VISUAL CHARACTERISTICS FOR COMPUTATIONAL PREDICTION OF AESTHETICS AND EVOKED EMOTIONS

A Dissertation in
Information Sciences and Technology
by
Xin Lu

© 2016 Xin Lu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2016
The dissertation of Xin Lu was reviewed and approved* by the following:

James Z. Wang  
Professor of Information Sciences and Technology  
Dissertation Advisor, Chair of Committee

Mary Beth Rosson  
Professor of Information Sciences and Technology

Reginald B. Adams, Jr.  
Associate Professor of Psychology

Michelle G. Newman  
Professor of Psychology

Jia Li  
Professor of Statistics

Mary Beth Rosson  
Professor of Information Sciences and Technology  
Associate Dean of Graduate and Undergraduate Studies

*Signatures are on file in the Graduate School.
Abstract

Human emotions and aesthetic feelings that are aroused by natural photographs have been actively studied during the past decades due to their potential applications to the development of intelligent computer systems and to broad areas of science and technology related to human emotion and aesthetics. In this dissertation, investigations of visual characteristics that evoke human emotion and aesthetic feelings are presented. First, shape features were studied in natural images in terms of how they influence emotions aroused in human beings. Shapes and their characteristics—such as roundness, angularity, simplicity, and complexity—have been found to evoke emotions in human perceivers, with evidence from psychological studies of facial expressions, dancing poses, and even simple synthetic visual patterns. Capturing these characteristics algorithmically to incorporate in computational studies, however, has proven difficult. Moreover, little prior research has modeled the dimensionality of emotions aroused by roundness and angularity. In this study, a collection of shape features was developed, which encoded the visual characteristics of roundness, angularity, and complexity using edge, corner, and contour distributions. Evaluation of those features were performed on the International Affective Picture System (IAPS) dataset, where evidence was provided regarding the significance of roundness-angularity and simplicity-complexity on predicting emotional content in images. Second, an investigation into three visual characteristics, i.e., roundness, angularity, and simplicity, of complex scenes that evoke human emotion was performed. Built upon the high-dimensional shape features, novel computational methods were developed to map visual content to the scales of roundness, angularity, and simplicity as three new constructs. The scope of the previous psychological hypothesis was, therefore, expanded by examining these three visual characteristics in computer analysis of complex scenes. The results produced by the three new constructs were compared to the hundreds of visual qualities examined by previous studies. The three constructs were completely
interpretable and could be used in other applications involving roundness, angularity, and simplicity of visual scenes. Meanwhile, a large collection of ecologically valid stimuli (i.e., photographs humans regularly encounter on the Web), containing more than 40K images crawled from web albums, was generated using crowdsourcing and was subjected to human subject emotion ratings. Critically, these three new visual constructs achieved classification accuracy comparable to the hundreds of shape, texture, composition, and facial feature characteristics previously examined. This reduces the number of features required for classification by about two orders of magnitude. In addition, our experimental results showed that the three constructs showed consistent capacity in classifying both dimensions of emotions. Finally, a novel deep learning algorithm was developed to automatically learn effective visual characteristics for image aesthetics assessment. The proposed RAPID (RAting PIctorial aesthetics using Deep learning) system, incorporates heterogeneous inputs generated from the image, which include a global view and a local view, and unifies the feature learning and classifier training using a double-column deep convolutional neural network. The experimental results showed that the RAPID system significantly outperformed the state of the art on the AVA dataset. The results of the three studies demonstrate (1) the capability of roundness, angularity, and complexity of complex scenes to evoke human emotions, and (2) the capability of global view and fine-grained details of complex scenes to evoke aesthetic feelings.
# Table of Contents

List of Figures ix

List of Tables xiii

Acknowledgments xiv

Chapter 1
Introduction 1
1.1 Emotion Recognition . . . . . . . . . . . . . . . . . . . . . . . . . . 2
   1.1.1 Shape and the Computability of Emotions . . . . . . . . . . 3
   1.1.2 The Three Visual Characteristics . . . . . . . . . . . . . . 4
1.2 Image Aesthetics Assessment . . . . . . . . . . . . . . . . . . . . . 6
1.3 Applications of Emotion Recognition and Image Aesthetics Assessment 8
   1.3.1 Psychological Applications . . . . . . . . . . . . . . . . . . 8
   1.3.2 Computational Applications . . . . . . . . . . . . . . . . 9
1.4 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
1.5 Structure of Dissertation . . . . . . . . . . . . . . . . . . . . . . . . 11

Chapter 2
Background: Psychology of Visual Characteristics and Emotions 12
2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
2.2 Psychological Background: Visual Characteristics . . . . . . . . . . 12
   2.2.1 Roundness and Complexity . . . . . . . . . . . . . . . . . . . 12
   2.2.2 Holistic Information and Fine-Grained Details . . . . . . 14
   2.2.3 Other Visual Characteristics . . . . . . . . . . . . . . . . 15
2.3 Psychological Background: Aesthetics . . . . . . . . . . . . . . . . . 16
2.4 Psychological Background: Emotion Representation . . . . . . . . . . 17
5.2.4 Dataset Statistics ........................................... 54
5.3 Constructs of Roundness, Angularity, and Simplicity ........... 55
  5.3.1 Roundness ............................................. 55
  5.3.2 Angularity ............................................. 58
  5.3.3 Simplicity ............................................. 58
5.4 Findings ..................................................... 59
  5.4.1 Statistical Correlations .................................. 60
  5.4.2 Emotion Classification ................................... 61
5.5 Summary ..................................................... 64

Chapter 6
RAPID: RAting PIctorial Aesthetics using Deep Learning 65
  6.1 Introduction ............................................... 66
  6.2 Deep Convolutional Neural Networks .......................... 67
    6.2.1 Convolutional Neural Network .......................... 67
    6.2.2 Double-Column Convolutional Neural Network ......... 70
    6.2.3 Regularized Double-Column Convolutional Neural Network 71
  6.3 Evaluation ................................................ 73
    6.3.1 SCNN Results .......................................... 74
    6.3.2 DCNN Results .......................................... 75
    6.3.3 Content-based Image Aesthetics ........................ 79
    6.3.4 Categorization with Style Attributes ................... 82
    6.3.5 Categorization with Semantic Attributes ............... 84
    6.3.6 The IAD dataset ...................................... 85
    6.3.7 Implementation Details and Computational Efficiency ... 87
  6.4 Summary ................................................... 87

Chapter 7
Conclusions and Future Work 89
  7.1 Summary ................................................... 89
  7.2 Limitations ............................................... 90
  7.3 Future Work .............................................. 91

Appendix
Image-Specific Prior Adaptation for Denoising 93
  1 Introduction .............................................. 93
    1.1 Related Work .......................................... 95
    1.2 Contributions .......................................... 96
  2 Modeling Patch Statistics using Gaussian Mixture Model ........ 97
    2.1 Gaussian Mixture Model .................................. 97
List of Figures

2.1 Example images from IAPS (The International Affective Picture System) dataset [41]. Images evoke positive affect are from left to right, and high arousal are from bottom to top. .......................... 13

2.2 Dimensional representation of emotions and an illustration of categorical emotions in the dimensional space. .......................... 17

4.1 Perceptual shapes of images with high valance. .......................... 33

4.2 Perceptual shapes of images with low valance. .......................... 34

4.3 Perceptual shapes of images with high arousal. .......................... 34

4.4 Perceptual shapes of images with low arousal. .......................... 35

4.5 The corner point and point of inflexion. .......................... 35

4.6 Images with low mean value of the length of line segments and their associated orientation histograms. The first row is the original images; the second row shows the line segments; and the third row shows the 18-bin histogram for line segments in the images. .......................... 38

4.7 Images with high mean value of the length of line segments and their associated orientation histograms. The first row is the original images; the second row shows the line segments; and the third row shows the 18-bin histogram for line segments in the images. .......................... 38

4.8 Images with highest and lowest number of angles. .......................... 39

4.9 The distribution of angles in images. .......................... 39

4.10 Images with highest degree of curving. .......................... 40

4.11 Images with lowest degree of curving. .......................... 40

4.12 Distribution of ratings in valence and arousal. .......................... 42

4.13 Classification accuracy (%) for emotional images and neutral images (Set 1 and Set 2 are defined in Section 4.3.1). .......................... 43

4.14 Examples of misclassification in Set 1. The four rows are original images, image contours, line segments, and continuous lines. .......................... 44

4.15 Examples of misclassification in Set 2. The four rows are original images, image contours, line segments, and continuous lines. .......................... 45
4.16 Experimental results. (a) Mean squared error for the dimensions of valance and arousal. (b) Accuracy for the classification task.

5.1 Examples of images in the EmoSet. Images in the EmoSet were retrieved from Flickr and triggered by 558 emotional words summarized in [97].

5.2 Mean value distributions of valence, arousal, dominance, and likeability in the EmoSet.

5.3 Examples of images and their scores of roundness. The images with highest roundness scores are shown in the first row; images with medium ranges of roundness scores in the second row; and images with lowest roundness scores in the third row.

5.4 Examples of images and their scores of angularity. The images with highest angularity scores are shown in the first row; images with medium ranges of roundness scores in the second row; and images with lowest roundness scores in the third row.

5.5 Examples of images and their scores of simplicity. The closer the score is to 1 means that the image is simpler.

5.6 Correlation between roundness and valence, arousal, dominance, and likeability in natural photographs.

5.7 Correlation between angularity and valence, arousal, dominance, and likeability in natural photographs.

5.8 Correlation between simplicity and valence, arousal, dominance, and likeability in natural photographs.

6.1 Global views and local views of an image. Global views are represented by normalized inputs: center-crop, warp, and padding (shown in the top row). Local views are represented by randomly-cropped inputs from the original high-resolution image (examples shown).

6.2 Single-column convolutional neural network for aesthetic quality rating and categorization. We have four convolutional layers and two fully-connected layers. The first and second convolutional layers are followed by max-pooling layers and normalization layers. The input patch of the size $224 \times 224 \times 3$ is randomly cropped from the normalized input of the size $256 \times 256 \times 3$ as done in [30].
6.3 Double-column convolutional neural network. Each training image is represented by its global and local views, and is associated with its aesthetic quality label: 0 refers to a low quality image and 1 refers to a high quality image. Networks in different columns are independent in convolutional layers and the first two fully-connected layers. The final fully-connected layer are jointly trained.

6.4 Regularized double-column convolutional neural network (RDCNN). The style attributes $\mathbf{x}_s$ are generated through pre-trained Style-SCNN and we leveraged the style attributes to regularize the training process of RDCNN. The dashed line indicates that the parameters of the style column is fixed during RDCNN training. While training the RDCNN, we only fine-tuned the parameters in the aesthetic column and the learning process is supervised by the aesthetic label.

6.5 Filter visualization of DCNN for image aesthetics. In particular, 128 convolutional kernels of the size $11 \times 11 \times 3$ learned by the first convolutional layer. The first 64 are from the local view column (with the input $I^l_f$) and the last 64 are from the global view column (with the input $I^w_g$).

6.6 64 convolutional kernels of the size $5 \times 5 \times 3$ learned by the first convolutional layer of CNN for object classification on the CIFAR dataset.

6.7 Images ranked the highest and the lowest in aesthetics generated by DCNN. Differences between low-aesthetic images and high-aesthetic images heavily lie in the amount of textures and complexity of the whole image.

6.8 Test images correctly classified by DCNN but misclassified by SCNN. The first row shows the images that are misclassified by SCNN with the input $I^l_f$. The second row shows the images that are misclassified by SCNN with the input $I^w_g$. The label on each image indicates the ground-truth aesthetic quality.

6.9 Classification accuracy of image aesthetics in the eight semantic categories.


6.11 32 convolutional kernels of the size $11 \times 11 \times 3$ learned by the first convolutional layer of Style-SCNN for style classification.

