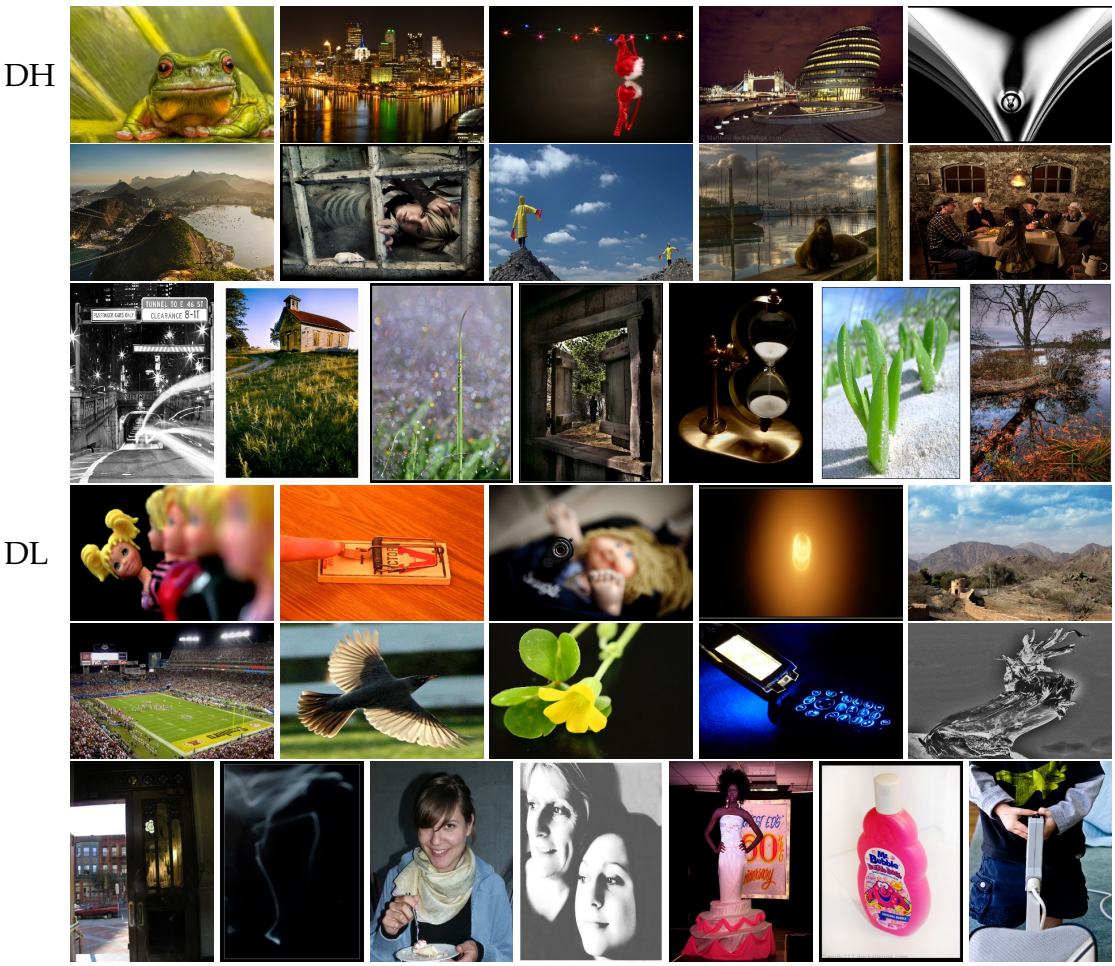


FIGURE 5.5: Sample photographs with dissented aesthetics judgements. DP: Dissented High quality. DL: Dissented Low quality.

5000 photographs are selected and the training and testing sets are also defined using half of the samples. The categories were "Animal," "Architecture," "Cityscape," "Floral," "Food/drink," "Portrait," and "Still life."

We constructed our visual complexity dataset using these eight subsets to retain sufficient semantic variety. We also tested the aesthetic estimation performance of the proposed features based on these categorical subsets.

Using this categorized dataset, we first designed a small-scale preliminary experiment to test whether human sensation on complexity is congruous. Then we use the proposed VCPC and VCHA features to set up the complexity models. After that we calculated complexity levels for large-scale photo database, and analysed the relationship between beauty expectation and complexity level.

As application, we used the proposed visual complexity features to predict aesthetics scores using gradient boost trees regression, and to determine aesthetic quality using random forest as classifier.

5.4 Relationship between visual complexity and aesthetics

In this section we apply the visual complexity model obtained in Section 3.2 to the training sets in AVA dataset (each category has 2500 training photos), calculate the expectation of aesthetic score, and explore its relationship with complexity level.

We calculate the visual complexity level for photos from the training sets in AVA data. As illustrated in Table 4.2, participants tend to agree with the complexity for photos from "Cityscape" category, so we expect more accurate complexity evaluation on "Cityscape" category than other categories. Example photos from "Cityscape" category with different complexity levels are shown in Figure 5.6.

FIGURE 5.6: Example photos from "Cityscape" category of 5 complexity levels calculated by proposed visual complexity model. The two images from the same column share the same complexity level.

The complexity level calculated using our proposed method is rounded to integrate levels. We employ one-way analysis of variance (ANOVA) to compare the expectation of beauty experience along with complexity level. ANOVA results suggest that the beauty score distribution of at least one complexity level is significantly different from those of other complexity levels ($p < .05$ for each category), and box plots for all 8 categories are shown in the left of Figure 5.7.

Due to the large variance ranges, the differences between the beauty score means of different complexity level is not clear. To further test the statistical significance of beauty score expectations, we conduct multiple comparison, group by group t-test, and show the results in the right part of Figure 5.7. The vertical axis is the aesthetic score in the range of 4.5 to 6, and the horizontal axis is the complexity level. Aesthetic score expectations of the complexity level coloured as red are significantly different from the one coloured as blue.

Ascending trends could be observed on the right column of Figure 5.7 in "Cityscape" and "Landscape" categories, and descending trends are shown in "Floral" and "Fooddrink" categories, while in the other categories only weak

ascending or descending trends could be observed. In "Portarit" and "Architecture" categories, we could only observe the ascending trend for the middle 3 complexity levels. And in "Animal" category, the descending trend is not clear for complexity levels 4 and 5.

For "Cityscape" and "Landscape" categories, the ascending trends have clear statistical significance, except that aesthetic assessment expectations of photos with intermediate complexity levels may be easily confused with those of adjacent complexity level. Taking "Cityscape" category for example, mean beauty score of simple photos (complexity level 2) is significantly different from those of extreme simple, complex and extreme complex photos (complexity levels of 1, 4 and 5), while it is hard to tell mean beauty scores of simple photos from that of medium photos (complexity level 2 and 3).

We evaluate the relationship between aesthetic experience and complexity level on AVA dataset training photos (each category has 2500 training photos). Based on our results, we only observed ascending or descending parts of Berlyne's invert-U curve for different categories. As for the ascending trends in "Architecture", "Cityscape" and "Landscape", this is because buildings or landscape scenes are already complex considering the lines and components and few photographer would like to produce too complex photos in these categories. Thus the optimal complexity level in the Berlynes inverted-U curve may be not included in the photo, and the drop of aesthetic experience when complexity level is higher than the optimal level is not observed. As "Animal", "Floral" and "Fooddrink" categories in which most photos focus on single or small number of objects, the descending trends of beauty expectation is understandable. Too complex photo would lead to distraction and difficulty to focus onto the content of the photo. Simple photo is better to express the beauty of these categories. However "Portrait" and "Stilllife" categories are a little different, as photos convey more semantic meanings and are difficult to model by

only low level features.

5.5 Aesthetic quality prediction by visual complexity features

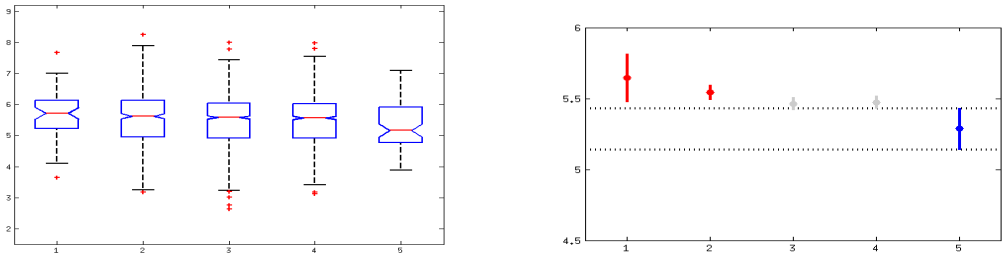
As verified in Section 5.5, visual complexity is closely related to aesthetic experience of photos. In this section we try to predict beauty scores for photos using visual complexity features.

5.5.1 Performance on categorized photos

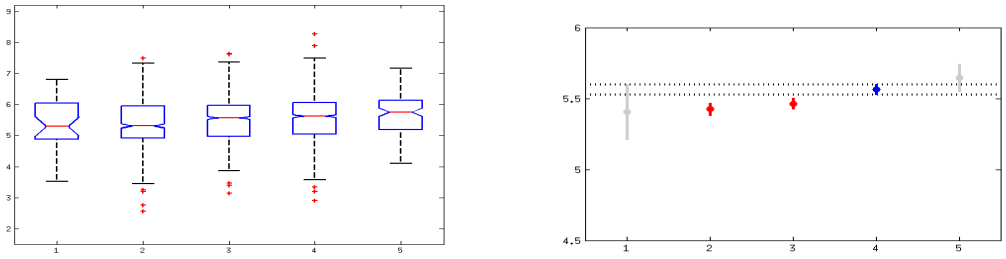
In this part, we conduct the experiments on categorized photographs and compare the proposed features with existing complexity features and also with other aesthetic features.

Comparison with complexity features

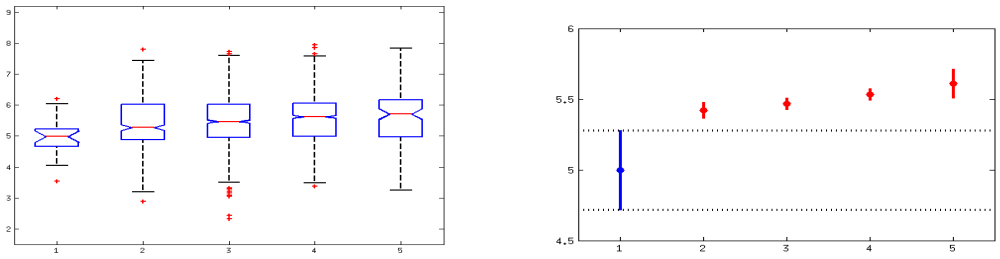
Visual complexity features are first extracted as illustrated in Section 3.2. We employ gradient boosted tree to train the regression model. Parameters are optimized through 5-fold validation similar to Section 4.3.3. The regression accuracy is measured using RMSE, and the correlation coefficient between the predicted beauty score and the one labelled by human beings. As shown in Table 5.1, the proposed visual complexity features outperforms compression file size related features in [83] and sum of gradient used in [80]. Considering the fact that beauty scores range from 1 to 10 and the average error of the proposed method is 0.70 for the best case (“Landscape” category) and 0.97 for the worst case (“Animal” category), the proposed method is capable of giving a reasonable estimation of aesthetic experience with.



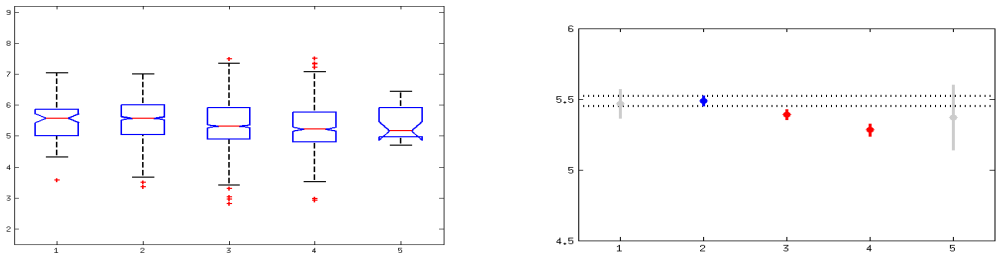
(A) "Animal"



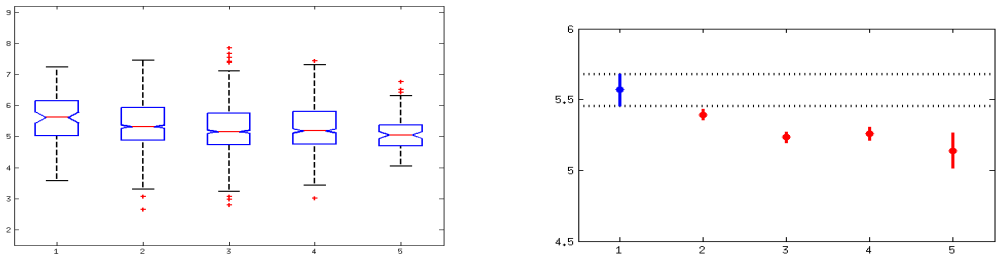
(B) "Architecture"