6.12 Test images correctly classified by RDCNN$_{style}$ and misclassified by DCNN. The label on each image indicates the ground truth aesthetic quality of images.
6.13 Test images correctly classified by RDCNN\textsubscript{semantic} and misclassified by DCNN. The label on each image indicates the ground truth aesthetic quality of images. ................. 84

6.14 The Mean Score Distributions of images collected from DPChallenge and PHOTO.NET. Left: The mean score distributions of the 1.2 million PHOTO.NET images; Right: The mean score distributions of the 300K DPChallenge images. ..................... 85
List of Tables

4.1 Summary of Shape Features ........................................ 36
4.2 Average number of curves in terms of the value of fitness in positive and negative images. ........................................ 41
4.3 Average number of curves in terms of the value of circularity in positive and negative images. ........................................ 41
4.4 Significant Features for Emotion Classification .................. 47

5.1 Classification Results of High Valence vs. Low Valence ...... 62
5.2 Classification Results of High Arousal vs. Low Arousal ...... 63

6.1 Accuracy of Different SCNN Architectures ....................... 74
6.2 Accuracy of Aesthetic Quality Categorization with Different Inputs .......................................................... 75
6.3 Accuracy of Aesthetic Quality Categorization with Different Methods .......................................................... 75
6.4 Accuracy of Different Network Architectures for Style Classification .......................................................... 82
6.5 Accuracy of Style Classification with Different Inputs .......... 82
Acknowledgments

Pursuing a Ph.D. degree has been a fantastic journey in my life through which I learned to be creative and imaginative, positive and reflective, and active in learning both mentally and physically. I would like to express my gratitude to those who have been with me and helped to shape my life in these years.

Many thanks go first to my advisor, Professor James Z. Wang, who always encourages me to look at the positive side of whatever I encounter in life and to explore ambitiously any excitement that life might bring. I cherish the freedom in research and study he has allowed me, and appreciate all the support he has provided me with during the pursuit of my Ph.D.

Next I would like to thank my committee members. I am grateful to Professor Jia Li for her valuable comments from a statistical perspective regarding my dissertation. I am indebted to Professor Reginald B. Adams, Jr. and Professor Michelle G. Newman for their insightful discussion and invaluable guidance from a psychological perspective. I would also like to thank Professor Mary Beth Rosson for serving on my committee and additional assistant on my thesis.

I am grateful to have worked with Adobe Systems, Inc. during the past three summers and to have collaborated with talented researchers there beyond these internships. These experiences have been some of the most rewarding, impressive, and influential of my Ph.D. studies. I would like to take this opportunity to express my great appreciation to Dr. Zhe Lin, Dr. Hailin Jin, Dr. Jianchao Yang, Dr. Radomír Měch, Dr. Xiaohui Shen, and Dr. Jonathan Brandt for their great support, invaluable advice, and generous help.

My gratitude also extends to professors in the College of Information Sciences and Technology and my colleagues in the Intelligent Information Systems Laboratory. Many thanks to Professor Guoray Cai, Professor Michael McNeese, and Professor John Yen, who guided and supported me during my first year in the Ph.D. program. I am indebted to Poonam Suryanarayan, Neela Sawant, Dhiraj Joshi, Michael J. Costa, Jason M. Frederick, Baris Kandemir, Hanjoo Kim, and Jian Wu for our joint efforts on research projects and papers. I am also grateful to Lei Yao, Mu Qiao, Yu Zhang, Siqiong He, Jianbo Ye, Yukun Chen, Zhaohui Wu,
and Pucktada Treeratpituk for all inspiring discussions.

My Ph.D. study was funded by the National Science Foundation, the College of Information Sciences and Technology, and gift funds from Adobe Systems Inc. I am truly grateful for the support.

I owe sincere appreciation to the friends that I have made while pursuing my Ph.D. You make my life so wonderful and colorful, and I will keep all our beautiful memories sealed forever in my heart. My sincere thanks goes to all of you for being part of my life over the past several years.

Most importantly, I would like to thank my parents for their infinite and unconditional love. Thanks Mom and Dad for providing me with a warm, creative, and encouraging family atmosphere, which endows me with the confidence and courage to build my own life.
Chapter 1

Introduction

Automated identification of human emotion and aesthetic feelings triggered by pictorial scenes has become an active research area in the past decades due to the emerging need to allow computers to perceive emotions and aesthetics. Potential usage of methods developed for these tasks could be foreseen towards wide applications, ranging from intelligent computer systems to real-time, mobile apps. For instance, emotion recognition tools might be leveraged to determine which picture might be used to decorate workplaces, hospitals, gymnasias, and schools. It could also be used to prevent children from viewing images of bad taste. Automated assessment or rating of pictorial aesthetics also has many applications. In an image retrieval system, the ranking algorithm can incorporate aesthetic quality as one of the factors. In picture editing software, aesthetics can be used in producing appealing polished photographs.

As defined by the New Oxford American Dictionary, human emotion is “a natural instinctive state of mind deriving from one’s circumstances, mood, or relationships with others.” Meanwhile, aesthetics is generally defined as the concept of beauty and artistically pleasing appearance, expression of beauty, and response to beauty [1]. Problems of image aesthetics and emotions have long been studied in visual arts and psychology under controlled experimental settings, due to their subjective nature. With the increasingly rapid growth of computational power and intelligent computational systems, we have seen the dawn of hope for endowing computers with the capabilities of perceiving emotions and aesthetics, when encountering visual scenes appear in daily life.
Most recently, interdisciplinary approaches have been commonly adopted to study the problems of image aesthetics and evoked emotions. One research approach is the investigation of findings from psychological or visual arts studies and the development of computational tools in order to leverage those findings to help predict aesthetics and emotions. Another direction for research is the development of computational algorithms in order to learn from large-scale labeled data collections and predict aesthetics and emotions. In this dissertation, both research approaches have been adopted to study image aesthetics and emotions, respectively.

Image aesthetics and emotion recognition are closely related research problems because both of the problems target human visual perception beyond the identification of the content of images. Emotion recognition, compared with image aesthetics, is an even more challenging research problem because of the tremendous variations of human emotion and the scarceness of large-scale labeled data collections for emotion recognition. Investigations regarding emotion recognition, therefore, have been conducted in close collaboration with professors in the psychology department in order to leverage psychological insights gained from previous studies.

Many previous computational studies have paved way for image aesthetics assessment leveraging photographic or psychological findings. Motivated by a recently released large-scale labeled dataset, we have investigated image aesthetics by developing computational algorithms to automatically identify visual representations and perform predictions. This study was a joint work with researchers in Adobe Systems, Inc, and the developed algorithm might be adopted for emotion recognition.

The remaining chapter is organized as follows: In Sections 1.1 and 1.2, we introduce the problems of emotion recognition and image aesthetics assessment that this dissertation has targeted. In Section 1.3, we discuss the potential applications of approaches developed in this dissertation. The contributions of this dissertations are summarized in Section 1.4. The structure of remaining chapters in this dissertation is presented in Section 1.5.

1.1 Emotion Recognition

In the past decades, a computational perspective to the problem of emotion recognition has attracted many researchers and resulted in articles on modeling
the emotional and aesthetic content in images [2–4]. However, there is a wide gap between what humans can perceive and feel and what can be explained using the state-of-the-art computational image features. Bridging this gap is considered the “holy grail” of computer vision and the multimedia community.

In this dissertation, we adopt both the dimensional and the categorical representations of emotion to extend our understanding of some of the low-level features which have not been explored in the study of visual affect through extensive statistical analyses. In contrast to prior studies which estimated the level of visual appeal [2], we leverage some of the psychological studies on visual characteristics of roundness, angularity, and complexity and their effect on human emotions.

In the first study (Section 1.1.1), we designed and developed shape features to encode roundness, angularity, and complexity; in the second study (Section 1.1.2), we developed three constructs by mapping images to the scales of roundness, angularity, and simplicity. Both collections of features were applied to emotion recognition.

1.1.1 Shape and the Computability of Emotions

Researchers [3,5–9] investigated factors such as color, texture, composition, and simple semantics to understand emotions, but have not quantitatively addressed the effect of perceptual shapes. The study that did explore shapes by Zhang et al. [10] predicted emotions evoked by viewing abstract art images, where low-level features like color, shape, and texture were used. This work only handles abstract images, and focuses on the representation of textures with little accountability of shape. Meanwhile, in previous studies, categorial representations were commonly adopted, i.e., emotions were represented through discrete categories like anger, fear, disgust, amusement, awe, and contentment, among others. In this dissertation, we predict emotions aroused from images by adopting both categorical and dimensional representations.

Investigating the quantitative relationship between perceptual shapes and emotions aroused from images is non-trivial. First, emotions aroused by images are subjective. Individuals may not have the same affective response to a given image, making the representation of shapes in complex images highly challenging. Second, images are not merely composed of simple and regular shapes, and complex shapes
occurred in natural images are difficult to model [11].

In this dissertation, we systematically investigate how shapes contribute to perceived emotions aroused from images through modeling the visual properties of roundness, angularity and simplicity using shapes. To model those visual properties in images, we developed a novel approach that statistically analyzed the line segments and curves extracted from strong continuous contours and generated a collection of shape features. Leveraging the proposed shape features, we automatically distinguished the images with strong emotional content from emotionally neutral images and predicted both categorical and dimensional emotions.

1.1.2 The Three Visual Characteristics

Everyday pictorial scenes are known to evoke emotions. The ability of a computer program to predict evoked emotions from visual content will have a high impact on social computing. For instance, before a photograph is shared on social media, the photographer could obtain an automatic assessment on the types of emotions the photo will evoke. Further, users of social networks or clouds could search for photos that evoke certain emotions. In entertainment, movie directors and advertising agencies could use such technologies to predict the potential emotions of the viewers.

It remains unclear, however, what specific visual characteristics of scenes are associated with specific emotions, such as calmness, dynamism, turmoil, or happiness. Finding such associations is arguably one of the most fundamental research problems in visual arts, psychology, and computer science, which has resulted in many relevant research articles [2–6,10–17] over past decades. In particular, hundreds of visual characteristics, ranging from edge distributions and color histograms to SIFT, GIST, and Fisher vectors had been examined by recent computational studies to predict human emotional responses.

Beyond those visual characteristics, many empirical studies in psychology and visual arts have investigated roundness, angularity, simplicity, and complexity (Complexity is essentially the opposite of simplicity. We thus omitted complexity in the later presentations), and their capacity to evoke emotion. By performing experiments using facial expressions [12], dancing poses [12], and synthetic visual patterns [14], studies [12,13,42] indicated that more rounded properties led to
positive feelings, such as warmth and affection, whereas more angular properties tended to convey threat. Meanwhile, studies in [11, 122, 123] showed that humans preferred simple visual scenes and stimulus patterns.

To expand the scope of previous studies, herein, we investigate the computability of these three visual characteristics of complex scenes that evoke human emotion. Describing these characteristics mathematically or computationally, with the purpose of predicting emotion and understanding the capacity of these characteristics to evoke emotion, is nontrivial. In our earlier research, we developed a collection of shape features that encoded the visual characteristics of roundness, angularity, and simplicity using edge, corner, and contour distributions [16]. Those features were shown to predict emotion to some extent on the IAPS dataset\(^1\), but they were of a high dimension making them difficult to interpret. Under classification or regression frameworks, representations in different dimensions are intertwined for emotion prediction. Therefore, articulating what specific visual characteristics might be associated with a certain emotion turns out to be extremely difficult.