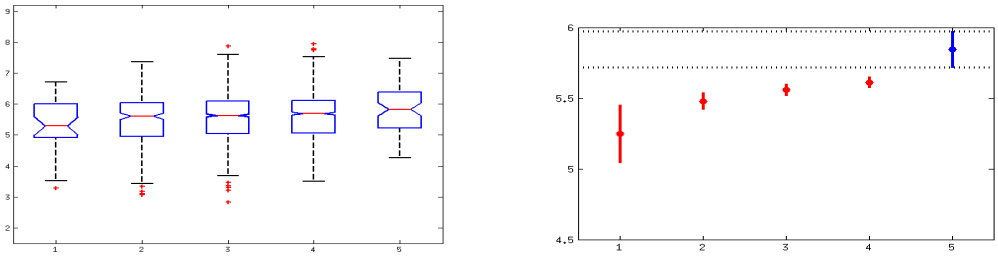
(C) "Cityscape"



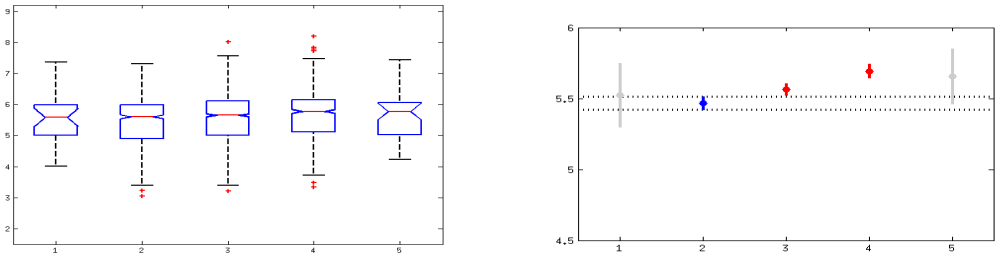
(D) "Floral"



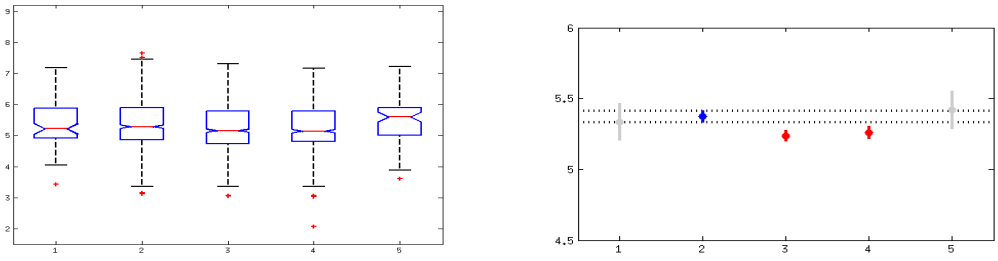
(E) "Fooddrink"



(F) “Landscape”



(G) “Portrait”



(H) “Stilllife”

FIGURE 5.7: Relationship between aesthetic experience and complexity level. Distribution of beauty experience along complexity level is represented by box plot in the left. And the difference significance is shown in the right.

TABLE 5.1: Comparison of beauty prediction results

Category	VCPC		Compression related [83]		Sum of Gradient [80]	
	RMSE	Correlation	RMSE	Correlation	RMSE	Correlation
Animal	0.97	0.16	**0.74**	**0.22**	1.05	0.04
Architecture	0.83	**0.21**	**0.71**	0.17	0.94	0.06
Cityscape	0.83	**0.30**	**0.82**	0.18	0.82	0.13
Floral	0.83	**0.21**	**0.77**	0.18	0.88	0.01
Fooddrink	**0.74**	**0.31**	0.80	0.20	0.83	0.04
Landscape	**0.70**	**0.38**	0.85	0.15	0.76	0.12
Portrait	**0.74**	**0.26**	0.78	0.15	0.84	0.07
Stilllife	0.78	0.23	**0.72**	**0.24**	0.83	0.04

Beauty scores are in the range of [1,10].

We also tested the visual complexity features under the high/low quality classification task. Photos are divided into high or quality class by introducing a threshold parameter δ. Photos with beauty scores higher than $5.5 + \delta$ is considered as of high quality, while photos with beauty scores lower than $5.5 - \delta$ is considered as of low quality. Higher δ leads to more unambiguous training samples making the classification easier, and when $\delta = 0$ the whole training set is used. We employ random forests as classifier. The maximum tree depth is set as 5, and the maximum number of trees in the forest is set as 100.

We set the threshold δ as $[1.0, 0.9, ..., 0.1, 0.0]$. For $\delta = 1$, there are several hundreds of photos left for each category, which is large enough to test the proposed method. And the performance is shown in Figure 5.8. These results are consistent with the performance in regression task. Visual complexity features have best performance for photos from "Landscape" category, and worst performance for "Portrait" category. This is because features indicating complexity in landscape photos mostly refer to more objects, and complex topographies and landforms could be easily summarized through composition and statistical features. On the contrary, the complexity of portrait photos which are mainly

human faces is difficult to measure using only low-level features. Semantic interpretations are necessary, and familiarity may be a predominate factor in complexity detection.

The proposed VCPC features averagely outperform 8.5% compression related features in [83] and 14.7% over sum of gradient in [80] for the "Landscape" category, and for the "Portrait" category 3.9% over compression related features in [83] and 6.8% over sum of gradient in [80].

Comparison with other features

In the following, we show the classification comparison on categorized photographs with the threshold set as 0. Here we compare the proposed VCPC and VCHA feature with the color features [49] and the Caffe CNN features [36, 41]. The Caffe features are of 4096 dimensions and are pre-trained on the ILSVRC2012 object classification dataset [41]. It is proved to have good accuracy in classifying 1000 objects classes, and here we adopt the features directly to check whether such object information contribute to the aesthetic estimation task.

We used the linear SVM as the classifier. And the results are shown as in Table 5.2. The proposed VCHA features have the best performance for most categories, especially in "Stilllife" and "Cityscape". The VCPC features work well for "Animal" and "Fooddrink". Caffe features give a comparative results but quite limited. Color features have the best performance in "Portrait" which indicate that for such photographs of rich semantic meaning color is important due to the inspired emotions.

TABLE 5.2: Comparison of categorized classification performances

Categories	Color [49]	Caffe [36]	VCPC	VCHA
Animal	58	61.92	**63.81**	62.80
Architecture	57	57.19	60.96	**61.60**
Cityscape	64	61.75	65.60	**66.32**
Floral	61	60.40	63.73	**64.73**
Fooddrink	65	58.44	**63.66**	64.20
Landscape	64	64.46	64.34	**65.02**
Portrait	**62**	57.35	61.41	60.21
Stilllife	61	56.24	61.94	**62.38**

The aesthetic rates range from 1 to 10 and threshold to distinguish high/low quality is set as 5.5.

5.5.2 Comparison with state-of-art aesthetics features on generic photos

Since developed in 2012, the AVA dataset has been used in many different ways concerning the thresholds to separate high and low quality classes, as well as various number of samples for training and testing.

In order to ensure a fair comparison with state-of-the-art methods, we conduct experiments under the following conditions.

- AVA20K5. This was the experiment setting used in [65, 52], where 20k samples were used for testing and the remaining 230k samples are used for training. To define high or low aesthetic quality, the threshold was set at 5, which yielded 167k positive samples and 68k negative samples in the training set, and 14k positive and 5.7k negative samples in the testing set.

- AVA20Km and AVA230Km. In this experiment, we adjusted the threshold to the mean value of the public aesthetic scores S_m for the samples in the training set, which was around 5.38. Thus, the numbers of positive and negative samples were almost the same in both the training

and testing sets. The training and testing samples in AVA20km were the same as those in AVA20K5. [65] defined another generic category training set containing 20k samples, which did not overlap with the testing set in the AVA20k dataset. Using the AVA230Km experimental settings, we managed to train the aesthetic model with only a small proportion of the dataset and we tested it with the remaining 230k samples.

- AVA40Kh. Tian et al. downloaded a subset of the AVA dataset containing 190k samples, and they used the top and bottom 10% samples in their experiments [96]. The score gap between the positive and negative samples is 1.83. Thus, this experiment was performed with around 40k samples, where the training and testing sets were comprised equal halves.

- AVA11KC. This is the consented samples shown as in Section 5.3. This subset is a little unbalanced, with 62% samples are of low quality. We divide it into train and test sets half-by-half.

The proposed features are scaled according to the training set, making sure that feature of each dimension is within the range of $[0, 1]$. We tested various classification methods and found that support vector machine (SVM) gave the best performance but require a long time to build the model especially when there are large-scale training samples, and that adaptive boosting (AdaBoost) gave a comparable performance as SVM while saving a lot of training time. Thus, in the following, if not specifically noticed, the experiments for generic photos were performed with AdaBoost for the proposed visual complexity features. Parameters of the classifier were determined through 5-fold cross-validation.

TABLE 5.3: Comparison of the performance of various aesthetic features and classification models using AVA20K5.

Features	Generic [65]	CNN [51]	DeepMA [51]	VCPC	VCHA
Accuracy (%)	66.7	72.32	74.46	72.74	73.41

AVA20K5

In this experimental setting, we compared our complexity features with generic features [65], convolutional neural network (CNN), and deep multi-patch aggregation (DeepMA) features [51].

In [65], a linear support vector machine (SVM) model was trained based on Fisher Vector signatures computed from the generic features and SIFT descriptors. The DeepMA and CNN [51] methods employed neural networks, with a soft max classifier at the end of the framework.

The evaluation metric was the overall accuracy, as used by [65, 51]. The results of the comparisons are shown in Table 5.3. Our complexity feature and our previous features with preprocessing performed better than the generic features[65], and even better than CNN, the accuracy of which was reported as the baseline by [51]. The accuracy of our VCHA feature was also comparable to that of DeepMA [51].

AVA20Km and AVA230Km

As found with AVA20K5, VCHA obtained better predictions than VCPC. Thus, we use the performance of VCHA to compare the performance under balanced and unbalanced conditions, together with the influence of different sample numbers in the training and testing test.

Our VCHA features obtained 64.23% accuracy for AVA20Km. Compared with the unbalanced condition (AVA20K5), there was a decrease in accuracy of

about 10%. The performance obtained with AVA230Km was 61.10%, which was 3% lower than that using the AVA20Km.

Figure 5.9 shows the change in performance with an aesthetic score gap of $2\delta \in [0.0, 0.2, ..., 3.6, 4.0]$, as well as the numbers of samples used in the training and testing sets. When the aesthetic score gap 2δ was 2.4, for AVA20Km, 23,350 photos were used for training and 2034 photos for testing, and the accuracy was 82.12%, as shown in Figure 5.9a. For AVA230Km, the numbers of samples used for training and testing were 1996 and 23,655, respectively, and the accuracy was 74.01%, as shown in Figure 5.9b. The accuracy with AVA20Km was 2% better than that with AVA230Km averagely for different aesthetic score gaps.

In Figure 5.9a and 5.9b, we can observe the tendency of increasing in accuracy along the score gap increase. Increase of accuracy is due to the clearer distinguish between the positive and negative samples. However as the number of photographs in testing set decrease, the number of samples that the proposed features do not work well also decreases. But the correctly and wrongly predicted samples are not distributed evenly along the aesthetic scores in the testing set. For a specific training and testing set, the ratio of such samples and the total testing samples may increase, as shown in Figure 5.9a, when the aesthetic score gap is increased over 2.4 there are some decrease in accuracy. If the training and testing sets are divided differently for many times and the averaged accuracy along the aesthetic score gap would be perfectly increasing.

In contrast to the methods used in previous studies to present their prediction results, where only a small proportion of the prediction results are generally displayed in the manuscript due to page length restrictions, we present our prediction results for AVA20Km on a web page [2], as shown in Figure 5.10. All of the prediction results for the 20,000 samples can be accessed via this page and a user can check the performance of our method for photos with specific

[2]Prediction results webpage link: to be made public after acceptance

aesthetic ranges by setting the low and high aesthetic scores. The copyrights of the photos belong to the photographers, so we only show thumbnails on our site. Information is shown by hovering over the images such as the aesthetic scores. After clicking on the thumbnails, users are redirected to the original page of the photo.

AVA40Kh

Tian et al. [96] implemented many manually prepared features and compared them with using an SVM as the classifier. We followed their protocol and reported the performance of the proposed features. We implemented two other complexity features: compression-related features [83] and histogram gradient-related features (PHOG) [80].

As shown in Table 5.4, the complexity features obtained comparable performance to the photography rule-based features. Compression size file-related features [83] had better accuracy than Luo et al.'s features [55]. In addition, PHOG features obtained similar performance to "Efficiency" [48].

Our VCHA features performed better (with 76.28% accuracy) than the VCPC features (75.01% accuracy) by using well extracted contours, textures, and other information, as shown in Figure 3.4.

Both of the proposed VCPC and VCHA features performed better than all of the hand-crafted features and even better than the deep features RAPID [52]. VCPC feature has comparable accuracy with Tian et al.'s [96] result. Our VCHA features gave the best prediction.

Figure 5.11 shows the sample photographs predicted correctly by VCHA features. And Figure 5.12 shows the sample photographs predicted wrongly by VCHA features.

TABLE 5.4: Comparison of the classification accuracy with AVA40Kh.