To better understand the role of roundness, angularity, and simplicity in evoking human emotion, we target the following three questions: (1) whether visual characteristics of roundness, angularity, and simplicity could be calculated with one number; (2) for natural setting photos, whether roundness, angularity, and simplicity strongly correlated with human emotion as stated in controlled psychological studies; (3) whether those single dimensional representations are equally as good as high-dimensional representations in emotion classification tasks, such as classifying images that arouse positive and negative emotions and distinguishing images that evoke calmness from excitement. To answer these questions, we propose novel computational methods to map visual content to the scales of roundness, angularity, and simplicity as three new constructs. Meanwhile, we collected a large collection of ecologically valid stimuli by rating natural setting photos, which allows the correlation study between roundness, angularity, and simplicity and human emotion on natural setting photos. In addition, we examined the effectiveness of those single dimensional representations in emotion classification tasks.

\(^{1}\)The International Affective Picture System (IAPS) dataset were developed by Lang et al. [41] through examining human affective responses to color photographs with varying degrees of emotional content. The IAPS dataset contains 1182 images, wherein each image is associated with an empirically derived mean and standard deviation of valance, arousal, and dominance ratings. A small subset of images in the IAPS dataset are also associated with categorical emotion labels.
A central difference between this work and previous studies is that questions posted in this work are on the scale of natural setting photos in contrast to controlled psychological studies. As a result, studies in this work were performed on a large collection of natural photos instead of on a small collection of manually selected photos. In particular, we have been working on developing a large collection of ecologically valid stimuli (named the EmoSet). Unlike the commonly used The International Affective Picture System (IAPS), for which 1082 images were rated with emotional responses, the EmoSet contains more complex scenes that humans regularly encounter in life. We have collected about 43,837 images with manually-supplied emotion labels (both dimensional and categorical) following strict psychological subject study procedures and validation approaches. These images were crawled from more than 1000 users’ Web albums using 558 emotional words as search terms summarized by Averill [97]. We used the 558 words to search Flicker. We introduce the data collection approach in Section 5.2, where the derived EmoSet marks the first time a human subject study, examining perceived emotion triggered by visual stimuli, was performed using natural photographs of complex scenes. The EmoSet is also different from large-scale affective datasets introduced in [23] and [98], where researchers crawled user-generated content, including pictorial ratings or associated affective tags, to indicate the affective intention of images. Whereas those datasets were of large scale, the emotional labels were not generated by human subjects under strict psychological procedures. For instance, psychological conventions were applied when recording human perceived emotions, i.e., aroused emotions are best recorded within six seconds after subjects view each visual stimulus [41]. More details will be provided later.

1.2 Image Aesthetics Assessment

Effective visual features are essential for computational aesthetic quality rating systems. Existing methods used machine learning and statistical modeling techniques on handcrafted features or generic image descriptors. Specifically, Datta et al. [25] and Ke et al. [26] formulated the problem of image aesthetics as a classification or regression problem where a given image is mapped to an aesthetic rating, which is normally quantized with discrete values. Under this framework, the effectiveness of the image representation, or the extracted features, can often be the
accuracy bottleneck. Various handcrafted aesthetics-relevant features have been proposed \cite{15,18,19,25-29}, including low-level image statistics such as distributions of edges and color histograms, and high-level photographic rules such as the rule of thirds.

While these handcrafted aesthetics features are often inspired from the photography or psychology literature, they share some known limitations. First, the aesthetics-sensitive attributes are manually designed, hence have limited scope. It is possible that some effective attributes have not yet been discovered through this process. Second, because of the vagueness of certain photographic or psychologic rules and the difficulty in implementing them computationally, these handcrafted features are often merely approximations of such rules. There is often a lack of principled approach to improve the effectiveness of such features.

Generic image features \cite{22-24} were proposed to address the limitations of the handcrafted aesthetics features. They used well-designed common image features such as SIFT and Fisher Vector \cite{20,22}, which have been successfully used for object classification tasks. The generic image features have been shown to outperform the handcrafted aesthetics features \cite{22}. However, because these features are meant to be generic, they may be unable to attain the upper performance limits in aesthetics-related problems.

In this dissertation, we intend to explore beyond generic image features by learning affective aesthetics features from images directly. We are motivated by the recent work in large scale image classification using deep convolutional neural networks \cite{30} where the features are automatically learned from RGB images. The deep convolutional neural network takes pixels as inputs and learns a suitable representation through multiple convolutional and fully connected layers. The originally proposed architecture cannot be directly applied to our task. Image aesthetics relies on a combination of local and global visual cues. For example, the rule of thirds is a global image cue while sharpness and noise levels are local visual characteristics. Given an image, we generated two heterogeneous inputs to represent its global cues and local cues respectively. To support network training on heterogeneous inputs, we extended the method in \cite{30} by developing a double-column neural network structure which takes parallel inputs from the two columns. One column takes a global view of the image and the other column takes a local view of the image. We integrated the two columns after some layers of transformations.
to form the final classifier. We further improved the aesthetic quality categorization by exploring style and semantic attributes associated with images. We named our system RAPID, which stands for RAting PIctorial aesthetics using Deep learning. We demonstrated the superior performance of our approach on the AVA dataset (a recently-published large dataset).

1.3 Applications of Emotion Recognition and Image Aesthetics Assessment

1.3.1 Psychological Applications

The research approach and findings from this dissertation may have a long-term impact on studies related to image aesthetics and evoked emotion in psychology, visual arts, and computer science. In previous psychological and visual arts studies, many hypotheses were made regarding the visual characteristics and their capabilities to evoke emotion. From the perspective of color vision, Changizi et al. indicated that different colors led to different emotional states [31]. Emotions of various gender, age, and ethnic groups were investigated by Rodgers et al. through examining the frames in news photographs on three dimensions of emotion [32]. From the perspective of roundness and angularity, Aronoff et al. showed that increased roundedness leads to more warmth, and increased linearity, diagonality, and angularity of forms lead to feeling threatened [33]. Aronoff et al. later confirmed that geometric properties of visual displays conveyed emotions such as anger and happiness [12]. Similar hypotheses were made and demonstrated by Bar et al. [14] showing that curved contours led to positive feelings and sharp transitions in contour triggered a negative bias. Visual arts studies also indicated that humans visually preferred simplicity. Any stimulus pattern was always perceived in the most simplistic structural setting.

In these studies, conventional approaches, such as human subject studies and interviews, were adopted, which limits the generalizability of the study in complex scenes. The computational approaches of emotion recognition and image aesthetics assessment developed in this dissertation could be widely adopted in future studies related to image aesthetics and emotions in psychology and visual arts. The findings and indications regarding holistic information and fine-grained details presented in
this dissertation may also advance emotion and aesthetics related studies. Possible future research questions are as following:

1. In which scenarios do roundness, angularity, and simplicity have the strongest capability of evoking human emotion.

2. How does demographic information affect the aroused emotion when human encountering complex scenes.

3. In which scenarios do fine-grained details matter most regarding the aesthetic feeling of complex scenes and in which scenarios does holistic information matter most?

1.3.2 Computational Applications

Researchers have applied the idea of image aesthetics and emotions to real-world systems and related research topics. Yao et al. developed the system OSCAR to help photographers to generate high-quality photos [34]. OSCAR provides on-site analyses of photos in terms of the composition and aesthetic quality, based on which the system automatically generates feedback through high-quality examples. Bhattacharya et al. proposed a learning-based framework to help improve the quality of photos with spatial recommendation techniques [35]. For photos with a distinct foreground, the work automatically recommends an alternative position for the foreground to enhance the aesthetic quality of the photo. For other photos, the framework provides recommendations to crop or expand the photos to make it aesthetically more appealing.

Studies on image aesthetics assessment and aroused emotion prediction provide hints for the image retrieval and scene recognition. By considering the attractiveness of images, a study [36] has shown that performance of image search engines is improved in terms of the online ranking, interactive re-ranking, and offline index selection. Redi and Meritaldo explored the discriminative power of visual features that represents the composition of images to tackle the task of scene recognition [37]. By integrating semantic, aesthetic, and affective features, an averagely 13% – 15% improvement was demonstrated for the task of scene recognition on various, diverse, and large-scale datasets.
The findings in this dissertation may facilitate potential applications of identifying image affect in computer vision and multimedia systems. For instance, in image retrieval systems, the constructs of simplicity and roundness could be incorporated into the ranking algorithm to help arouse positive feelings of users, whereas they are using the image search engine. The construct of angularity could be utilized to help protect children from viewing pictures that contain anger, fear, disgust, or violence. Similarly, image editing softwares could take the findings of our work into account when making design suggestions to photographers. Meanwhile, findings from this study may contribute to relevant research studies in computer science, such as the automatic predictions of image memorability [38], interestedness [29,39], and popularity [40]. As we have shown the capability of roundness, angularity, and simplicity on evoking human emotion and the capability of holistic information and fine-grained details on affecting aesthetic feelings, these visual characteristics may also shed light on interpreting the memorability, interestedness, and popularity of an image.

1.4 Contributions

We summarize the main contributions as follows:

- **Shape Features.** We systematically investigated the correlation between visual shapes and emotions aroused from images and quantitatively model the concepts of roundness-angularity and simplicity-complexity from the perspective of shapes using a dimensional approach. Moreover, we distinguished images with strong emotional content from those with weak emotional content.

- **The Three Constructs of Complex Scenes.** We quantized, investigated, and computed roundness, angularity, and simplicity with three numbers. The three constructs are completely interpretable and could be used in other applications involving roundness, angularity, and simplicity of visual scenes. For natural setting photos, roundness, angularity, and simplicity have rather weak relationship with emotion although in controlled psychological studies, they were found related to emotional responses. The capacity of the three constructs to classify the positivity of emotional responses was established. When combined with color features, the three constructs achieve comparable
classification accuracy on the positivity of emotions as a set of over 200 shape, texture, composition, and facial features. This reduces the number of features required for classification by about two orders of magnitude. In addition, our experimental results showed that the three constructs showed consistent capacity in classifying both dimensions of emotions.

- Deep Learning For Image Aesthetics Assessment  We conducted systematic evaluation of the single-column deep convolutional neural network approach with different types of input modalities for aesthetic quality categorization and developed a double-column deep convolutional neural network architecture to jointly learn features from heterogeneous inputs. Moreover, we developed a regularized double-column deep convolutional neural network to further improve aesthetic categorization using style and semantic attributes.

1.5 Structure of Dissertation

The remainder of the dissertation is organized as following: Chapter 2 details psychological foundations of this dissertation. Chapter 3 reviews previous studies related to the computational modeling of aesthetics and emotions. In Chapter 4, shape features is introduced and its effectiveness is demonstrated on emotion recognition. Chapter 5 presents computational approaches to compute the three visual characteristics using three constructs, where findings are presented regarding the relationship between the three novel constructs and human emotion. Chapter 6 introduces novel deep learning approaches for image aesthetics assessment, which leverages both the global information and fine-grained details of images. Chapter 7 concludes the dissertation and discusses future work.
Chapter 2

Background: Psychology of Visual Characteristics and Emotions

2.1 Introduction

In psychology, the development of aesthetics and evoked emotions involves research work in sub-areas of experimental aesthetics, psychology of art, neuroaesthetics, and social psychology. Research problems in psychology include three major questions that center on humans’ affective responses: How to represent human’s affective response? What kind of visual patterns associated with visual scenes arouse those responses? How are those responses generated? In this chapter, we review psychological literatures related to image aesthetics and evoked emotions.

2.2 Psychological Background: Visual Characteristics

2.2.1 Roundness and Complexity

The study of human visual preferences and emotions imparted by various works of art and natural images has long been an active research topic in the field of visual arts and psychology. Many psychological theories suggest a link between human affective responses and the low-level features in images. In particular, studies indicate that roundness and complexity of shapes are fundamental to understanding
Figure 2.1: Example images from IAPS (The International Affective Picture System) dataset [41]. Images evoke positive affect are from left to right, and high arousal are from bottom to top.

...emotions.