Feature category	Algorithm	Accuracy (%)
Photography rule-based	Luo [55]	61.49
	Efficiency [48]	68.13
	Datta [14]	68.67
	Ke [38]	71.06
Content description	Generic [58]	68.55
	RAPID (CNN) [52]	74.54
	Tian (CNN) [96]	75.89
Complexity	PHOG [80]	67.50
	Compression [83]	65.19
Visual complexity	Proposed VCPC	75.01
	Proposed VCHA	76.28

TABLE 5.5: Performance on AVA11KC

SVM parameters	Feature	Accuracy	Precision	Recall	F1-score
Linear, $C = 1$	VCPC	0.799	0.81	0.80	0.80
	VCHA	0.811	0.82	0.81	0.81
RBF, $C = 1000$	VCPC	0.817	0.82	0.82	0.82
	VCHA	0.838	0.84	0.84	0.84

AVA11KC

We test the proposed VCHA and VCPC features on the consented samples AVA11KC dataset. We first used linear SVM with class margin set as $C = 1$, then refined the classifier with RBF kernel and class margin as $C = 1000$. The performance is summarized as in Table 5.5. Both features could predict correctly for over 80% samples.

Example results of the prediction by VCHA and VCPC features using linear SVM are shown as in Figure 5.13 and 5.14.

TABLE 5.6: Classification accuracy for AVA40Kh using part of VCHA features.

Color	(3,50)	(6,150)	(9,300)	(12,500)	(3,500)	(12,50)
69.54	73.22	72.42	70.89	69.71	69.76	73.0
Color + Diagonal 4 abstractions				75.17		
Composition				70.01		
Shape				71.07		
Distribution				67.81		

Analysis of VCHA and VCPC features

The prediction accuracy of VCHA features on AVA40Kh is 75.41%. In the following we show the contribution of different parts of VCHA features. We first show the classification accuracy predicted using features extracted from different abstractions in Table 5.6. Color features are calculated from the original image and have an accuracy of 69.54%, which is lower than the features extracted from edge and contour information from the abstractions. Among features from the abstractions, smaller contrast scale generally leads to better performance. And if we only select the diagonal four abstractions, the accuracy is 75.17%, which is close to the performance 75.41% by all the VCHA feature. We also summarize the performance of using different category features. Shape features give better performance 71.07% comparing with 70.01% by composition, and 67.81% by distribution.

In order to compare the importance of the VCHA features clearly, we use linear SVM as the classifier and group the features into color for the original image and features extracted from edge and contour information belonging to the other six hierarchical abstractions. As shown in Figure 5.15, color features and features from abstraction layers $(3, 50)$, $(6, 150)$, and $(3, 500)$ are more important, considering both of the average importance and the number of features marked

as outliers with high importance. Abstractions with smaller spatial scale generates more important features. Although features from abstraction $(12, 50)$, $(12, 500)$, and $(9, 300)$ have lower average importances, some features (marked as outliers) from these abstraction layers also have high importances, which are comparable to the most important features from the abstractions with smaller spatial scales.

We further analyze the feature importances in VCHA and VCPC features using AVA11KC dataset.

VCHA Features with top 20 importance values are summarized as in Table 5.7. We divide the VCHA features into different layers, color information from the original image and the features extracted from edge and contour perceptions for six abstractions as illustrated in Figure 4.1. Most important features are from abstractions with spatial scale and contrast scale (3, 50), (6, 150), (9, 300). Some features from abstraction with parameters (3, 500) also have high importance. Color distribution and edge orientation features contribute a lot to the aesthetics judgement. As for the category of the features, the most intuitive complexity-related features are the object number which ranks 5th, and the ellipse analysis features, such as the average and standard deviation of bounding box width-height-ratio (ranked 1st and 6th), and ellipse solidity standard deviation (ranked 8th and 15th). Besides, contour curvature also joins the top 20 list.

The 20 features with most important values in VCPC features are shown as in Table 5.8, which is quite different from that in VCHA features. Ellipse angle mean is the most important features. The second important feature is one from texture orientation histogram. Color composition and histogram also ranks a lot in the top 20 list.

TABLE 5.7: Top 20 important features in VCHA

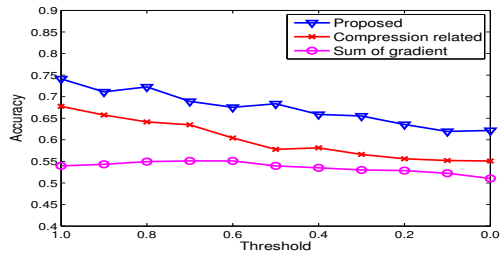
Rank	Layer (σ_s, σ_r)	Description
1	(6, 150)	(+) Ellipse bounding rect WH ratio mean
2	(6, 150)	(-) Contour curvature
3	Color	(+) Histogram
4	(3, 50)	(+) Edge orientation
5	(6, 150)	(-) Object number
6	(3, 500)	(+) Ellipse bounding rect WH ratio std
7	Color	(+) Histogram
8	(6, 150)	(-) Ellipse solidity std
9	Color	(-) Histogram
10	(3, 50)	(+) Edge orientation histogram
11	Color	(+) Histogram
12	(3, 50)	(+) Edge orientation histogram
13	(6, 150)	(+) Edge composition
14	(9, 300)	(+) Contour curvature
15	(9, 300)	(-) Ellipse solidity std
16	Color	(-) Histogram
17	Color	(+) Relative color composition
18	(6, 150)	(-) Contour curvature
19	Color	(-) Histogram
20	(6, 150)	(+) Contour curvature

TABLE 5.8: Top 20 important features in VCPC

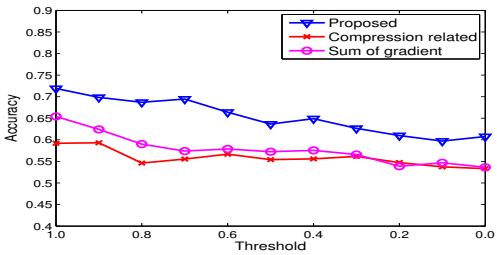
Rank	Description	Rank	Description
1	(-) Texture entropy	11	(+) Line orientation histogram
2	(+) Ellipse angle mean	12	(-) Curvature
3	(+) Texture orientation histogram	13	(+) Ellipse area std
4	(+) Lightness composition	14	(-) Hue composition
5	(-) Curvature	15	(+) Lightness composition
6	(-) Chroma composition	16	(+) Color histogram
7	(-) Texture orientation histogram	17	(+) Hue composition
8	(+) Texture orientation histogram	18	(+) Chroma composition
9	(+) Relative color composition	19	(+) Curvature
10	(+) Relative color composition	20	(+) Color histogram

5.6 Summary

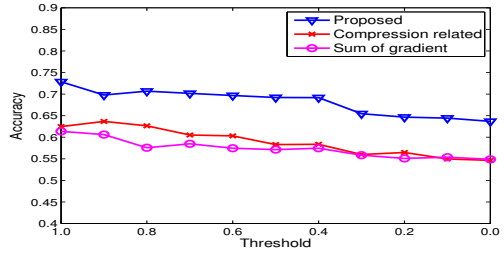
In this chapter, we evaluated the role of complexity played in aesthetic assessment and verified the Berlyne's inverted-U curve on thousands of photos through computational methods. We collected human labels on complexity for photographs from different categories. We found that human beings' judgement on complexity levels are congruous, hence complexity levels of photo is measurable. Then we trained the proposed VCHC and VCHA features into complexity models, and calculated complexity level for large-scale photo database to explore the relationship between beauty expectation and complexity level. Our analysis confirmed the ascending part of Berlyne's inverted-U curve and the importance of complexity in aesthetic assessment. The proposed visual complexity features are proved to be efficient in both beauty prediction and quality classification tasks.



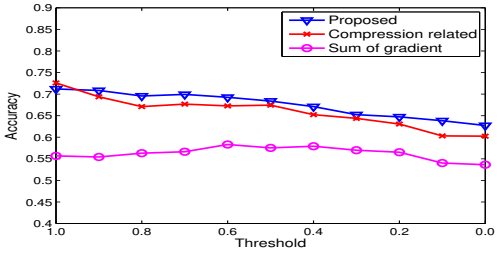
(A) "Animal"



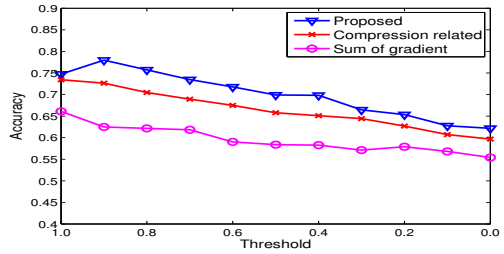
(B) "Architecture"



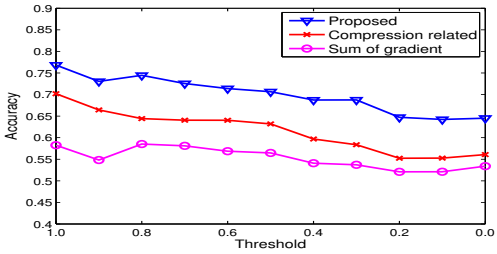
(C) "Cityscape"



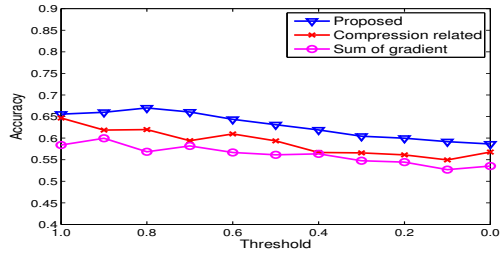
(D) "Floral"



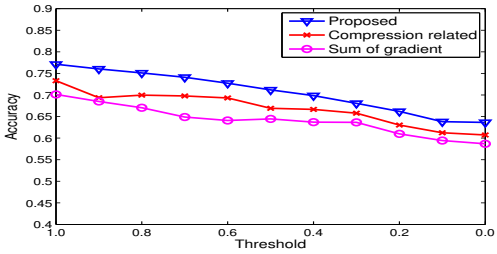
(E) "Fooddrink"



(F) "Landscape"

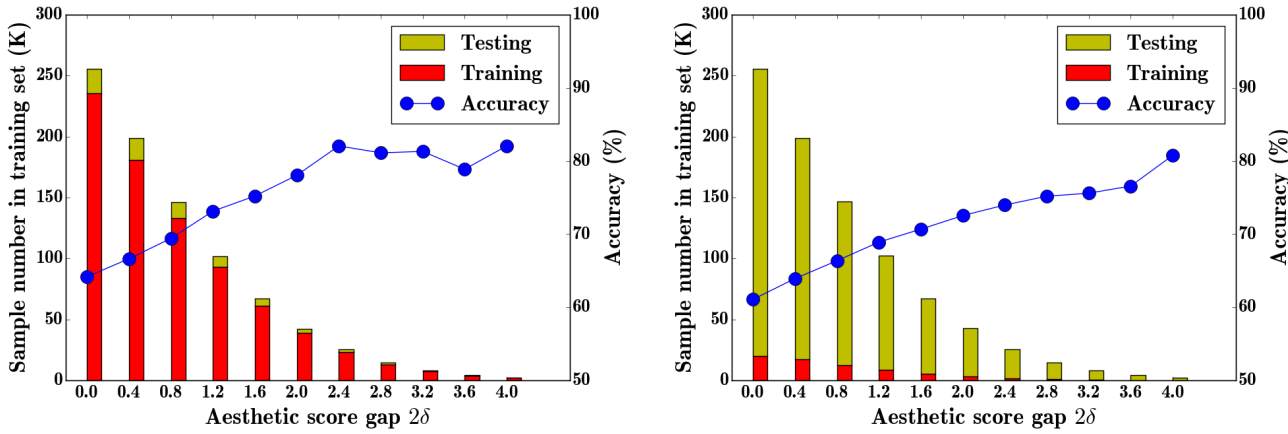


(G) "Portrait"



(H) "Stilllife"

FIGURE 5.8: Performance comparisons on high/low-quality classification task.



(A) AVA20Km (B) AVA230Km

FIGURE 5.9: Performance of VCHA features when the aesthetic score gap changed 2δ. The sample numbers in the training and testing set are shown by the bars.

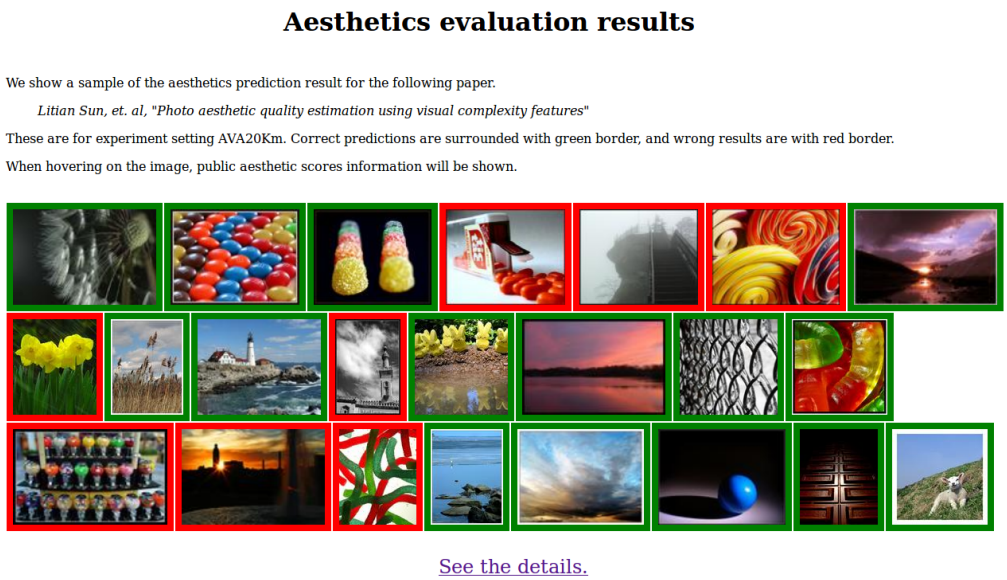
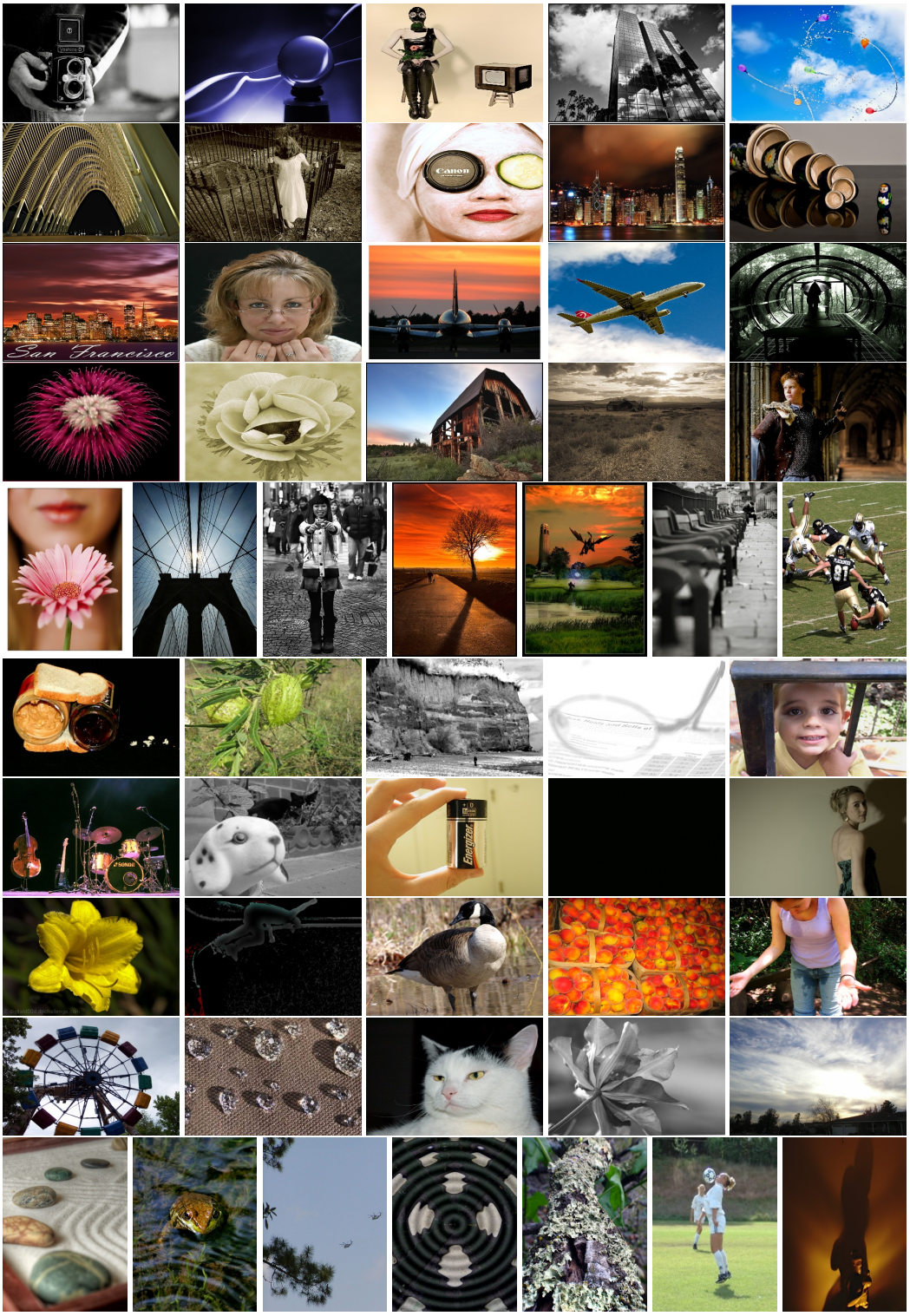


FIGURE 5.10: Screenshot of the web page showing the prediction results.

H



L

FIGURE 5.11: Sample photographs predicted correctly by VCHA for AVA40Kh. H: High quality photographs. L: Low quality photographs.

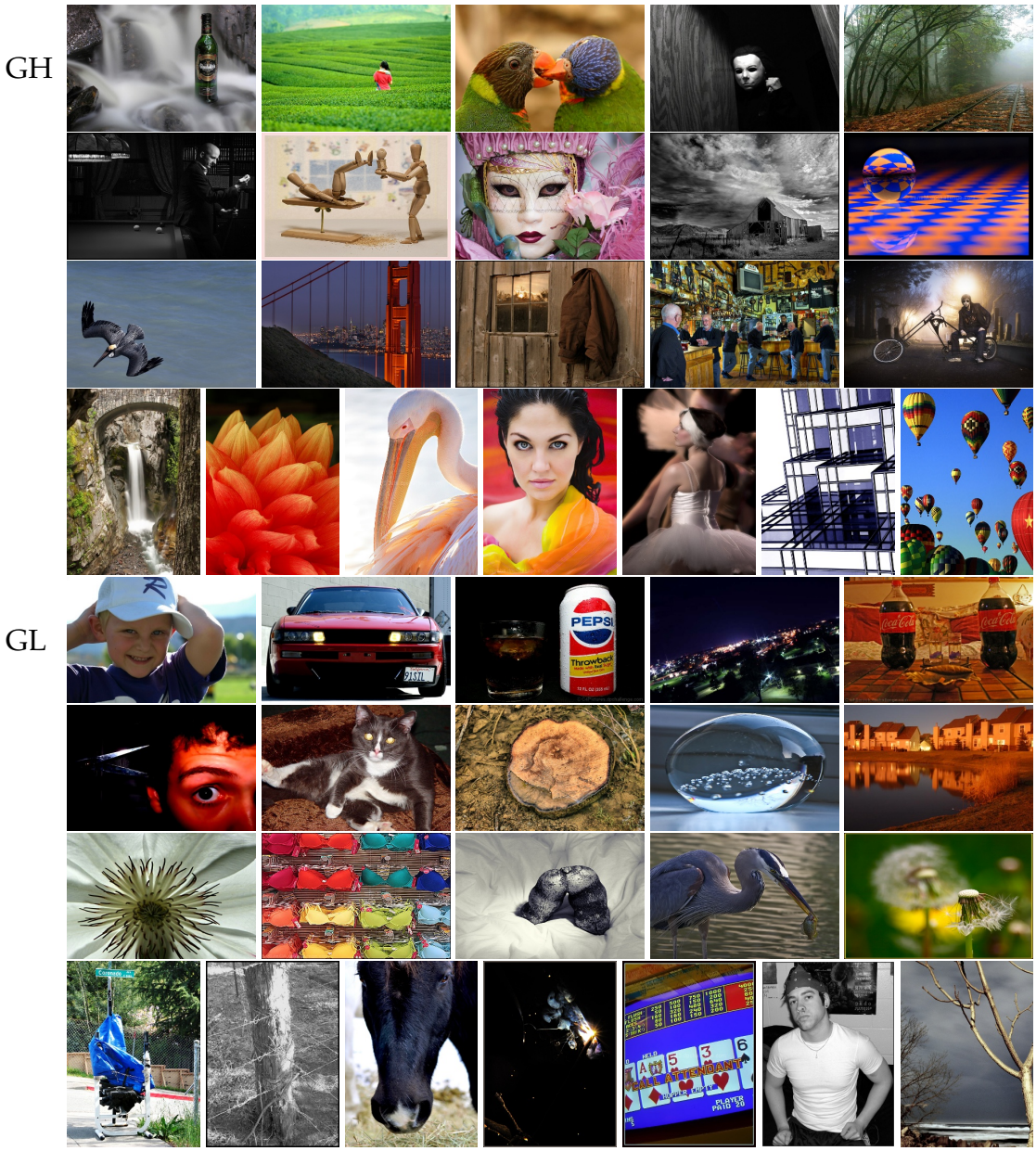


FIGURE 5.12: Sample photographs predicted wrongly by VCHA for AVA40Kh. GH: Ground truth is High quality but predicted as of low quality. GL: Ground truth is Low quality but predicted as of high quality.

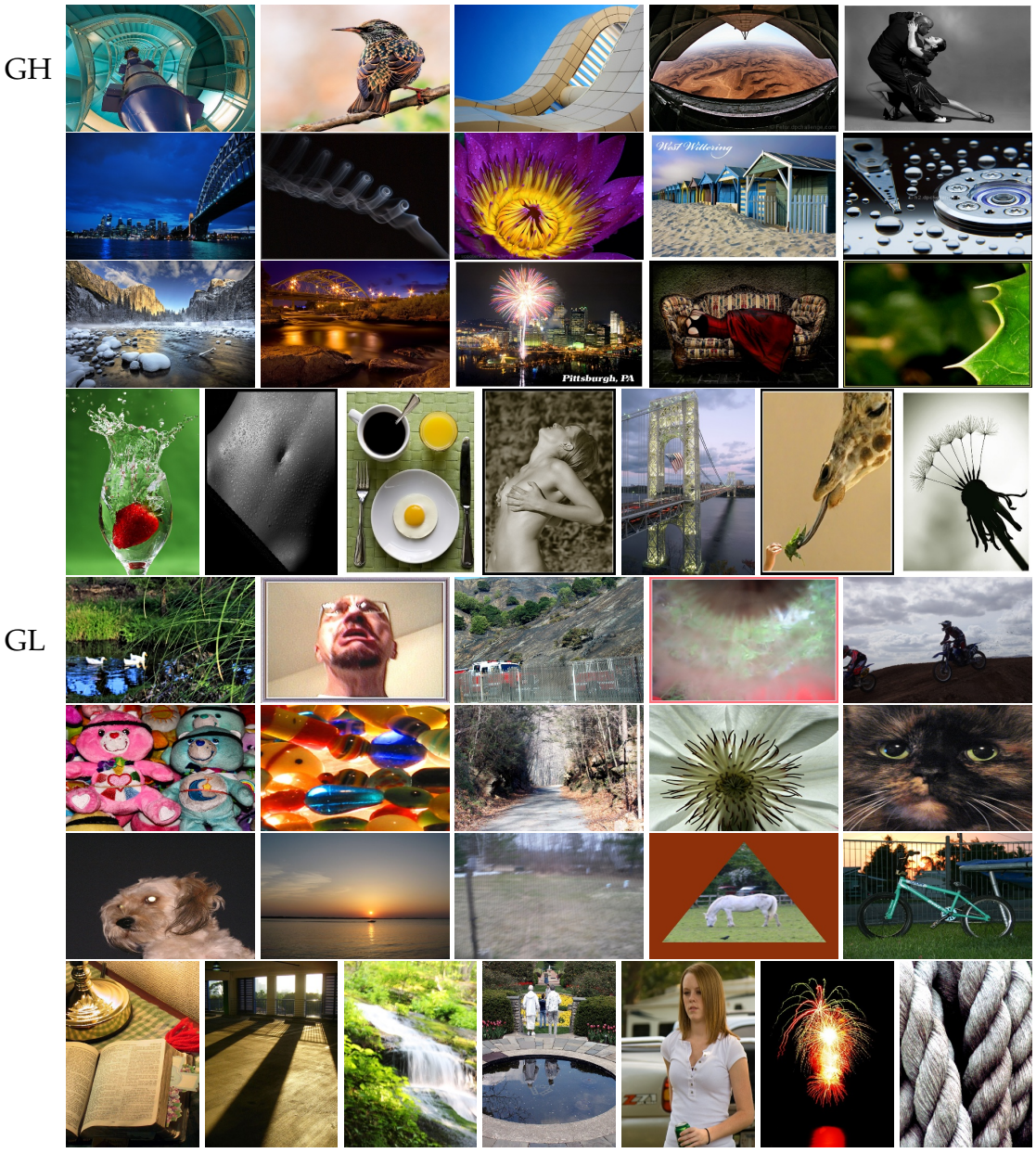


FIGURE 5.13: Sample photographs predicted wronly by VCHA but correctly by VCPC for AVA11KC. GH: Ground truth is High quality. GL: Ground truth is Low quality.