- **Roundness** - Aronoff et al. indicated that increased roundedness leads to more warmth, and increased linearity, diagonality, and angularity of the form lead to feeling threatened [33]. Findings from [13] implied that beauty is reflected through the fluency for perceivers to process an object, and the more fluently it does, the more positive the response. This work reviewed factors that may have an impact on aesthetic response, including figural goodness, figure-ground contrast, stimulus repetition, and symmetry, and confirmed the findings by monitoring the influences introduced by changes of those factors. Meanwhile, by performing experiments using facial expressions [12], dancing poses [12], and synthetic visual patterns [14], studies [12, 13, 42] indicated that more rounded properties led to positive feelings, such as warmth and affection, whereas more angular properties tended to convey threat.

- **Complexity of shapes** - As enumerated in various works of art, humans visually prefer simplicity. Any stimulus pattern is always perceived in the most simplistic structural setting. Although the perception of simplicity is
partially subjective to individual experiences, it could also be highly affected by two objective factors, parsimony and orderliness. Parsimiony refers to the minimalistic structures that are used in a given representation, whereas orderliness refers to the simplest way of organizing these structures [11].

2.2.2 Holistic Information and Fine-Grained Details

Patterns in aesthetic photographs indicate human beings’ visual preferences. Among those patterns, composition [43] and pictorial balance [44] (i.e., visual weights on local elements) are two of the most important factors. Popular composition principles include the rule of thirds, diagonal lines, and golden ratio [1], whereas pictorial balance is affected by position, form, size, tone, color, brightness, contrast, and proximity to the fulcrum [44]. Lines of different direction have completely different indications, where horizontal lines and vertical lines suggest peace and strength respectively, and position of lines may indicate the associated relationship with other elements in the visual design [45]. Common shapes in the photographs include the circle, the square, the oval, the triangle, and the rectangle. Associated with other elements, shapes have an influence on the feeling of proportions in photographs. Visual balance also bears upon the texture and perspective of photographers. Imbalance, in contrast, may suggest dominance and tension [45].

Compared with photography, paintings allow more “Dynamics” [46] in terms of a point, a direction of the brush, and intensity. Though the Gestalt principle of simplicity presents that the visual pattern is prone to the simplest context-aware configuration, the dynamic property inherently exists in every visual object [11]. In El Greco’s St. Jerome [11], points of various intensity and direction of brushes are employed to express the movement of the beard. In Géricault’s Derby at Epsom and Moslem architecture the horseshoe arch [11], the artist uses the direction of the shape to depict the dynamics. Also, in Hans Thoma’s work Illustration from Quickborn [11], the composition of the painting indicates the dynamics. Meanwhile, the visual patterns from paintings do not merely lie in the content but also in the media, “the sheet of paper”, “the canvas”, “the stone”, and “the tools and materials” [11]. These media provide the natural textures to generate visual balance and convey different feelings that associate with other elements in paintings.
2.2.3 Other Visual Characteristics

From the perspective of color vision, a study showed that different colors lead to different emotional states [31]. Emotions of various gender, age, and ethnic groups were investigated in [32] through examining the frames in news photographs on three dimensions of emotion. Findings from this work suggested that emotions are highly correlated with gender, ethnicity, and age. The findings also suggested that (1) women in the photos show more positive visual representations than do men; (2) African Americans are usually depicted as happy, excited, whereas Latinos are commonly portrayed as sad, calm, and submissive; (3) the average teens are described as happy, calm and dominant, while senior citizens and children are sad, calm, and submissive. Also, in [47], emotional reactions to color hue, saturation, and brightness were presented in terms of the three dimensions of emotion, i.e., valence, arousal, and dominance. Blue, blue-green, green, red-purple, purple, and purple-blue were shown to be the most pleasant hues, and yellow, green-yellow were the least pleasant ones in contrast. Likewise, green-yellow, blue-green, and green were demonstrated as the most arousing, and the least arousing were colors of purple-blue and yellow-red. Red-purple was indicated to be of less dominance than green-yellow.

Literatures from both visual arts and psychology show the influence of visual stimuli from the scene and object level, respectively. Scene-level visual stimuli include factors like visual weight (e.g., position, depth, and color) and visual direction (e.g., weights attraction, structural skeleton, subject matter, and movement). The object-level stimuli include factors like symmetry, prototypically, contrast, complexity, perceptual fluency, and curved visual expression. In addition to aforementioned factors, space and movement were shown to affect humans’ visual perception [11]. The spatio-temporal feeling in visual arts were presented through three dimensions, where the spatial concept is limited to a line in the first dimension, enriched by shape of different sizes (round and angular, large and small, and regular and irregular) in two dimensions and extended by unlimited arrangement of visual objects in the three dimensional space. In open space, motion can produce strong visual appeal and demand attention [11].
2.3 Psychological Background: Aesthetics

Experimental aesthetics is among the earliest subjects contributing to the development of aesthetics in psychology [4]. The beginning of a significant movement in experimental aesthetics was made by Gustav Theodor Fechner who studied the problem of utilizing metrics to investigate art and aesthetics [124–126]. Finding proper metrics to evaluate aesthetics then became one of the most fundamental problems in empirical aesthetics. The pionerring rectangle experiment conducted by Fechner showed that the metric of the golden ratio is related to special aesthetic properties. In this experiment, Fechner placed ten rectangles of different length-to-width ratios on the top of an object, where human subjects were asked to select one of the most pleasing rectangles. Human subjects’ selections were centered on rectangles with ratios of 1.75, 1.62, and 1.5, with a peak at the 1.62, numerically very close to the golden ratio \(\frac{1+\sqrt{5}}{2}\).

Another two major advances, Gestalt psychology and Berlyne’s thoughts in neural mechanisms of perception associated with aesthetic feelings, drove the development of experimental aesthetics in the early decades of the twentieth century [4]. Gestalt psychology, also known as the Law of Simplicity, has the central principle that the mind forms a global understanding of external stimuli rather than the sum of each individual part [127]. Many laws of grouping were then formulated by Kohler and Koffka, such as Law of Proximity, Law of Similarity, Law of Closure, Law of symmetry, Law of Common Fate, Law of Continuity, Law of Good Gestalt, and Law of Past Experience.

Daniel Berlyne developed Psychobiological Aesthetics during 1960s and 1970s with the purpose of explaining the preference of human or animals given external stimuli, which marks a starting point for contemporary experimental aesthetics. In particular, Berlyne et al. found a close relationship between the complexity perceived by the viewers and their preferences [128,129]. Berlyne also indicated that perceived complexity is related with factors such as regularity/irregularity of the patterns, the amount of elements in the scene, and their heterogeneity [130].
2.4 Psychological Background: Emotion Representation

In psychology, emotion is represented using a categorical approach or a dimensional approach. Categorical representation distinctively classifies emotions into categories, such as anger, fear, disgust, amusement, awe, and contentment. Dimensional representation [48] describes emotions using three dimensions (Figure 2.2), i.e., valence, arousal, and dominance. Valence presents the positive or negative aspect of human emotions, where common emotions, like joy and happiness, are positive, whereas anger and fear are negative. Arousal describes the human physiological state of being reactive to stimuli. A higher value of arousal indicates higher excitation. Dominance tells the controlling nature of the emotion. For instance, anger can be more controlled than fear.

The development of dimensional emotions stems from the idea of two systems in the brain, i.e., appetitive and defensive, which refers to the primacy of valence and arousal dimensions [49]. Recent studies have indicated the strengths of dimensional approaches for studying human emotions. Bradley and Lang introduced the development and use of picture stimuli [50] provided by IAPS data sets [41], whose measurements followed the SAM (i.e., the self-assessment manikin [51]), a
culture-free and language-free instrument. According to Bradley and Lang [50], categorized emotions do not provide one-to-one relationships between the content and emotion of an image since different emotions may be evoked while viewing the same image. This highlights the utility of a dimensional approach, which controls for the intercorrelated nature of human emotions aroused by images. From the perspective of neuroscience studies, it has been demonstrated that the dimensional approach is more consistent with how the brain is organized to process emotions at their most basic levels [52,53]. Dimensional approaches also allow the separation of images with strong emotional content from images with emotionally neutral images\(^1\).

### 2.5 Psychological Background: Dataset

Psychological findings provide an intuitive understanding of visual stimuli that evoke affective responses, but the fact that those findings were made from small-scale studies makes the results less convincing. In order to investigate visual stimuli that evoke human’s affective responses from a more generalized perspective, psychologists created the standard International Affective Picture System (IAPS) [41] dataset. In particular, the IAPS dataset were developed by examining human affective responses to color photographs with varying degrees of emotional content. The IAPS dataset contains 1,182 images, wherein each image is associated with an empirically derived mean and standard deviation of valance, arousal, and dominance ratings. A small subset of images in the IAPS dataset are also associated with categorical emotion labels. Image examples in the IAPS dataset were shown in Figure 2.1. Images evoke positive affect are from left to right, and high arousal are from bottom to top.

Research on human responses to visual stimuli widely leverages the IAPS dataset [41]. Basic findings, including subjective, psychophysiological, behavioral, and neurophysiological reactions when viewing affective stimuli, were presented in [50]. In [54], evidence was provided that greater affective arousal leads to higher probability of response to emotionally evocative images. Barrett [55, 56] introduced the effect of individual differences on dimensional properties of emotions, i.e., valence focus and arousal focus. As presented in [55], increases in valence focus

\(^1\)In psychology, emotionally neutral images refer to images which evoke very weak or no emotions in humans.
lead to strong correlations between like-valenced emotions, and increases in arousal focus lead to the opposite.

In addition, most recent research work heavily replied on the IAPS to evaluate the effectiveness of computational algorithms on emotion recognition [6,16,17].

2.6 Summary

Existing studies in psychology provide foundations of emotion/aesthetics representation, validated image stimuli set, and findings on visual characteristics that evoke emotion and aesthetic feelings. In this chapter, we reviewed psychological studies that are closely related to the target of this dissertation, especially the visual characteristics of roundness, angularity, complexity, and holistic information and fine-grained details that evoke human’s affective responses.
Chapter 3

Background: Computational Modeling of Aesthetics and Emotions

3.1 Introduction

From a computational perspective, image aesthetics assessment and emotion recognition are both emerging research topics in computer vision and multimedia analysis. Image aesthetics assessment is a technique to identify the aesthetic value of a given pictorial scene. Emotion recognition is a technique to predict perceived emotion of human beings when provided with a visual stimuli. In this chapter, we review computational approaches of emotion recognition and image aesthetics assessment in Sections 3.2 and 3.3, respectively. In addition, we review deep learning approaches in Section 3.4 to facilitate understanding of the deep learning approaches developed in this dissertation for image aesthetics assessment.

3.2 Computational Background: Emotion Recognition

Previous work [3, 6, 9] predicted emotions aroused by images mainly through training classifiers on visual features to distinguish categorical emotions, such as happiness, anger, and sadness. Low-level stimuli such as color and composition have been widely used in computational modeling of emotions. Affective concepts were modeled using color palettes, which showed that the bag of colors and Fisher
vectors (i.e., higher order statistics about the distribution of local descriptors) were effective [5]. Zhang et al. [10] characterized shape through Zernike features, edge statistics features, object statistics, and Gabor filters. Emotion-histogram and bag-of-emotion features were used to classify emotions by Solli et al. [57]. These features were extracted based on findings from psychophysiological experiments, which indicated that homogeneous regions and transitions among them provide a hint for identifying evoked emotions.

The first work that comprehensively modeled categorical emotions, Machajdik and Hanbury [6] used color, texture, composition, content, and semantic level features such as the number of faces to model eight discrete emotional categories. In addition, to model categorized emotions, adjectives or word pairs were used to represent human emotions. The earliest work based on the Kansei system employed 23 word pairs (e.g., like-dislike, warm-cool, cheerful-gloomy) to establish the emotional space [7]. Along the same lines, researchers enumerated more word pairs to reach a universal, distinctive, and comprehensive representation of emotions in Wang et al. [8].