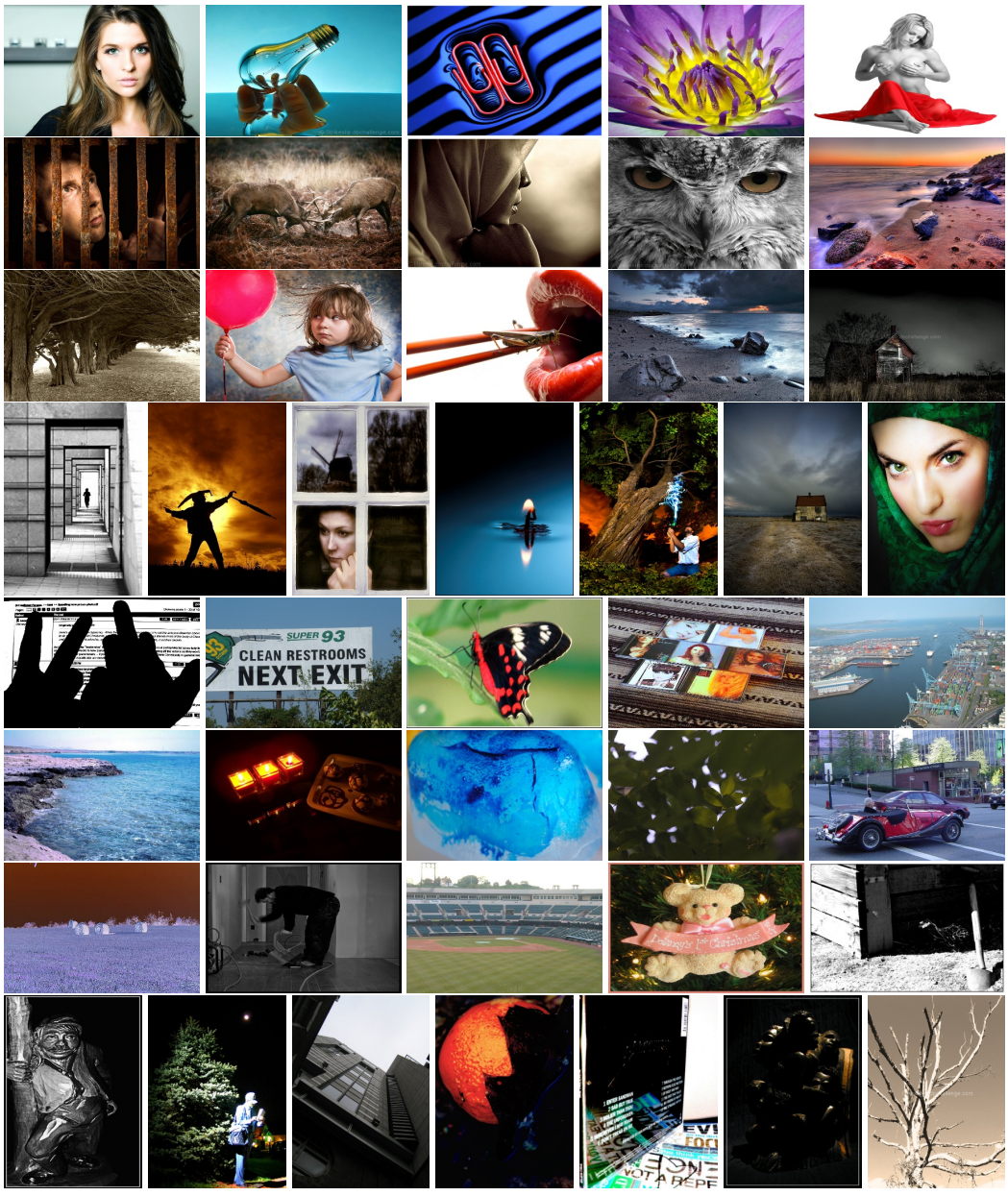


FIGURE 5.14: Sample photographs predicted wronly by VCPC but correctly by VCHA for AVA11KC. GH: Ground truth is High quality. GL: Ground truth is Low quality.

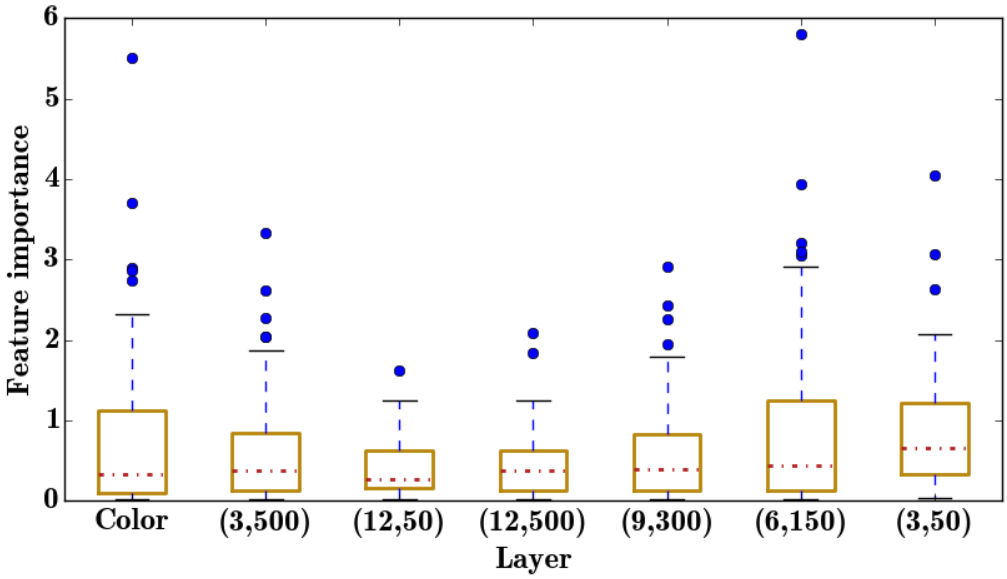


FIGURE 5.15: Boxplot of VCHA features for color and hierarchical abstraction layers.

Chapter 6

Video lectures personal preference prediction by gaze information

6.1 Introduction

As the recent explosion in massive online open courses (MOOC), predicting the viewers impression on the video talks is becoming more and more important.

Previous works made great efforts to understand the content of the video lectures [109], improve the quality of video lectures and the learning performance of students [26, 40, 39], and recommend possibly preferred videos based on viewer's historical or social behaviours [63, 53, 15]. However, viewer's preference for video lectures is more difficult to profile due to the limited number of viewer's historical behaviour and available materials. In addition, unlike the entertainment purpose videos, viewers tend to watch the lecture videos until it ends and the ratings not only depends on viewer's knowledge background and experience but also is closely related with their mental status and viewing environment, which are not predictable from viewer's profile.

EEG [111] and fMRI [84] are widely used to monitor and predict internal brain dynamics status. Comparing with EEG and fMRI, eye movement is believed to be a cognitive indicator that could be non-invasively collected with

low cost and compact devices. Gaze information is used to infer visualization task and user cognitvie abilities in [92], and gaze pattern would differ in decision-making task and a search task as reported in [25].

Recently, gaze is applied as an indicator in more subtle cognitive processes, such as personal preference profiling and artwork appreciation. The work in [108] uses eye tracking results to build a recommendation system for online multimedia materials. As suggested in [11], viewers' eye movements during the process of appreciating artworks are closely related with the attractiveness of the artworks. And gaze features are proved to be accurate to predict viewer's preference for images under controlled viewing environment by [94]. In [77], eye tracking method is applied to explore how the viewers' previous training is related to their aesthetic viewing in various interactions. Gaze is also used to infer user's latent interest [76] and language expertise [43]. All of these works suggest that gaze information could serve as an efficient indicator of cognitive process, including viewer's knowledge background, previous experience, mental status, emotion, and etc. In this work we step further to deal with individual rating for on-line video lectures by using gaze.

Conventional gaze features refer to statistics of the two event types of gaze, fixation and saccades. Features such as fixation count and duration, saccades velocity and angle, and etc., are widely used to characterize viewer's cognitive process [92, 107, 45]. Considering the lack of the analysis of the original content in the statistics features, more works use pre-defined areas of interest (AOIs) derived from the content, and convert gaze points to the sequence of AOI hits. For example, in [11] image is divided into two or three AOIs, and the transition entropy is calculated from the gaze point shifts between AOIs. The dynamic video stimuli make the AOI identification much more challenging. One method would be to determine the AOIs by clustering gaze points as described in [44].

In this chapter we use features extracted from gaze information to predict

individual rating for video talks. First a small-scale experiment is conducted to collect eye movements during watching TED Talks together with viewer's preference measured in rating values. Then a system is developed, especially with a set of gaze features, to predict individual preference for video lectures. To the best of our knowledge, this is the first work to predict personal preference for video lectures using gaze information.

The reminder of this chapter is organized as follows. The proposed features are described in detail in Section 6.2, the experiment is introduced in Section 6.3, and the analysis of the gaze and rating prediction are described in Section 6.4. Finally conclusions are given in Section 6.5.

6.2 Gaze features

We focus on two categories of features, one of which is the gaze statistics ($f_{gzstats}$) and the other is visual saliency related features (f_{sc}). Gaze statistics features focus on the information that eye movements provides and extract mental status indicators such as fixation duration, saccades length and etc. On the other hand, saliency related features focus on the difference between the visual saliency and viewer's actual gaze points.

6.2.1 Gaze statistics features

Gaze statistics is designed to include three parts; fixation and saccades statistics, and shift length. Besides the events (fixation and saccades) defined by the Tobii I-VT fixation filter [98], we also investigate the shift between different events, for instance, shift from one fixation to another fixation, or from fixation to saccades. Statistics features for fixation and saccades include average and standard deviation for event count, event duration and position. As for the shift, we mainly measure the length of shifts within a specific time period,

TABLE 6.1: Summary of gaze statistics features

Category	Details	Dimensions
Fixation	Count	1
	Duration (mean and std)	2
	Position at x and y axis (mean and std)	4
Saccades	Count	1
	Duration (mean and std)	2
	Position at x and y axis (mean and std)	4
Shift length	Histogram (10-bins)	10
	Statistics (mean, std, max and min)	4
Total		28

we calculated the histogram of shift length and also other statistics. The gaze statistics features are summarized as in Table 6.1.

Videos are treated as a collection of shots, and gaze statistics features are extracted from each shot. Then shot features for training set is clustered by a k-means clustering method and represented using its cluster index. In this way, the gaze for each video is represented by the histogram of shot features, which in short is a shot-based Bag-of-Feature (BoF) representation. Videos are divided into shots based on motion and color differences between frames [88]. Most shots are in the range of 30 to 400 frames, while some shots consist of up to 1000 frames. Figure 6.1 shows the statistics of shots for different videos.

As the length of shots are different, values in the features are normalized according to the shot length. Considering that the features as shown in Table 6.1 have different scales, for example fixation duration and gaze point position are measured in completely different scales, and k-means clustering uses Euclidean distance, we first normalize all gaze statistics features to the range of $[0, 1]$, then conduct the clustering calculation. Besides, the number of clusters is determined in a way that each cluster contain 30 samples on average. Because

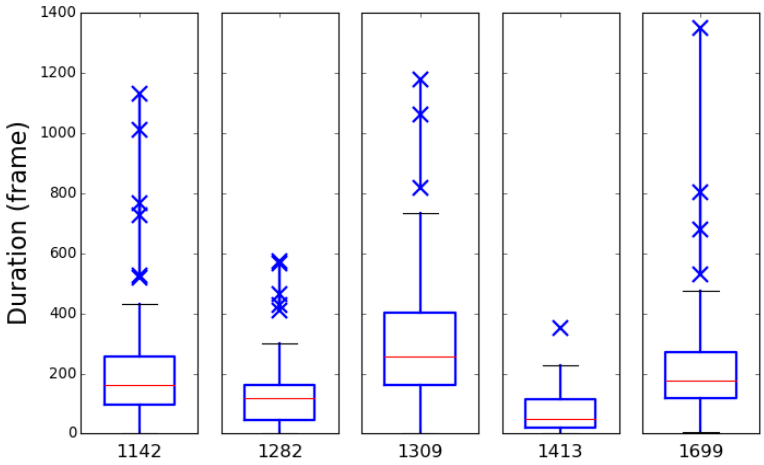


FIGURE 6.1: Boxplot of shot length. The numbers below are IDs in TED Talks.

the number of shots within each video is different, the histogram representations are normalized according to video and shot length.

6.2.2 Saliency related features

To represent the relationship between visual saliency and actual viewer's eye movements, we propose a feature, saliency hit f_{schit}. For each gaze point, a visual saliency map I_{sc} is generated from its corresponding frame using Nick's Machine Perception Toolbox (NMPT) [69, 112], which reflects the expected attention region based purely on visual cues including color, brightness, orientation, human faces and motion. The basic idea for the saliency relationship feature is shown in Figure 6.2. For a single gaze point located at C_{gz}, we first determine a set of circular masks M_i, as shown in the left of Figure 6.2. The intensity of the mask is defined according to the distance from the center gaze point as

$$M_i(d) = \begin{cases} 1 & \text{if} \quad r \times i \leq |d - C_{gz}| < r \times (i+1) \\ 0 & \text{else} \end{cases} \quad , i \in [0, N] \tag{6.1}$$

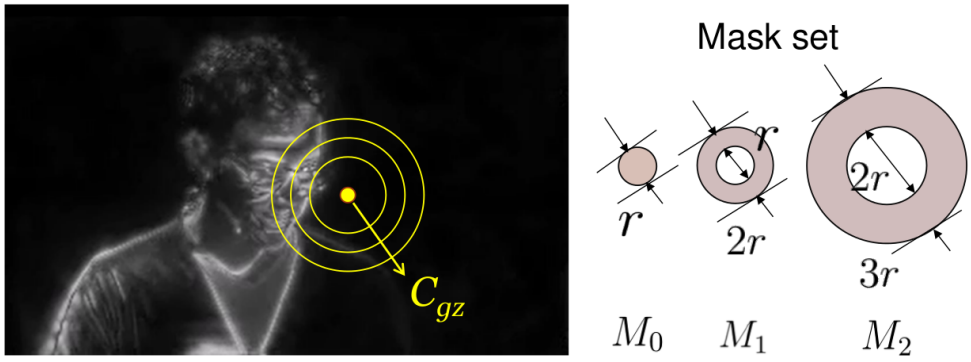


FIGURE 6.2: Illustration of the saliency hit feature.

where $M_i(d)$ indicates the value of M_i at location d, and r is the diameter unit. The diameter r of the smallest mask M_0 is determined according to Tobii I-VT fixation filter, so that all gaze points belonging to the same fixation fall into this smallest mask region. The number of masks is set in such a way that the largest mask could cover a quarter of the screen, which make the outer diameter of the largest mask exceed half of the screen width. In this work, the diameter r is set as 120 pixel, and the number of mask is seven, so f_{schit} feature is 7-dimension vectors. Then the saliency hit feature is calculated as the average saliency intensity within the overlapping region of each circular mask and the saliency map. The saliency hit feature f_{schit} is calculated as equation 6.2.

$$f_{schit} = \left\{ f_i | f_i = avg\left(M_i \bigcap I_{sc}\right), i \in [0, N] \right\} \tag{6.2}$$

Comparing with the gaze statistics features, saliency hit feature treat the gaze points as the center and takes the surrounding visual saliency cues into consideration. Thus it could reflect how the gaze points are attracted by the visual saliency cues and sparser saliency hit features may lead to a wondering mental status or an attention influenced more by audio rather than visual features.