Most existing emotion recognition approaches were evaluated on small-scale datasets and did not establish a clear association between visual characteristics and human emotion. In this dissertation, we built a large image stimuli dataset and addressed the question of what specific visual characteristics might be associated with a certain emotion. Features we have designed are complementary to commonly adopted visual features, such as color, texture, and composition.

In the remaining subsections, we review commonly adopted features for emotion recognition.

### 3.2.1 Color

When modeling the factor of color, factors such as hue, saturation, brightness, and colorfulness are commonly investigated.

- **Hue, Saturation, and Brightness** Findings in psychology show that saturation and brightness directly affect the aroused emotion in the dimensions of pleasure (valence), arousal, and dominance [47]. In particular, Valdez and Mehrabian [47] presented an empirical formula to compute this influence as
Pleasure = 0.69Y + 0.22S ,
Arousal = − 0.31Y + 0.6S ,
Dominance = 0.76Y + 0.32S ,

where Y refers to brightness and S refers to saturation. As hue is an angular value, circular statistical descriptors were commonly utilized to describe the hue statistics of an image [131,132]. Denote by $H_i$ where $i = 1, \ldots, n$ as n hue values, mean direction $\bar{H}$ is computed as:

$$\bar{H} = \arctan \left( \frac{B}{A} \right) ,$$

where $A = \sum_i \cos H_i$ and $B = \sum_i \sin H_i$. Due to the close relationship between hue and saturation, saturation-weighted hue statistics is also used in describing hue statistics.

$$\bar{H}_S = \arctan \left( \frac{B_S}{A_S} \right) ,$$

where $A_S = \sum_i S_i \cos H_i$ and $B_S = \sum_i S_i \sin H_i$.

• **Color Distribution.** A color histogram is a typical representation of color distributions in an image, which quantizes image’s color spaces into several ranges and shows the number of pixels falling into each of the color ranges. The color histogram is commonly computed within the L*a*b or HSA space. An alternative approach to describe color distribution is to quantize the color space into some basic colors, such as black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow, and then compute a pixel-wise color distribution on these basic images. Weijer et al. trained a PLSA [134] model to learn color names from real-world images [133]. With the trained model, the probability of a color name given one pixel $P(z|w)$ could be computed as following:

$$P(z|w) \propto P(z)P(w|z) ,$$

where $z$ refers to each of the basic colors, and $w$ refers to a pixel. Denote by $P(z)$ the prior distribution of basic colors and $P(w|z)$ the probability of a pixel value generated by a basic color. $P(z)$ and $P(w|z)$ were learned when
training the PLSA model in [133].

- **Colorfulness.** Colorfulness was formulated utilizing the Earth Mover’s Distance (EMD) in [25] and commonly deployed in emotion/aesthetics prediction problems. The motivation is quantizing the color space into several ranges and forming an ideal color distribution of a ‘colorful’ image, i.e., the color equally distributed across all quantized ranges, and then comparing the real color distribution with the ideal colorful color distribution through the EMD metric. The more similar the two distribution is, the more colorful the image is. In [25], the color space were divided as 64 cubic blocks, denote by $D_1$ the ideal colorful color distribution and $D_2$ the real color distribution. The colorfulness is then computed as:

$$
\text{colorfulness} = \text{emd}(D_1, D_2, \{d(a, b) \mid 0 \leq a, b \leq 63\}) ,
$$

(3.5)

where $d(a, b) = ||rgb2luv(c_a) - rgb2luv(c_b)||$, and $c_a$ and $c_b$ are the geometric centers of each cube $i$.

### 3.2.2 Texture

Texture in images represents visual content and meanwhile have an impact on evoked emotion. We review typical textual features utilized in both image content analyses and aesthetics/emotion analyses in this subsection.

- **Wavelet-based features.** Wavelet-based features were commonly used to describe the smoothness in images [6,25,135]. In [6,25], a three-level wavelet transform on all color channels was computed as following:

$$
f_i = \frac{\sum_{x,y} w_i^h(x,y) + \sum_{x,y} w_i^v(x,y) + \sum_{x,y} w_i^d(x,y)}{|w_i^h| + |w_i^v| + |w_i^d|},
$$

(3.6)

where $w_i^h$, $w_i^v$, and $w_i^d$ refer to coefficients of wavelet transform in level $i$ ($i = 1, 2, 3$ in [6]).

- **Coarseness, Contrast, and Directionality.** Coarseness, Contrast, and Directionality are the first three features defined in [136,137], corresponding to human visual perception.
Coarseness intends to identify the scale of a texture. For each pixel \((x, y)\), an average \(A_k(x, y)\) over its neighborhood is first computed in Equation 3.2.2.

\[
A_k(x, y) = \frac{\sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}}^{y+2^{k-1}-1} I(i, j)}{2^{2k}}.
\]  

(3.7)

Then at each pixel \((x, y)\), differences between non-overlapping neighborhoods in terms of \(A_k(x, y)\) are computed both horizontally and vertically.

\[
E_{k,h}(x, y) = |A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y)|,
\]

\[
E_{k,v}(x, y) = |A_k(x, y + 2^{k-1}) - A_k(x, y - 2^{k-1})|.
\]

(3.8)

Coarseness is then the computed in Equation 3.2.2.

\[
\text{coarseness} = \frac{1}{N} \sum_x \sum_y 2^{k_{opt}},
\]

(3.9)

where \(k_{opt} = \max_k \{E_{k,h}(x, y), E_{k,v}(x, y)\}\), and \(N\) is the total number of pixels in the image.

Contrast measures how the intensity varies in an image. Denote by \(q\) the unique intensity value in an image \(I\) and \(\mu\) the mean intensity value of \(I\), and then contrast is computed as following:

\[
\text{contrast} = \frac{\sigma}{(\alpha_4)^n},
\]

(3.10)

where \(n = 0.25\), \(\sigma^2 = \sum_q (q - \mu)^2 \Pr(q|I)\), and \(\alpha_4 = \frac{1}{\sigma^4} \sum_{q=0}^{q_{max}} (q - \mu)^4 \Pr(q|I)\).

Directionality describes a distribution of degree of directionality of an image. The histogram distribution \(H_{dir}(a)\) is computed by examining the edge strength \(e(x, y)\) and the directional angle \(a(x, y)\), computed in Equation 3.11.

\[
e(x, y) = 0.5(|\nabla_x (x, y)| + |\nabla_y (x, y)|),
\]

\[
a(x, y) = \tan^{-1} \frac{\nabla_y (x, y)}{\nabla_x (x, y)},
\]

(3.11)
where \( \nabla_x (x, y) = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \) and \( \nabla_y (x, y) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \).

- **Gray-Level Co-occurrence Matrix (GLCM).** Contrast, correlation, energy, and homogeneity of an image were commonly computed given the co-occurrence matrix [138]. Gray-level co-occurrence matrix is a basic measure to capture co-occurrence of adjacent pixels in terms of intensity. Given the co-occurrence matrix, contrast, correlation, energy, and homogeneity are computed as following:

\[
\begin{align*}
\text{Contrast} &= \sum_i \sum_j (i - j)^2 c_{ij}, \\
\text{Entropy} &= -\sum_i \sum_j c_{ij} \log c_{ij}, \\
\text{Energy} &= \sum_i \sum_j c_{ij}^2, \\
\text{Homogeneity} &= \sum_i \sum_j \frac{c_{ij}}{1 + |i - j|},
\end{align*}
\]

(3.12)

where \( c_{ij} \) refers to the element value in the co-occurrence matrix.

### 3.2.3 Composition

Composition is an essential factor in understanding image semantics and aesthetics/emotions aroused from images. We briefly review three representative computational descriptions of image composition: Low Depth of Field (DOF), Dynamics, and Rule of Thirds.

- **Low Depth of Field (DOF).** An intentional blur may make an image more appealing [25]. In image aesthetics/emotion literatures [6, 25], images are divided into 16 equal rectangular blocks \( \{M_1, \ldots, M_{16}\} \) in row-major order. DOF features are computed as Equation 3.13.

\[
f_i = \frac{\sum_{(x,y) \in M_6 \cup M_7 \cup M_{10} \cup M_{11}} w^i_3(x,y)}{\sum_{i=1}^{16} \sum_{(x,y) \in M_i} w^i_3(x,y)},
\]

(3.13)

where denote by \( w^i_3 \) the set of wavelet coefficients in the level 3 (high-frequency level) of the image, and \( i = \{H, S, V\} \).
• **Dynamics.** Psychological findings showed that horizontal lines convey calmness and peacefulness; vertical lines show direct and communicate dignity; and slant lines indicate unstable and dynamism [6]. Computationally, researchers [6] first detected important lines using Hough transform, then classified those detected lines into static or slant lines according to their tilt angle $\theta$. Lines are treated as static if $-15^\circ < \theta < 15^\circ$ or if $75^\circ < \theta < 105^\circ$, and are treated as slant otherwise.

• **Rule of Thirds.** The principle of rule of thirds could be treated as an approximation to the golden ratio [25]. In [25], an image was uniformly divided into nine equal rectangular blocks, the rule of thirds feature is computed by averaging pixel values in the central cubic in the H, S, V channels, respectively (shown in Equation 3.14).

\[
f_i = \frac{9}{XY} \sum_{x=X/3}^{2X/3} \sum_{y=Y/3}^{2Y/3} I_i(x,y),
\]

where $X, Y$ are the width and height of an image and $i = \{H, S, V\}$.

In addition, facial expression in images could have an influence on aroused emotion, facial features, such as number of frontal faces has been included in emotion recognition tasks [6]. More facial features could be included in the future with the development of algorithms in facial emotion recognition.

### 3.3 Computational Background: Image Aesthetics Assessment

Earlier visual aesthetics assessment research focused on examining hand-designed visual features based on common cues such as color [18, 19, 25], texture [25, 26], composition [15, 28, 29], and content [28, 29], as well as generic image descriptors [22, 23, 58]. Commonly investigated color features include lightness, colorfulness, color harmony, and color distribution [18, 19, 25]. Texture descriptors vary from wavelet-based texture features [25], distribution of edges, to blur descriptors and shallow depth-of-field descriptors [26]. Composition features typically include the rule of thirds, size and aspect ratio [28], and foreground and background composition [15,
There have been attempts to represent the content of images using people and portrait descriptors [28, 29], scene descriptors [29], and generic image features such as SIFT [20], GIST [21], and Fisher Vector [22–24].

Despite the success of handcrafted and generic visual features, the usefulness of automatically learned features have been demonstrated in many vision applications [30, 59–61]. Recently, trained deep neural networks were used to build and associate mid-level features with class labels. Convolutional neural network (CNN) [62] is one of the most powerful learning architectures among the various types of neural networks (e.g., Deep Belief Net [63] and Restricted Boltzmann Machine [64]). Krizhevsky et al. [30] significantly advanced the 1000-class classification task in ImageNet challenge with a deep architecture of CNN in conjunction with dropout and normalization techniques, Sermanet et al. [61] achieved the-state-of-the-art performance on all major pedestrian detection datasets, and Ciresan et al. [59] reached a near-human performance on the MNIST\(^1\) dataset.

Few studies have examined the potential of CNN for image aesthetics assessment. As image aesthetics depends on both the global and local characteristics, in this dissertation, we addressed the general problems of learning holistic information and fine-grained details from natural images. Methods we have developed in this dissertation are applicable to applications that highly depend on information extracted from multiple image scales.

In particular, we categorize features designed for image aesthetics assessment into three divisions: feature designed following principles in scene vision, photographic practices, and psychological findings.