Besides the analysis of static relationship between visual saliency and gaze point, i.e., saliency hit feature, we also propose another feature focusing on the

dynamic aspect the the relationship between visual saliency and gaze points, saliency hit transition (f_{schit_trs}). We first group the saliency hit vectors f_{schit} into clusters $\{C_i, i \in [1, K_{schit}]\}$. And each saliency hit vector is represented by its cluster center. Eye movements between each pair of gaze point is represented as the encoded transition from one saliency hit cluster to another cluster, as shown in equation 6.3. Histogram of the encoded transition is calculated to represent the saliency hit transition feature for a specific time period.

$$\begin{aligned} f_{schit\_trs}\left(t \rightarrow t+1\right) &= E\left(f_{schit}\left(t\right), f_{schit}\left(t+1\right)\right) \\ &= ID_{p \rightarrow q} \\ f_{schit}\left(t\right) \in C_p, \quad f_{schit}\left(t+1\right) \in C_q, \quad p, q \in [1, K_{schit}] \end{aligned} \quad , \tag{6.3}$$

The saliency related features are calculated from gaze points one by one, and we treat the video as one shot. The final representation is also in the BoF format. We set the saliency hit features cluster number K_{schit} as 100. In this way, the BoF representation of video using f_{schit} feature is 100-dimension vectors, so the saliency hit transition feature f_{schit_trs} is 10000 ($100 \times 100 = 10000$). And we set the cluster number for f_{schit_trs} also 100, and make the final BoF representation using f_{schit_trs} also 100-dimension.

6.3 Experiment setup

There are 649 English videos in the TED Talk channel on YouTube. To measure the general preference, we propose a preference indicator, which is the log of the ratio between the voting number of "like" and "dislike for each video on YouTube, as shown in equation 6.4.

$$P_{vid} = log\left(\frac{N_{like}}{N_{dislike}}\right) \tag{6.4}$$

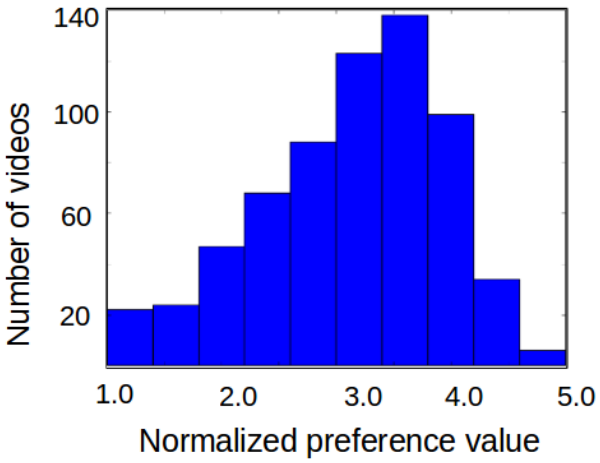


FIGURE 6.3: Distribution of public preference.

We normalized the preference indicators to the range of $[1, 5]$ in order to compare with the ratings collected in our experiment. "1" refers to relatively bad quality and disliked, and "5" refers to relatively high quality and liked videos. Although all TED Talks have more "like" voting than "dislike", not all of them share the same high preference level. As shown in Figure 6.3, the distribution of the public preference is similar to a normal distribution. We take the median normalized preference value as threshold and split the 649 video into high or low public preference levels.

To cover the various range of TED Talks, and to exploit gaze as indicator of viewers' mental process to the maximum extent, five videos are carefully selected. We chose talks that contains more multimedia content, for example slides, movie clips, and etc., rather than music and pure presenter. And these videos belong to different preference levels so that the possibility that participants in our experiment give different ratings could be high. Detailed information about the videos used in our experiment is summarized as in Table 6.2, two of which are of high preference, other three are of lower preference level. And the topics of the talks vary from technology to design.

Eight participants (two female and six male, aged from 21 to 38) are invited

TABLE 6.2: Video information.

TED ID	P_{vid}	Length	Tags
1142	High	9min40s	biology, biomimicry, design, green, invention, life, medicine, science
1282	Low	3min48s	creativity, design, entertainment, storytelling
1309	Low	10min	medical research, medicine, science, sight
1413	Low	4min	architecture, design, smell
1699	High	7min20s	agriculture, biodiversity

FIGURE 6.4: Illustration of the gaze collection experiment environment.

to this experiment. Although all of them are non-native English speaker, their listening comprehension level is enough to understand the TED Talks content.

Videos were rescaled to 1920×1080 resolution and shown on a 14 inch laptop. We used Tobii X2-60 eye tracker to track participants' eye movements during watching. And the viewing environment is illustrated as in Figure 6.4 As the videos vary from 5 minutes to 10 minutes, we conducted calibration before playing each video and asked the viewers to rest between videos. In order to eliminate the influence of watching order, the five videos were arranged randomly for different participants. After watching each video, participants were required to give their English comprehension level and concentration level in

TABLE 6.3: Viewers' rating summary.

TED ID	Participants								Avg	P_{vid}
1142	4	3	4	5	3	5	5	3	3.9	High
1282	4	5	5	4	4	4	3	4	4.1	Low
1309	2	4	3	4	4	4	5	3	3.6	Low
1413	4	2	5	3	3	3	3	2	3.1	Low
1699	4	5	5	5	4	5	4	5	4.6	High

percent, to rate the video within the range of $[1, 5]$, and choose up to three impression labels out of 14 which are the same as those on TED website[1].

6.4 Analysis

6.4.1 Collected data

A part of viewers' raw gaze points along x and y axes are shown as in Figure 6.5. Despite of a lot of similar patterns among viewers' eye movements, different gaze patterns could be clearly observed, such as pt_1 for x axis at the beginning of the shot. Fixations and saccades are determined by Tobii I-VT filter [98].

Participants' ratings for different videos are shown in Table 6.3. Although we allow the ratings in the range of 1 to 5, the minimum rating given by participants is 2, and in most situations are from 3 to 5. Individual rating values are quite different from the public preference. For example the video 1282 is generally disliked while the average ratings in our experiment shows that it is even better 1142.

6.4.2 Framework

The individual ratings as shown in Table 6.3 are binarized using the average value as the threshold. In this way, we manage to predict a positive or negative

[1]https://www.ted.com/talks

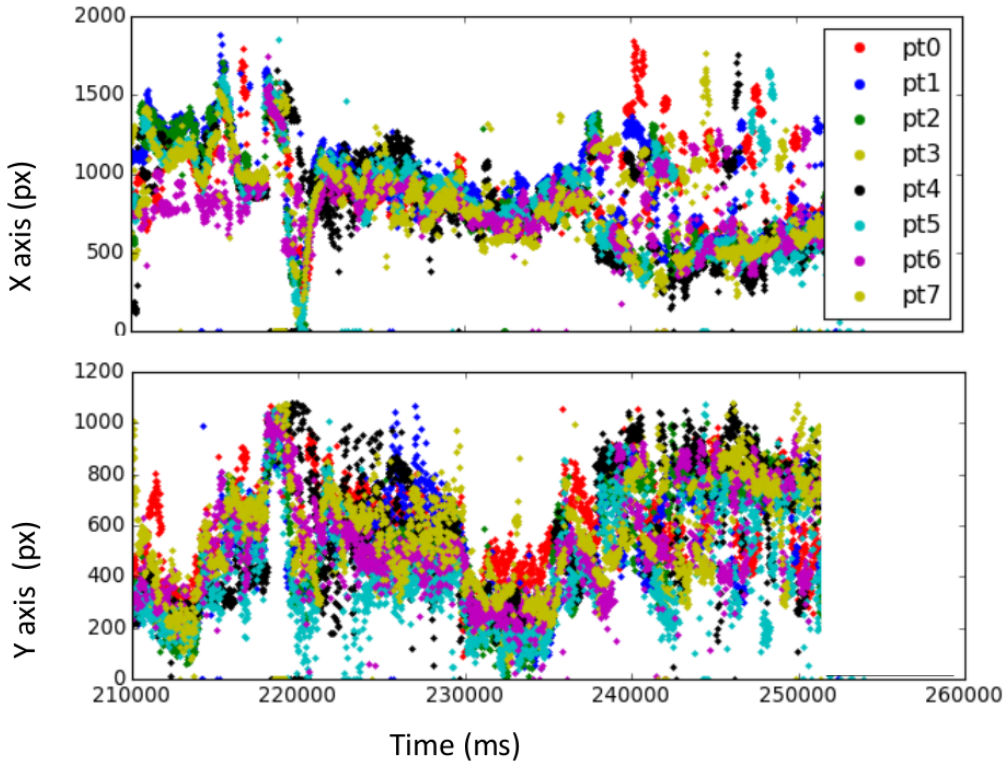


FIGURE 6.5: Example of gaze position for different participants.

impression for each viewer. Finally, the individual ratings are predicted using the SVM classifier.

6.4.3 Rating prediction

The total 40 samples are randomly split into training and testing set according to different test ratio from 0.1 to 0.5. We also did an additional leave-one-out test, which make the test ratio 0.025. For each test ratio setting we conducted the classification 200 times using the SVM classifier. The parameters of SVM classifier are optimized each time through grid search, including both RBF and linear kernel.

In order to show the effectiveness of the proposed features, the performance using random features and the chance level accuracy are also included. A 100-dimension random feature is generated for each sample (same dimension as saliency related features described in Section 6.2.2), and followed the same train

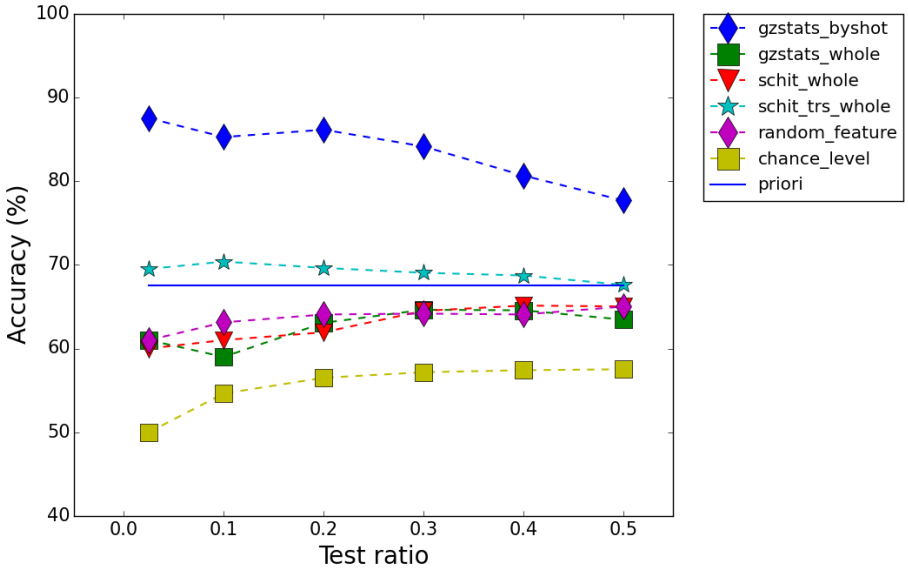


FIGURE 6.6: Classification accuracy for different features.

and test protocol as other gaze features. Chance level accuracy is calculated as comparison of the random prediction result with the ground truth. The difference of random feature performance and the chance level accuracy shows the gain from SVM classifier. As the numbers of positive and negative samples are not balanced, if we know the priori distribution of the sample and predict all the test samples as the majority value, then the accuracy would be 67.5% ($27/40 = 0.675$).

As shown in Figure 6.6, for all train-test-split settings, the proposed features could achieve around 20% improvement than chance level accuracy. The gaze statistics features using shot-based representation lead to the best performance, achieving over 85% accuracy in the leave-one-out test and over 75% in the half-half train-test-split setting, while using the same gaze statistics features but treat the video as single shot (shown as"gzstats_whole" in Figure 6.6) leads to no better performance than using random feature. The saliency hit transition feature also perform well and has higher accuracy than priori knowledge in almost all settings.

6.5 Summary

In this chapter, we conducted a small scale experiment to collect gaze information for video talk viewing process together with viewers' ratings of the talks. And we proposed a set of gaze features and used these to predict viewers' negative or positive impression of the video talks. Besides the conventional statistics features, we also analyzed the relationship between visual saliency and gaze points in both static and dynamic aspects. The analysis results showed that the proposed method could achieve for a over 85% accuracy in the leave-one-out test and over 75% in the half-half train-test-split setting.

Chapter 7

Conclusions

Concerning the large-scale on-line multimedia information, both photograph and videos, how people are appreciating these content? Is it possible that the computer can think in a similar way with us and predict viewer's impression? These are the questions that this thesis managed to answer.

We estimated the viewer impressions for photograph and video lectures by exploring the cognitive process. We proposed a series of complexity related features, verified the relationship between complexity and impression, and estimated the subjective impressions.

We will summarize this thesis in detail in the following and discuss the limitation of the current work together with future directions.

7.1 Summary

Different from the previous methods that designed features by mimicking the content production rules or the generic features that are proved to be efficient in object classification task, we proposed a series of complexity-related features based on psychological theories.

We predicted the subjective impressions towards photographs and videos through analysing both the content and the viewer. For photographs we mainly predicted the general aesthetic preference by analysing the visual complexity of

the photographs. And for video lectures, we predicted the individual viewer's preference using gaze information collected from the viewer during watching. Specifically, we implemented the following tasks:

1. **Proposed a series of complexity-related features.**

   We took the factors that are used in the psychological experiments to control the visual complexity level of the stimulus and proposed a set of visual complexity operators that can extract the complexity characteristics in composition, shape and distribution. And we proposed a set of VCPC features to apply the complexity operators to various perception cues.

   Furthermore, we explored the component of the complexity as "the hierarchical structure of the elements", and proposed another set of visual complexity features VCHA, which applies the visual complexity operators to hierarchical abstractions derived from the photograph. We used rolling guidance filter and by adjusting the spatial and intensity scales we divide the input photograph into a set of hierarchical abstractions, in which elements of different size and contrast are kept in different images.

   In addition, we noticed that viewer's cognitive load could be measured by the eye movement of the user viewing the content. We also extracted complexity-related features from gaze information to predict viewer's individual impression towards video lectures. We proposed two sets of features. One is the gaze statistics, which put the emphasize on indicators that are close to human reaction on complexity, such as the fixation duration and the saccades length. The other set of features analyses the deviation of the gaze points from the saliency map, which is the expected gaze from the producer's view.

2. **Verified the relationship between complexity and impression.**

Due to the limitations of the psychological theories, we verified the relationship between complexity and aesthetics in the following aspects: the vague definition of visual complexity, the applicable content range and the measurement methods.

We evaluated the role of complexity played in aesthetic assessment and verified the Berlyne's inverted-U curve on thousands of photos through computational methods. We collected human labels on complexity for photographs from different categories. We found that human beings' judgement on complexity levels are congruous, hence complexity levels of photo is measurable. Then we trained the proposed VCPC and VCHA features into complexity models, and calculated complexity level for large-scale photo database to explore the relationship between beauty expectation and complexity level. Our analysis confirmed the ascending part of Berlyne's inverted-U curve and the importance of complexity in aesthetic assessment. The proposed visual complexity features are proved to be efficient in both beauty prediction and quality classification tasks.

We conducted the verification on large-scale dataset by computational methods. We set up a visual complexity model using the proposed complexity features together with a complexity dataset we built. By comparing the visual complexity levels computed by our model with the pubic aesthetics ratings, we found that there are monotonic relationship in different categories. We observed the ascending trends in "Architecture", "Cityscape" and "Landscape" categories and descending trends in "Animal", "Floral" and "Fooddrink" categories.

3. **Used complexity-related features for impression prediction.**

   We conducted extensive experiments using the public aesthetics dataset

AVA, with unbalanced, balanced, categorized, generic, and consented subsets. The experimental results show that the proposed VCPC and VCHA features could very well predict the general aesthetic impressions. Under the balanced experiment settings (AVA40Kh), VCHA features give the highest prediction accuracy comparing with existing photography-role-based features, generic features (including deep features), complexity-related features. And for the consented samples, both VCPC and VCHA features have over 80% prediction accuracy.

As for the individual impression prediction for the video lectures, we applied the SVM classifier on the proposed two sets of features to predict whether the viewer like or dislike the content of the video lectures. The best performance is given by the gaze statistics feature. Our system has 85% accuracy for leave-one-out testing condition.

7.2 Future challenges

We overview some possible future directions of this thesis.

Visual complexity concerning semantic understanding is a interesting topic. As interpreted in the processing fluency theory [79, 78], the easier viewer could understand the content, the more they are prone to enjoy it. With the fast development of deep neural network and the increasing accuracy in object detection and recognition tasks, it would be easier to get more correct and detail analysis of the individual elements.

Further analysis of the relationship and interaction between the objects could be included into the visual complexity estimation. As different objects may have different influence on visual complexity due to their function and interaction with other objects. Such analysis is important to estimate the cognitive load that human brain is baring during viewing the photograph and the videos.

Besides, applications of the proposed methods will be of great help for content producers.

Based on the feature importance analysis in Section 5.5.2, we can know which features of a given photograph are contributing to a positive aesthetic impression, and which are leading to a negative one. Future work could be done to integrate the features into instructional suggestions, for example, "the main object should be put more left", "adjust the focal length to reduce the noise from the grass" and etc. Such instructions will greatly improve amateur photographers user experience and reward the users with more share and appreciations.

Viewer-aware lecture systems and feedback systems for lecturers are interesting applications. We divided the video lecture into a sequence of shots in Chapter 6. Gaze features for each shot could be analyzed to understand the cognitive load of the viewer. If the user is confused and feel the content too difficult, and we may find fixation points far away from salient area or other situations in gaze features, the system could automatic mark the shot as not completely understood scene, and take some actions, for example, replay the confusing shot, or put more weight in the homework or question sheet about such part. Gaze analysis of a group of viewers could form a summary feedback for the lecturer, about which part of the lecture is not well explained and making the viewers dislike the content.

Bibliography

[1] Aysu Akalin et al. "Architecture and engineering students' evaluations of house façades: Preference, complexity and impressiveness". In: *Journal of Environmental Psychology* 29.1 (2009), pp. 124–132.

[2] Pablo Arbelaez et al. "Contour detection and hierarchical image segmentation". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.5 (2011), pp. 898–916.

[3] Tunc Ozan Aydin, Aljoscha Smolic, and Markus Gross. "Automated aesthetic analysis of photographic images". In: *Visualization and Computer Graphics, IEEE Transactions on* 21.1 (2015), pp. 31–42.

[4] Barry J Babin and Jill S Attaway. "Atmospheric affect as a tool for creating value and gaining share of customer". In: *Journal of Business research* 49.2 (2000), pp. 91–99.

[5] Michael W Beauvois. "Quantifying aesthetic preference and perceived complexity for fractal melodies". In: *Music Perception* 24.3 (2007), pp. 247–264.

[6] Daniel E Berlyne. *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation.* Hemisphere, 1974.

[7] DE Berlyne. "Aesthetics and psychobiology, 1971". In: *Appleton-Century-Crofts, New York* ().

[8] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. "A framework for photo-quality assessment and enhancement based on visual aesthetics". In: *Proceedings of the international conference on Multimedia*. ACM. 2010, pp. 271–280.

[9] George David Birkhoff. *Aesthetic measure*. Cambridge, Mass., 1933.

[10] Allan Campbell, Vic Ciesielksi, and AK Qin. "Feature Discovery by Deep Learning for Aesthetic Analysis of Evolved Abstract Images". In: *Evolutionary and Biologically Inspired Music, Sound, Art and Design*. Springer, 2015, pp. 27–38.

[11] Daniel J Campbell et al. "Saliency-based Bayesian modeling of dynamic viewing of static scenes". In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM. 2014, pp. 51–58.

[12] Susan F Chipman. "Complexity and structure in visual patterns." In: *Journal of Experimental Psychology: General* 106.3 (1977), p. 269.

[13] Ritendra Datta, Jia Li, and James Z Wang. "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition". In: *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE. 2008, pp. 105–108.

[14] Ritendra Datta et al. "Studying aesthetics in photographic images using a computational approach". In: *Computer Vision–ECCV 2006*. Springer, 2006, pp. 288–301.

[15] James Davidson et al. "The YouTube video recommendation system". In: *Proceedings of the fourth ACM conference on Recommender systems*. ACM. 2010, pp. 293–296.

[16] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. "High level describable attributes for predicting aesthetics and interestingness". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 1657–1664.

[17] David M Diamond et al. "The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law". In: *Neural plasticity* 2007 (2007).

[18] Gwyneth Doherty-Sneddon and Fiona G Phelps. "Gaze aversion: A response to cognitive or social difficulty?" In: *Memory & Cognition* 33.4 (2005), pp. 727–733.

[19] Don C Donderi. "Visual complexity: a review." In: *Psychological Bulletin* 132.1 (2006), p. 73.

[20] Hans Jurgen Eysenck. "The experimental study of the'good Gestalt'a new approach." In: *Psychological Review* 49.4 (1942), p. 344.

[21] Alex Forsythe et al. "Predicting beauty: fractal dimension and visual complexity in art". In: *British Journal of Psychology* 102.1 (2011), pp. 49–70.

[22] Jay Friedenberg and Marco Bertamini. "Aesthetic Preference for Polygon Shape". In: *Empirical Studies of the Arts* 33.2 (2015), pp. 144–160.

[23] Mariano García, Albert N Badre, and John T Stasko. "Development and validation of icons varying in their abstractness". In: *Interacting with Computers* 6.2 (1994), pp. 191–211.

[24] Pascal WM van Gerven et al. "Modality and variability as factors in training the elderly". In: *Applied cognitive psychology* 20.3 (2006), pp. 311–320.

[25] Kerstin Gidlöf et al. "Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment". In: *Journal of Eye Movement Research* 6.1 (2013), pp. 1–14.

[26] Vijay N Gohokar and Mangal H Dhend. "Adaptive, cognitive and innovative tools & techniques for improving learning performance: A case study". In: *MOOC, Innovation and Technology in Education (MITE), 2014 IEEE International Conference on*. IEEE. 2014, pp. 47–52.

[27] Xiaoying Guo et al. "Analysis of texture characteristics associated with visual complexity perception". In: *Optical review* 19.5 (2012), pp. 306–314.

[28] Simon Harper, Eleni Michailidou, and Robert Stevens. "Toward a definition of visual complexity as an implicit measure of cognitive load". In: *ACM Transactions on Applied Perception (TAP)* 6.2 (2009), p. 10.

[29] Xiao-Chen He and Nelson HC Yung. "Curvature scale space corner detector with adaptive threshold and dynamic region of support". In: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 2. IEEE. 2004, pp. 791–794.

[30] Christopher Heaps and Stephen Handel. "Similarity and features of natural textures." In: *Journal of Experimental Psychology: Human Perception and Performance* 25.2 (1999), p. 299.

[31] Tom Heath, Sandy G Smith, and Bill Lim. "Tall Buildings and the Urban Skyline The Effect of Visual Complexity on Preferences". In: *Environment and Behavior* 32.4 (2000), pp. 541–556.

[32] Donald Olding Hebb. "Drives and the CNS (conceptual nervous system)." In: *Psychological review* 62.4 (1955), p. 243.

[33] Bruce A Huhmann. "Visual complexity in banner ads: The role of color, photography, and animation". In: *Visual Communication Quarterly* 10.3 (2003), pp. 10–17.

[34] Shintchi Ichikawa. "Quantitative and structural factors in the judgment of pattern complexity". In: *Perception & psychophysics* 38.2 (1985), pp. 101–109.

[35] Çagri Imamoglu. "COMPLEXITY, LIKING AND FAMILIARITY: ARCHITECTURE AND NON-ARCHITECTURE TURKISH STUDENTS' ASSESSMENTS OF TRADITIONAL AND MODERN HOUSE FACADES". In: *Journal of Environmental Psychology* 20.1 (2000), pp. 5–16.

[36] Yangqing Jia et al. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *arXiv preprint arXiv:1408.5093* (2014).

[37] Bruce F Katz. "What makes a polygon pleasing?" In: *Empirical studies of the Arts* 20.1 (2002), pp. 1–19.

[38] Yan Ke, Xiaoou Tang, and Feng Jing. "The design of high-level features for photo quality assessment". In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 1. IEEE. 2006, pp. 419–426.

[39] Juho Kim et al. "Data-driven interaction techniques for improving navigation of educational videos". In: *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM. 2014, pp. 563–572.

[40] Joris Klerkx, Katrien Verbert, and Erik Duval. "Enhancing learning with visualization techniques". In: *Handbook of Research on Educational Communications and Technology*. Springer, 2014, pp. 791–807.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[42] BM Kudielka et al. "HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: impact of age and gender". In: *Psychoneuroendocrinology* 29.1 (2004), pp. 83–98.

[43] Kai Kunze et al. "Towards inferring language expertise using eye tracking". In: *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2013, pp. 217–222.

[44] Kuno Kurzhals, Florian Heimerl, and Daniel Weiskopf. "ISeeCube: Visual analysis of gaze data for video". In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM. 2014, pp. 43–50.

[45] Laura N van der Laan et al. "Do you like what you see? The role of first fixation and total fixation duration in consumer choice". In: *Food Quality and Preference* 39 (2015), pp. 46–55.

[46] Helmut Leder et al. "A model of aesthetic appreciation and aesthetic judgments". In: *British journal of psychology* 95.4 (2004), pp. 489–508.

[47] Guo Lihua and Li Fudi. "Image aesthetic evaluation using paralleled deep convolution neural network". In: *arXiv preprint arXiv:1505.05225* (2015).

[48] Kuo-Yen Lo, Keng-Hao Liu, and Chu-Song Chen. "Assessment of photo aesthetics with efficiency". In: *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE. 2012, pp. 2186–2189.