- **Scene Vision.** Following principles in scene vision, the developed features for image aesthetics assessment covers descriptors of color and texture utilizing channels of hue, saturation, brightness, and wavelet-based transforms. Features in this category are shared among aesthetics assessment and aroused emotion recognition. As we have detailed those features in Sections 3.2.1 and 3.2.2, we will not repeat it here.

- **Photographic Practices.** Many hand-designed features for image aesthetics could date back to photographic practices, such as the rule of thirds, the golden ratio, low depth of field, and colorfulness. Similar to features in the

\(^1\)http://yann.lecun.com/exdb/mnist/.
first category, the effectiveness of these features have also been demonstrated in both aesthetics prediction and emotion recognition. Computation details were presented in Sections 3.2.1 and 3.2.3.

- **Psychological Findings.** In addition, according to psychological findings, size and aspect ratio and shape complexity, and region composition have an impact on image aesthetics assessment. These features were briefly discussed in [25]. In this dissertation, we were inspired by these psychological findings, and studied (1) shape complexity for emotion recognition by developing 219 dimension shape features (Chapter 4); (2) region composition to model overall simplicity of an image (Chapter 5); (3) the size and aspect ratio by developing deep learning approaches for image aesthetics assessment (Chapter 6).

### 3.4 Computational Background: Deep Learning Approaches

As designing handcrafted features has been widely considered an appropriate approach in assessing image aesthetics, insufficient effort has been devoted to automatic feature learning on a large collection of labeled ground-truth data. The recently-developed AVA dataset [23] contains roughly 250,000 images with aesthetic ratings and a 14,000 subset with style labels (e.g., rule of thirds, motion blur, and complementary colors), making automatic feature learning using deep learning approaches possible.

The break-through made by the deep neural network approach for image classification [30] has inspired hundreds of follow-on studies on deep learning and their applications to vision. We humbly review the studies that are closely related to this dissertation.

Recent work [65–72] focused on adapting deep neural network training to various vision applications. They mostly were able to show some improvement with slight modification of the network structures (e.g., adding a layer or adding a column) or changing the training strategy (e.g., fine-tuning). In addition to the useful techniques such as ReLU, dropout, and data augmentation introduced in [30], we noticed two key ideas that have led to promising results in classification problems:

1. **Multiple Image Resolutions.** Different vision applications require informa-
tion from different image resolution. In image classification, deep convolutional neural network (CNN) achieved great success by training on $256 \times 256 \times 3$ images. However, in image aesthetics, training deep neural networks on relatively high-resolution images helps improve the performance significantly because image aesthetics depend on both holistic and local cues.

When using imagenet feature as a generic image descriptor for recognition tasks, researchers found that aggregating descriptors from different image resolutions helps boost the classification performance. In [74–76], researchers computed ImageNet features (i.e., features extracted by the neural network trained in the ImageNet Challenge [30]) from multi-scale image pyramid for object recognition, scene recognition, and object detection. In [77], Gong et al. improved the geometric invariance of imagenet activations by pooling features extracted from multi-resolution patches. In addition, recent study has demonstrated the significance of maintaining the aspect ratio of images through spatial pyramid pooling [78] in object detection.

2. Multi-Column Neural Network. Multi-column neural network [59, 73, 79] has been demonstrated as an efficient approach to improve performance of single-column neural networks in various classification problems. Motivated by part-based approaches (e.g., [80–83]), recent studies attempted to train multiple convolutional neural networks on aligned parts. In [84], Zhang et al. trained pose-normalized CNNs on semantically aligned part patches, whose learned features are associated with certain parts under specific poses. A similar approach has been applied to fine-grained category detection in [85]. In [86], Bourdev et al. applied trained CNN to extract features on poselet patches.

In this dissertation, we trained deep neural networks on the AVA dataset to categorize image aesthetic quality. Specifically, we built upon [30] and proposed a double-column CNN architecture to automatically discover effective features that capture image aesthetics from two heterogeneous input sources. The proposed architecture is different from the recent work in multi-column neural networks [59, 79]. Agostinelli et al. [79] extended stacked sparse autoencoder to a multi-column version by computing the optimal column weights and applied the model to image denoising. Ciresan et al. [59] averaged the output of several columns trained on inputs with different standard preprocessing methods. Our architecture is different from that work because the two columns in our architecture are jointly trained using two different inputs: The first column of the network takes global image representation
as the input, while the second column takes local image representations as the input. This strategy allows us to leverage both compositional and local visual information. In addition, we proposed regularized double-column CNN architecture to learn and predict image aesthetics by leveraging additional style or semantic information.

### 3.4.1 Feed Forward Neural Networks

The basic feed forward neural network models includes a series functional transformations. Given input variables \(x_1, \ldots, x_D\), the first layer output is constructed by a linear transform of the input and a nonlinear activation function. The linear transform function on input variables is shown in Equation 3.15.

\[
a_j = \sum_{i=1}^{D} w^{(1)}_{ji} x_i + b^{(1)}_{j0},
\]

where denote by \(j = 1, \ldots, M\) the index of output neurons, \(w_{ji}(1)\) the weights and \(b^{(1)}_{j0}\) the biases. The superscript \((1)\) indicates the weights and the biases are in the first layer of the network.

A classic activation function to model a neuron’s output is through \(tanh\) function \((h(x) = tanh(x))\) or \(sigmoid\) function \((\sigma(x) = (1 + \exp^{-x})^{-1})\). The breakthrough made in [30] showed that a non-saturating nonlinearity function \(f(x) = \max(0, x)\) greatly reduces training time compared with typical activation functions such as \(tanh\) and \(sigmoid\). This function was called Rectified Linear Units (ReLUs).

When applying ReLU in a two-layer feed forward network given the input variables \(x_1, \ldots, x_D\), the output is

\[
y_k(x, w, b) = f\left(\sum_{j=1}^{M} w^{(2)}_{kj} f\left(\sum_{i=1}^{D} w^{(1)}_{ji} x_i + b^{(1)}_{j0}\right) + b^{(2)}_k\right),
\]

where denote by \(M\) the number of output in the first layer.

The goal of network training is to learn the weights and biases in all network layers. Parameters are commonly optimized through backpropagation. Since backpropagation utilizes the gradient descent method, the derivative of the loss function is computed with respect to the weights of the network and backpropagated following chain rule.
3.4.2 Regularization in Neural Networks

One of the major practical difficulties of network training is overfitting due to the large collection of parameters. In addition to reduce the number of neurons and number of layers of the network, we review the two techniques that contribute to the breakthrough made in [30] for image classification utilizing deep neural network training.

- **Data Augmentation** Training with transformed data has long been treated as an effective technique to avoid network training [140]. In [30], when training convolutional neural networks, images were randomly cropped with a small margin and flipped horizontally with probability of 0.5. Another data augmentation strategy performed in [30] was altering the intensities of the RGB channels in training images by analyzing the principle components derived from the entire training images.

- **Dropout.** Dropout is a recently-developed technique to regularize the network and avoid overfitting [30,139] by setting the output of hidden neurons to zero with a certain probability, commonly 0.5. When dropout is applied, the real network structure in each training epoch is a subsampling of the original network structure setting, whereas the weights were shared in training. This strategy reduces the dependency among neurons during training and forces the network to learn more robust network parameters. In testing, a common strategy is to multiply outputs of all neurons with 0.5 to approximate the geometric mean of the predictive distributions produced by the network utilizing dropout.

3.5 Summary

It is the goal of researchers working on areas of image aesthetics and emotions to endow computers with the capability of perceiving aesthetics and emotions as human vision systems do. In this chapter, we reviewed and discussed approaches developed so far for automatic aesthetics assessment and emotion recognition given a visual stimuli. We also reviewed recent development in deep learning to facilitate understanding of proposed deep learning approach for aesthetics prediction in this dissertation.
Chapter 4

On Shape and the Computability of Emotions

4.1 Introduction

Shapes and their characteristics such as roundness, angularity, and simplicity have been postulated to affect emotional responses of human beings in the field of visual arts and psychology. In this section, we investigated how shape features in natural images influence emotions aroused in human beings. Our contributions include an in-depth statistical analysis to understand the relationship between shapes and emotions. Through experimental results on the International Affective Picture System (IAPS) dataset we provide evidence for the significance of roundness-angularity and simplicity-complexity on predicting emotional content in images. We combined our shape features with other state-of-the-art features to show a gain in prediction and classification accuracy. We model emotions from a dimensional perspective in order to predict valence and arousal ratings which have advantages over modeling the traditional discrete emotional categories. Finally, we developed a method to distinguish images with strong emotional content from emotionally neutral images with high accuracy.
4.2 Shape Features

For decades, numerous theories have been promoted about the relationship between emotions and the visual characteristics of simplicity, roundness, and angularity. Despite these theories, researchers have yet to resolve how to model these relationships quantitatively. In this dissertation, we proposed shape features to capture those visual characteristics. By identifying the link between shape features and emotions, we are able to determine the relationship between the aforementioned visual characteristics and emotions.

Shapes in images are difficult to capture, mainly due to the perceptual and merging boundaries of objects which are often not easy to differentiate using even state-of-the-art segmentation or contour extraction algorithms. In contemporary computer vision literatures [88, 89], there are a number of statistical representations of shape through characteristics like the straightness, sinuosity, linearity, circularity, elongation, orientation, symmetry, and the mass of a curve. We chose roundness, angularity, and simplicity characteristics because they have been found previously by psychologists to influence the affect of human beings through controlled human subject studies.

To make it more convenient to introduce the shape features proposed, we first define the four terms used: line segments, angles, continuous lines, and curves. The
framework for extracting perceptual shapes through lines and curves is derived from [90]. The contours are extracted using the algorithm in [91], which used color, texture, and brightness of each image for contour extraction. The extracted contours are of different intensities and indicate the algorithm’s confidence on the presence of edges. Considering the temporal resolution of our vision system, we adopted a threshold of 40%. Example results are presented in Figures 4.1, 4.2, 4.3, and 4.4. Pixels with an intensity higher than 40% are treated equally, which results in the binary contour map presented in the second column. The last three columns show the line segments, continuous lines, and curves.

**Line segments** - Line segments refer to short straight lines generated by fitting nearby pixels. We generated line segments from each image to capture its structure. From the structure of the image, we propose to interpret the simplicity. We extracted locally optimized line segments by connecting neighboring pixels from
the contours extracted from the image [92].

**Angles** - Angles in the image are obtained by calculating angles between each of any two intersecting line segments extracted previously. According to Julian Hochberg’s theory [11], the number of angles and the number of different angles in an image can be effectively used to describe its simplicity. The distribution of angles also indicates the degree of angularity of the image. A high number of acute angles makes an image more angular.

**Continuous lines** - Continuous lines are generated by connecting intersecting line segments having the same orientations with a small margin of error. Line segments of inconsistent orientations can be categorized as either corner points or points of inflexion. Corner points, shown in Figure 4.5(a), refer to angles that are lower than 90 degrees. Inflexion points, shown in Figure 4.5(b), refer to the midpoint of two angles with opposite orientations. Continuous lines and the degree of curving can be used to interpret the simplicity of the image.

**Curves** - Curves are a subset of continuous lines, the collection of which are employed to measure the roundness of an image. To achieve this, we consider each curve as a section of an ellipse, thus we use ellipses to fit continuous lines. Fitted
curves are represented by parameters of its corresponding ellipses.

We now present the details of the proposed shape features: line segments, angles, continuous lines, and curves. A total of 219 shape features are summarized in Table 4.1.