[49] Hong Lu et al. "Leveraging Color Harmony and Spatial Context for Aesthetic Assessment of Photographs". In: *Advances in Multimedia Information Processing–PCM 2014*. Springer, 2014, pp. 323–332.

[50] Peng Lu et al. "Towards aesthetics of image: A Bayesian framework for color harmony modeling". In: *Signal Processing: Image Communication* (2015).

[51] Xin Lu et al. "Deep Multi-Patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 990–998.

[52] Xin Lu et al. "Rapid: Rating pictorial aesthetics using deep learning". In: *Proceedings of the ACM International Conference on Multimedia*. ACM. 2014, pp. 457–466.

[53] Hangzai Luo et al. "Personalized news video recommendation". In: *Advances in Multimedia Modeling*. Springer, 2009, pp. 459–471.

[54] Wei Luo, Xiaogang Wang, and Xiaoou Tang. "Content-based photo quality assessment". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 2206–2213.

[55] Yiwen Luo and Xiaoou Tang. "Photo and video quality evaluation: Focusing on the subject". In: *Computer Vision–ECCV 2008*. Springer, 2008, pp. 386–399.

[56] Sonia J Lupien et al. "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition". In: *Brain and cognition* 65.3 (2007), pp. 209–237.

[57] Birgit Mallon, Christoph Redies, and Gregor U Hayn-Leichsenring. "Beauty in abstract paintings: perceptual contrast and statistical properties". In: *Frontiers in human neuroscience* 8 (2014).

[58] Luca Marchesotti et al. "Assessing the aesthetic quality of photographs using generic image descriptors". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 1784–1791.

[59] Manuela M Marin and Helmut Leder. "Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music". In: *PloS one* 8.8 (2013), e72412.

[60] Lucy Markson and Kevin B Paterson. "Effects of gaze-aversion on visual-spatial imagination". In: *British Journal of Psychology* 100.3 (2009), pp. 553–563.

[61] José Antonio Martín H, Matilde Santos, and Javier de Lope. "Orthogonal variant moments features in image analysis". In: *Information Sciences* 180.6 (2010), pp. 846–860.

[62] John W Mason. "A review of psychoendocrine research on the sympathetic-adrenal medullary system." In: *Psychosomatic medicine* 30.5 (1968), pp. 631–653.

[63] Tao Mei et al. "VideoReach: an online video recommendation system". In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2007, pp. 767–768.

[64] Harry Munsinger and William Kessen. "Uncertainty, structure, and preference." In: *Psychological Monographs: General and Applied* 78.9 (1964), p. 1.

[65] Naila Murray, Luca Marchesotti, and Florent Perronnin. "AVA: A large-scale database for aesthetic visual analysis". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2408–2415.

[66] Marcos Nadal et al. "The influence of visual complexity on aesthetic preference: An explanation of diverging results". In: *Proceedings of IAEA08* (2008), pp. 137–141.

[67] Marcos Nadal et al. "Visual complexity and beauty appreciation: Explaining the divergence of results". In: *Empirical Studies of the Arts* 28.2 (2010), pp. 173–191.

[68] JACK L NASAR. "What design for a presidential library? Complexity, typicality, order, and historical significance". In: *Empirical Studies of the Arts* 20.1 (2002), pp. 83–99.

[69] *Nick's Machine Perception Toolbox*. http://mplab.ucsd.edu/~nick/NMPT/.

[70] Masashi Nishiyama et al. "Aesthetic quality classification of photographs based on color harmony". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 33–40.

[71] Werner Nohl. "Sustainable landscape use and aesthetic perception–preliminary reflections on future landscape aesthetics". In: *Landscape and urban planning* 54.1 (2001), pp. 223–237.

[72] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. "Color compatibility from large datasets". In: *ACM Transactions on Graphics (TOG)*. Vol. 30. 4. ACM. 2011, p. 63.

[73] A Mack Oliva, ML Shrestha, and M Peeper. "A.(2004)". In: *Identifying the perceptual dimensions of visual complexity of scenes Proceedings of the 26th Annual Meeting of the Cognitive Science Society Meeting. Chicago*.

[74] Fred GWC Paas and Jeroen JG Van Merriënboer. "The efficiency of instructional conditions: An approach to combine mental effort and performance measures". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 35.4 (1993), pp. 737–743.

[75] Letizia Palumbo and Marco Bertamini. "The Curvature Effect A Comparison Between Preference Tasks". In: *Empirical Studies of the Arts* 34.1 (2016), pp. 35–52.

[76] Hye-Sun Park, Takatsugu Hirayama, and Takashi Matsuyama. "Gaze Mirroring-based Intelligent Information System for Making User's Latent Interest". In: *Journal of Intelligence and Information Systems* 16.3 (2010), pp. 37–54.

[77] Juyeon Park, Emily Woods, and Marilyn DeLong. "Quantification of aesthetic viewing using eye-tracking technology: the influence of previous training in apparel design". In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM. 2010, pp. 153–155.

[78] Rolf Reber. "Processing fluency, aesthetic pleasure, and culturally shared taste". In: *Aesthetic science: Connecting minds, brains, and experience* (2012), pp. 223–249.

[79] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. "Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience?" In: *Personality and social psychology review* 8.4 (2004), pp. 364–382.

[80] Christoph Redies et al. "PHOG-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects". In: *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer. 2012, pp. 522–531.

[81] Ranier Reisenzein. "Pleasure-arousal theory and the intensity of emotions." In: *Journal of Personality and Social Psychology* 67.3 (1994), p. 525.

[82] Jaume Rigau, Miquel Feixas, and Mateu Sbert. "Informational aesthetics measures". In: *IEEE Computer Graphics and Applications* 28.2 (2008), pp. 24–34.

[83] Juan Romero et al. "Using complexity estimates in aesthetic image classification". In: *Journal of Mathematics and the Arts* 6.2-3 (2012), pp. 125–136.

[84] Bruno Rossion, Bernard Hanseeuw, and Laurence Dricot. "Defining face perception areas in the human brain: a large-scale factorial fMRI face localizer analysis". In: *Brain and cognition* 79.2 (2012), pp. 138–157.

[85] James A Russell. "Evidence of convergent validity on the dimensions of affect." In: *Journal of personality and social psychology* 36.10 (1978), p. 1152.

[86] DH Saklofske. "Visual aesthetic complexity, attractiveness and diversive exploration". In: *Perceptual and motor skills* 41.3 (1975), pp. 813–814.

[87] Bo N Schenkman and Fredrik U Jönsson. "Aesthetics and preferences of web pages". In: *Behaviour & Information Technology* 19.5 (2000), pp. 367–377.

[88] *Shot detect master*. https://api.travis-ci.org/johmathe/Shotdetect.svg?branch=master.

[89] Florian Simond, Nikolaos Arvanitopoulos Darginis, and Sabine Süsstrunk. "Image Aesthetics Depends on Context". In: *IEEE Proceedings of the International Conference on Image Processing*. EPFL-CONF-212967. 2015.

[90] Dean Keith Simonton. "Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets". In: *Computers and the Humanities* 24.4 (1990), pp. 251–264.

[91] Arthur E Stamps III. "Entropy, visual diversity, and preference". In: *The Journal of general psychology* 129.3 (2002), pp. 300–320.

[92] Ben Steichen, Giuseppe Carenini, and Cristina Conati. "User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities". In: *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM. 2013, pp. 317–328.

[93] Hsiao-Hang Su et al. "Scenic photo quality assessment with bag of aesthetics-preserving features". In: *Proceedings of the 19th ACM international conference on Multimedia*. ACM. 2011, pp. 1213–1216.

[94] Yusuke Sugano et al. "Image preference estimation with a data-driven approach: A comparative study between gaze and image features". In: *Journal of Eye Movement Research* 7.3 (2014), p. 5.

[95] John Sweller. "Cognitive load during problem solving: Effects on learning". In: *Cognitive science* 12.2 (1988), pp. 257–285.

[96] Xinmei Tian et al. "Query-Dependent Aesthetic Model With Deep Learning for Photo Quality Assessment". In: *Multimedia, IEEE Transactions on* 17.11 (2015), pp. 2035–2048.

[97] Pablo PL Tinio and Helmut Leder. "Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization". In: *Acta psychologica* 130.3 (2009), pp. 241–250.

[98] *Tobii I-VT fixation filter*. http://www.tobii.com/eye-tracking-research/global/library/white-papers/the-tobii-i-vt-fixation-filter/.

[99] Carlo Tomasi and Roberto Manduchi. "Bilateral filtering for gray and color images". In: *Computer Vision, 1998. Sixth International Conference on*. IEEE. 1998, pp. 839–846.

[100] Alexandre N Tuch et al. "The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments". In: *International Journal of Human-Computer Studies* 70.11 (2012), pp. 794–811.

[101] Alexandre N Tuch et al. "Visual complexity of websites: Effects on users experience, physiology, performance, and memory". In: *International Journal of Human-Computer Studies* 67.9 (2009), pp. 703–715.

[102] José Van Dijck. "Memory matters in the digital age". In: *Configurations* 12.3 (2004), pp. 349–373.

[103] PWM Van Gerven et al. "Cognitive load theory and aging: Effects of worked examples on training efficiency". In: *Learning and Instruction* 12.1 (2002), pp. 87–105.

[104] Aldert Vrij et al. "Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order". In: *Law and human behavior* 32.3 (2008), pp. 253–265.

[105] Cuong T Vu, Thien D Phan, and Damon M Chandler. "A spectral and spatial measure of local perceived sharpness in natural images". In: *Image Processing, IEEE Transactions on* 21.3 (2012), pp. 934–945.

[106] Kirk L Wakefield and Julie Baker. "Excitement at the mall: determinants and effects on shopping response". In: *Journal of retailing* 74.4 (1998), pp. 515–539.

[107] Chia-Chien Wu and Eileen Kowler. "Timing of saccadic eye movements during visual search for multiple targets". In: *Journal of vision* 13.11 (2013), p. 11.

[108] Songhua Xu, Hao Jiang, and Francis Lau. "Personalized online document, image and video recommendation via commodity eye-tracking". In: *Proceedings of the 2008 ACM conference on Recommender systems*. ACM. 2008, pp. 83–90.

[109] Toshihiko Yamasaki et al. "Prediction of User Ratings of Oral Presentations using Label Relations". In: *ACM Workshop on Affect and Sentiment in Multimedia (ACMMM ASM)*. ACM. 2015.

[110] Robert M Yerkes and John D Dodson. "The relation of strength of stimulus to rapidity of habit-formation". In: *Journal of comparative neurology and psychology* 18.5 (1908), pp. 459–482.

[111] Jaewook Yoo, Jaerock Kwon, and Yoonsuck Choe. "Predictable internal brain dynamics in EEG and its relation to conscious states". In: *Frontiers in neurorobotics* 8 (2014).

[112] Lingyun Zhang et al. "SUN: A Bayesian framework for saliency using natural statistics". In: *Journal of vision* 8.7 (2008), p. 32.

[113] Luming Zhang et al. "Fusion of multichannel local and global structural cues for photo aesthetics evaluation". In: *Image Processing, IEEE Transactions on* 23.3 (2014), pp. 1419–1429.

[114] Qi Zhang et al. "Rolling guidance filter". In: *Computer Vision–ECCV 2014*. Springer, 2014, pp. 815–830.

List of Publications

Journal Papers

- Litian Sun, Toshihiko Yamasaki, Kiyoharu Aizawa, "Photo aesthetic quality estimation using visual complexity features", Multimedia Tools and Applications (under major revision) [1]

- Litian Sun, and Tadashi Shibata. "Unsupervised object extraction by contour delineation and texture discrimination based on oriented edge features." IEEE Transactions on Circuits and Systems for Video Technology 24, no. 5 (2014): 780-788. [2]

Reviewed Conference Papers

- Litian Sun, Toshihiko Yamasaki, and Kiyoharu Aizawa. "Personal rating prediction for on-line video lectures using gaze information." In 2015 10th International Conference on Information, Communications and Signal Processing (ICICS), pp. 1-5. IEEE, 2015. [1]

- Toshihiko Yamasaki, Yusuke Fukushima, Ryosuke Furuta, Litian Sun, Kiyoharu Aizawa, and Danushka Bollegala. "Prediction of user ratings of oral presentations using label relations." In Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia, pp. 33-38. ACM, 2015. [2]

[1] Related to this thesis
[2] Not related to this thesis, and done in the Univerity of Tokyo

- Litian Sun, Toshihiko Yamasaki, and Kiyoharu Aizawa. "Relationship Between Visual Complexity and Aesthetics: Application to Beauty Prediction of Photos." In Workshop at the European Conference on Computer Vision, pp. 20-34. Springer International Publishing, 2014. [1]

- Litian Sun, and Kiyoharu Aizawa. "Action recognition using invariant features under unexampled viewing conditions." In Proceedings of the 21st ACM international conference on Multimedia, pp. 389-392. ACM, 2013. (Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge) [2]

- Litian Sun, and Tadashi Shibata. "Unsupervised object extraction by contour delineation and texture-based discrimination." In Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pp. 1945-1949. IEEE, 2012. [2]

Non-reviewed Conference Papers

- Litian Sun, Toshihiko Yamasaki, Kiyoharu Aizawa, "Portrait aesthetic quality estimation by semantic complexity", ITE, 2016. [1]

- Litian Sun, Toshihiko Yamasaki, Kiyoharu Aizawa, "Prediction of subjective impression labels of TED talks based on viewers gaze information", ITE, 2015. [1]