<table>
<thead>
<tr>
<th>Category</th>
<th>Short Name</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line Segments</td>
<td>Orientation</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Length</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Mass of the image</td>
<td>4</td>
</tr>
<tr>
<td>Continuous Lines</td>
<td>Degree of curving</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Length span</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Line count</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mass of continuous lines</td>
<td>4</td>
</tr>
<tr>
<td>Angles</td>
<td>Angle count</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Angular metrics</td>
<td>35</td>
</tr>
<tr>
<td>Curves</td>
<td>Fitness</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Circularity</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Area</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Orientation</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Mass of curves</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Top round curves</td>
<td>18</td>
</tr>
</tbody>
</table>

### 4.2.1 Line Segments

Psychologists and artists have claimed that the simplicity of an image is determined not only by lines or curves, but also by its overall structure and support [11]. Based on this idea, we employed line segments extracted from images to capture their structure. Particularly, we used the orientation, length, and mass of line segments to determine the simplicity of the images.

**Orientation** - To capture an overall orientation, we employed statistical measures of minimum (min), maximum (max), 0.75 quantile, 0.25 quantile, the difference between 0.75 quantile and 0.25 quantile, the difference between max and min, sum, total number, median, mean, and standard deviation (we will later refer to these as {statistical measures}), and entropy. We experimented with both 6- and 18-bin histograms. The unique orientations were measured based on the two histograms to capture the simplicity of the image.
Among all line segments, horizontal and vertical lines are known [11] to be static and to represent the feelings of calm and stability within the image. Horizontal lines suggest peace and calm, whereas vertical lines indicate strength. To capture the emotions evoked by these characteristics, we counted the number of horizontal lines and vertical lines through an 18-bin histogram. The orientation $\theta$, of horizontal lines fall within $0^\circ < \theta < 10^\circ$ or $170^\circ < \theta < 180^\circ$, and $80^\circ < \theta < 100^\circ$ for vertical lines.

**Length** - The length of line segments reflects the simplicity of images. Images with simple structure might use long lines to fit contours, whereas complex contours have shorter lines. We characterized the length distribution by calculating the {statistical measures} of lengths of line segments within the image.

**Mass of the image** - The centroid of line segments may indicate associated relationships among line segments within the visual design [11]. Hence, we calculate the mean and standard deviation of the $x$ and $y$ coordinates of the line segments to find the mass of each image.

Some of the example images and their features are presented in Figures 4.6 and 4.7. Figure 4.6 presents the ten lowest mean values of the length of line segments. The first row shows the original images, the second row shows the line segments extracted from these images and the third row shows the 18-bin histogram for line segments in the images. The 18 bins refer to the number of line segments with an orientation of $[-90 + 10(i - 1), -90 + 10i]$ degrees where $i \in \{1, 2, ..., 18\}$. Similarly, Figure 4.7 presents the ten highest mean values of the length of line segments.

These two figures indicate that the length or the orientation cannot be examined separately to determine the simplicity of the image. Lower mean values of the length of line segments might refer to either simple images such as the first four images in Figure 4.6 or highly complex images such as the last four images in that figure. The histogram of the orientation of line segments helps us to distinguish the complex images from simple images by examining variation of values in each bin.

### 4.2.2 Angles

Angles are important elements in analyzing the simplicity and the angularity of an image. We capture the visual characteristics from angles through two perspectives.
Figure 4.6: Images with low mean value of the length of line segments and their associated orientation histograms. The first row is the original images; the second row shows the line segments; and the third row shows the 18-bin histogram for line segments in the images.

Figure 4.7: Images with high mean value of the length of line segments and their associated orientation histograms. The first row is the original images; the second row shows the line segments; and the third row shows the 18-bin histogram for line segments in the images.

- **Angle count** - We first calculate the two quantitative features claimed by Julian Hochberg, who has attempted to define simplicity (he used the value-laden term “figural goodness”) via information theory: “The smaller the amount of information needed to define a given organization as compared to the other alternatives, the more likely that the figure will be so perceived” [11]. Hence this minimal information structure is captured using the number of angles and the percentage of unique angles in the image.

- **Angular metrics** - We use the {statistical measures} to extract angular metrics. We also calculate the 6- and 18-bin histograms on angles and their entropies.

Some of the example images and features are presented in Figures 4.8 and 4.9. Images with lowest and highest number of angles are shown along with their corresponding contours in Figure 4.8. These examples show promising relationships between angular features and simplicity of the image. Example results for the
Figure 4.8: Images with highest and lowest number of angles.

Figure 4.9: The distribution of angles in images.

The histogram of angles in the image are presented in Figure 4.9. The 18 bins refer to the number of line segments with an orientation in \([10(i-1), 10i)\) degrees where \(i \in \{1, 2, ..., 18\}\).

4.2.3 Continuous Lines

We attempt to capture the degree of curvature from continuous lines, which has implications for the simplicity of images. We also calculated the number of continuous lines, which is the third quantitative feature specified by Julian Hochberg [11]. For continuous lines, open/closeness are factors affecting the simplicity of an image. In the following, we focus on the calculation of the degree of curving, the length span value, and the number of open lines and closed lines. The length span refers to the highest Euclidean distance among all pairs of points on the continuous lines.

\[
\text{Length Span}(l) = \max_{p_i \in l, p_j \in l} \text{EuclideanDist}(p_i, p_j), \tag{4.1}
\]

where \(\{p_1, p_2, ..., p_N\}\) are the points on continuous line \(l\).
Degree of curving - We calculated the degree of curving of each line as

\[
\text{Degree of Curving}(l) = \frac{\text{Length Span}(l)}{N},
\]

where \( N \) is the number of points on continuous line \( l \).

To capture the statistical characteristics of contiguous lines in the image, we calculated the {statistical measures}. We also generated a 5-bin histogram on the degree of curving of all continuous lines (Figures 4.10 and 4.11).

- **Length span** - We used {statistical measures} for the length span of all continuous lines.

- **Line count** - We counted the total number of continuous lines, the total number of open lines, and the total number of closed lines in the image.

### 4.2.4 Curves

We used the nature of curves to model the roundness of images. For each curve, we calculated the extent of fit to an ellipse as well as the parameters of the ellipse such as its area, circularity, and mass of curves. The curve features are explained in detail below.

- **Fitness, area, circularity** - The fitness of an ellipse refers to the overlap between the proposed ellipse and the curves in the image. The area of the
Table 4.2: Average number of curves in terms of the value of fitness in positive and negative images.

<table>
<thead>
<tr>
<th></th>
<th>(0.8,1]</th>
<th>(0.6,0.8]</th>
<th>(0.4,0.6]</th>
<th>(0.2,0.4]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive imgs</td>
<td>2.12</td>
<td>9.33</td>
<td>5.7</td>
<td>2.68</td>
</tr>
<tr>
<td>Negative imgs</td>
<td>1.42</td>
<td>7.5</td>
<td>5.02</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Table 4.3: Average number of curves in terms of the value of circularity in positive and negative images.

<table>
<thead>
<tr>
<th></th>
<th>(0.8,1]</th>
<th>(0.6,0.8]</th>
<th>(0.4,0.6]</th>
<th>(0.2,0.4]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive imgs</td>
<td>0.96</td>
<td>2.56</td>
<td>5.1</td>
<td>11.2</td>
</tr>
<tr>
<td>Negative imgs</td>
<td>0.73</td>
<td>2.19</td>
<td>4</td>
<td>9.75</td>
</tr>
</tbody>
</table>

The fitted ellipse is also calculated. The circularity is represented by the ratio of the minor and major axes of the ellipses. The angular orientation of the ellipse is also measured. For each of the measures, we used the statistical measures and entropies of the histograms as the features to depict the roundness of the image.

- **Mass of curves** - We used the mean value and standard deviation of \((x, y)\) coordinates to describe the mass of curves.

- **Top round curves** - To make full use of the discovered curves and to depict roundness, we included the fitness, area, circularity, and mass of curves for each of the top three curves.

To examine the relationship between curves and positive-negative images, we calculated the average number of curves in terms of values of circularity and fitness on positive images (i.e., the value is higher than 6 in the dimension of valance) and negative images (i.e., the value is lower than 4.5 in the dimension of valance).

The results are shown in Tables 4.2 and 4.3. Positive images have more curves with 60% – 100% fitness to ellipses and higher average curve count.
4.3 Evaluation

To demonstrate the relationship between proposed shape features and the felt emotions, the shape features were utilized in three tasks. First, we distinguished images with strong emotional content from emotionally neutral images. Second, we fit valence and arousal dimensions using regression methods. We then performed classification on discrete emotional categories. The proposed features were compared with the features discussed in Machajdik et al. [6], and overall accuracy was quantified by combining those features. Forward selection and Principal Component Analysis (PCA) strategies were employed for feature selection and to find the best combination of features.

4.3.1 Dataset

We used two subsets of the IAPS [41] dataset:

Subset A of the IAPS dataset includes many images with faces and human bodies. Facial expressions and body language strongly affect emotions aroused by images, slight changes of which might lead to an opposite emotion. The proposed shape features are sensitive to faces hence we removed all images with faces and human bodies from the scope of this study. In experiments, we only considered the remaining 484 images, which we labeled as Subset A. To provide a better understanding of the ratings of the dataset, we analyzed the distribution of ratings.
Within valence and arousal, as shown in Figure 4.12. We also calculated average variations of ratings in each rating unit (i.e., 1-2, 2-3, ..., 7-8). Valence ratings between 3 and 4, and 6 and 7, have the maximum variance for single images. Similarly, arousal ratings between 4 and 5 varied the most.

Subset B are images with category labels (with discrete emotions), generated by Mikels [96]. Subset B includes eight categories namely, anger, disgust, fear, sadness, amusement, awe, contentment, and excitement, with 394 images in total. Subset B is a commonly used dataset, hence we used it to benchmark our classification accuracy with the results mentioned in Machajdik et al. [6].

### 4.3.2 Identifying Strong Emotional Content

Images with strong emotional content have very high or very low valance and arousal ratings. Images with values around the mean values of valance and arousal lack emotions and were used as samples for emotionally neutral images.

Based on dimensions of valance and arousal respectively, we generated two sample sets from Subset A. In Set 1, images with valence values higher than 6 or lower than 3.5 were considered images with strong emotional content and the rest to represent emotionally neutral images. This resulted in 247 emotional images and 237 neutral images. Similarly, images with arousal values higher than 5.5 or lower than 3.7 were defined as emotional images, and others as neutral images. With similar thresholds, we obtained 239 emotional images and 245 neutral images in Set 2.

We used the traditional Support Vector Machines (SVM) with radial basis
function (RBF) kernel to perform the classification task. We trained SVM models using the proposed shape features, Machajdik’s features, and combined (Machajdik’s and shape) features. Training and testing were performed by dividing the dataset uniformly into training and testing sets. As we removed all images with faces and human bodies, we did not consider facial and skin features discussed in [6]. We used both forward selection and PCA methods to perform feature selection. In the forward selection method, we used the greedy strategy and accumulated one feature at a time to obtain the subset of features that maximized the classification accuracy. The seed features were also chosen at random over multiple iterations to obtain better results. Our analyses showed that the forward selection strategy achieved greater accuracy for Set 2, whereas PCA performed better for Set 1 (Figure 4.13). The feature comparison showed that the combined (Machajdik’s and shape) features achieved the highest classification accuracy, whereas individually the shape features alone were much stronger than the features from [6] (Machajdik’s features). This result is intuitive since emotions evoked by images cannot be well represented by shapes alone and can definitely be bolstered by other image features including their color composition and texture.

By analyzing valence and arousal ratings of the correctly classified images, we observed that very complex/simple, round and angular images had strong emotional content and high valence values. Simple structured images with very low degrees of curving also tends to portray strong emotional content as well as to have high arousal values.
By analyzing the individual features for classification accuracy we found that line count, fitness, length span, degree of curving, and the number of horizontal lines achieved the best classification accuracy in Set 1. Fitness and line orientation were more dominant in Set 2.

We present a few example images, which were wrongly classified based on the proposed shape features in Figures 4.14 and 4.15. The misclassification can be explained as a shortcoming of the shape features in understanding the semantics. Some of the images generated extreme emotions based on image content irrespective of the low-level features. Besides the semantics, our performance was also limited by the performance of the contour extraction algorithm.

4.3.3 Fitting Dimensional Emotions

Emotions can be represented by word pairs, as previously done in [7]. However, some emotions are difficult to label. Modeling basic emotional dimensions helps in alleviating this problem. We represented emotion as a tuple consisting of valence and arousal values. The values of valence and arousal were in the range of (1, 9). In order to predict the values of valence and arousal we proposed to learn a regression model for either dimension separately.

We used SVM regression with RBF kernel to model the valance and arousal values using shape, Machajdik’s features, as well as the combination of features. The mean squared error (MSE) was computed for each of the individual features as well as combined for both valence and arousal values separately. The MSE
Figure 4.16: Experimental results. (a) Mean squared error for the dimensions of valance and arousal. (b) Accuracy for the classification task.

values are shown in Figure 4.16(a). These figures show that the valance values were modeled more accurately by Machajdik’s features than our shape features. Arousal was well modeled by shape features with a mean squared error of 0.9. However, the combined feature performance did not show any improvements. The results indicated that visual shapes provide a stronger cue in understanding the valence as opposed to the combination of color, texture, and composition in images.

We also computed the correlation between quantified individual shape features and valence-arousal ratings. The higher the correlation, the more relevant the features were. Through this process we found that angular count, fitness, circularity, and orientation of line segments showed higher correlations with valance, whereas angle count, angle metrics, straightness, length span, and orientation of curves had higher correlations with arousal.

### 4.3.4 Classifying Categorized Emotions

To evaluate the relationship between shape features and emotions on discrete emotions, we classified images into one of the eight categories, anger, disgust, fear, sadness, amusement, awe, contentment, and excitement. We followed Machajdik et al. [6] and performed one-vs-all classification to compare and benchmark our classification accuracy. The classification results are reported in Figure 4.16(b). We used SVM to assign the images to one of the eight classes. The highest accuracy was obtained by combining Machajdik’s with shape features. We also observed a considerable increase in the classification accuracy by using the shape features alone, which proves that shape features indeed capture emotions in images more
effectively.

Table 4.4: Significant Features for Emotion Classification

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>Circularity</td>
</tr>
<tr>
<td>Disgust</td>
<td>Length of line segments</td>
</tr>
<tr>
<td>Fear</td>
<td>Orientation of line segments and angle count</td>
</tr>
<tr>
<td>Sadness</td>
<td>Fitness, mass of curves, circularity, and orientation of line segments</td>
</tr>
<tr>
<td>Amusement</td>
<td>Mass of curves and orientation of line segments</td>
</tr>
<tr>
<td>Awe</td>
<td>Orientation of line segments</td>
</tr>
<tr>
<td>Excitement</td>
<td>Orientation of line segments</td>
</tr>
<tr>
<td>Contentment</td>
<td>Mass of lines, angle count, and orientation of line segments</td>
</tr>
</tbody>
</table>
images. In the future, we hope to expand our experimental dataset and provide stronger evidence of established relationships between shape features and emotions.
A Computational Investigation into Three Visual Characteristics of Complex Scenes that Evoke Human Emotion

5.1 Introduction

Prior computational studies have examined hundreds of visual characteristics related to color, texture, and composition in an attempt to predict human emotional responses. Beyond those myriad features examined in the computer sciences, roundness, angularity, and visual simplicity have also been found to evoke emotions in human perceivers, as demonstrated in psychological studies of facial expressions, dancing poses, and even simple synthetic visual patterns. Capturing these characteristics algorithmically to incorporate in computational studies, however, has proven difficult. Here we expand the scope of previous computer vision work by examining these three visual characteristics in computer analysis of complex scenes, and compare the results to the hundreds of visual qualities previously examined. A large collection of ecologically valid stimuli (i.e., photographs humans regularly encounter on the web), containing more than 40,000 images crawled from web albums, were generated using crowdsourcing and subjected to human subject emotion ratings. We developed computational methods to the separate indices of roundness, angularity, and simplicity, thereby establishing three new computational constructs. Critically, these three new visual constructs achieve comparable
classification accuracy to the hundreds of shape, texture, composition, and facial feature characteristics previously examined. In addition, our experimental results show that roundness, angularity, and simplicity related similarly with both of these emotions dimensions.

A large collection of ecologically valid stimuli (named the EmoSet) were introduced in detail in this chapter. Unlike the commonly used The International Affective Picture System (IAPS), where 1082 images were rated with emotional responses, the EmoSet contains complex scenes that humans regularly encounter in life. Up to now, we have collected about 43,837 images with emotional labels (both dimensional and categorical labels) following strict psychological subject study procedures and validation approaches. The images were crawled from more than 1000 users’ Web albums, and associated with about 2K semantic tags and 558 emotional words. The EmoSet marks the first time a human subject study, examining perceived emotion triggered by visual stimuli, was performed in uncontrolled environment using natural photographs of complex scenes. The EmoSet is also different from large-scale affective datasets introduced in [23, 98], where researchers crawled user-generated content, including pictorial ratings or associated affective tags, to indicate the affective intention of images. Whereas those datasets were of a large scale, the emotional labels were not generated by human subjects under strict psychological procedures. For instance, psychological conventions were applied when recording human perceived emotions, i.e., the aroused emotions are best recorded within 6 seconds after subjects view each visual stimulus [41].

5.2 The EmoSet

To have a large collection of photographs with complex scenes, we crawled more than 50K images from Flickr, one of the most popular web albums. We performed a human subject study on those photographs and developed a large-scale set of ecologically valid image stimuli, i.e., The EmoSet. This human subject study was empowered by crowdsourcing and computational tools, where we incorporated strict psychological procedures into the User Interface (UI) design, in order to recruit a

---

1 The 558 emotional words were summarized by Averill [97]. We used the 558 words to trigger Flickr and crawled images. We introduced the data collection in the approach Section.

2 This human subjects study was IRB approved.

As a result, the EmoSet contains 43,837 color images associated with emotional labels, including dimensional labels, categorical labels, and likeability ratings. Subjects’ demographics were also collected such as age, gender, ethnic groups, nationality, educational background, and income level. In addition, we collected semantic tags and other metadata associated with the images in the EmoSet.

5.2.1 The Data Collection Approach

To collect image stimuli, we took 558 emotional words summarized by Averill [97] and used those words to retrieve images by triggering the Flickr image search engine (Examples are presented in Figure 5.1). For each emotional word, we took the top 100 returned images to ensure a high correlation between images and the query. The crawled images were generated by Web users and contained complex scenes that humans may encounter in daily life. We removed duplicate images, images of bad taste, and the ones fully occupied by text.
5.2.2 The Human Subject Study

In our efforts to establish a large-scale set of image stimuli, we leveraged crowd-sourcing and computational tools, and incorporated Lang and Bradley’s methods in creating and validating the IAPS [41] into our study design. We detail the design rationale in this section.

Inspired by the concept of semantic meaning, defined in Charles E. Osgood’s Measurement of Meaning as “the relation of signs to their significants” [99], we asked human subjects to evaluate a series of color images from three perspectives: (I) by rating them along the three dimensional scales, (II) by selecting one or more categorical emotions if relevant, and (III) by selecting their amount of like/dislike toward every presented image.

In part I, we adopted a dimensional approach in an attempt to understand the emotional characteristics that people associate with a vast array of images. The dimensional approach was also used in the creation of the IAPS [41], whose strengths have been indicated by recent studies in psychology [49,50,52,53]. In line with the IAPS study, we utilized the Self-Assessment Manikin (SAM) instrument, recording a rating for the three dimensions: valence, arousal, and dominance. A 9-point rating scale was employed to quantify the emotional ratings on the three dimensions. Instead of the static SAM instrument used in the IAPS, we implemented a dynamic SAM instrument, which could easily be manipulated by sliding a solid bubble along a bar. This was motivated by Peter J. Lang’s claim that “SAM was presented to subjects as an oscilloscope display under the control of a joy-stick on his chair arm” and “An initial instruction program associates the visual anchors of each display with the polar adjectives used by Mehrabian to define the semantic differential scales for the three affective factors” [51]. A gradually changed expression on the dynamic SAM allowed for a more “natural” rating experience for human subjects. As a result of making SAM dynamic, it was necessary to display a single SAM figure for each dimension, minimizing the clutter that would otherwise exist with three rows of static SAM figures, varying slightly in expression.

We collected categorical emotion labels in part II, where eight basic emotions discussed in [96] were included. We displayed the emotions with a “checkbox” next to the words “Click here if you felt any emotions from the displayed image”. Human

---

4Dr. Peter J. Lang is the Director at the Center for the Study of Emotion and Attention and the inventor of the SAM instrument.
subjects were allowed to enter one or more emotions that were not included in the list provided by selecting the checkbox next to the word “Other”, whereby a blank text box would appear.

Also, we collected likeability ratings in part III. We included likeability as an affective measure of images to indicate the extent to which subject liked images. To quantify the likeability, we included a scale for human subjects to select: like extremely, like very much, like slightly, neither like nor dislike, dislike slightly, dislike very much, and dislike extremely.

Motivated by the subjective nature of analyzing image affect, we also collected demographics of the human subjects, including their age, gender, ethnic groups, nationality, educational background, and income level. Such information helps understand the generality of our findings regarding the population.

5.2.3 Human Subject Study Procedures

Detailed procedures of the human subject study are introduced in this section. Once a subject clicked the “agree” button on the consent form, we presented him/her with the instructions for participating in the study. We allowed 5 – 10 minutes for participants to review the instruction, and each subject was asked to evaluate 200 images in total. We briefly summarize the human subject study procedures below:

Step 1 Click the “Start” button, after 5 seconds subjects will be presented with an image;

Step 2 View the image that displays for 6 seconds;

Step 3 A page with three sections will display, and we allow 13 – 15 seconds to fill out the 3 parts. For part I, subjects were asked to make a rating on each scale (valence, arousal, and dominance), based on how they actually felt while they observed the image. They were asked to complete part II only if they felt emotions by selecting one or more of the emotions they felt and/or by entering the emotion(s) they felt into “Other.” For part III, subjects were asked to rate how much they liked or disliked the image. They then clicked “Next” in the lower right hand corner when finished with all three parts;

Step 4 They repeated “Step 2” and “Step 3” until a button with the word “Finish” was displayed;

Step 5 They clicked the “Finish” button.
In *Step 1*, we followed psychological convention in [41], and set 6 seconds as the default value in *Step 2* for image viewing. This was because our intention was to collect immediate affective responses from participants given the visual stimuli. If subjects needed to refer back to the image, they were allowed to click “Reshow Image” in the upper left part of the screen, and click “Hide” to return to the three parts.

![Figure 5.2: Mean value distributions of valence, arousal, dominance, and likeability in the EmoSet.](image)

**5.2.4 Dataset Statistics**

We statistically analyzed the EmoSet, including the collected emotional labels and subjects’ demographics. Each image in the EmoSet was evaluated by at least three subjects. To reduce low-quality ratings, we removed ratings with viewing duration shorter than 2.5 seconds. Borrowing from the procedures used to norm the IAPS,
valence, arousal, and dominance were rated on a scale from 1 to 9 and the likeability was rated on a scale from 1 to 7, borrowing from the widely used photo.net Website. We showed the distributions of mean values in valence, arousal, dominance, and likeability in Figure 5.2 (a)-(d).

The human subject study involved both psychology students within Penn State University and users of the Amazon Mechanical Turk, which ensured a diverse population of emotional ratings. Among the 4148 human subjects we recruited, there were 2236 females and 1912 males, with ages ranging from 18 to 72, various ethnic groups including American Indian or Alaska Native, Asian, African American, native Hawaiian or Other Pacific Islander, Hispanic or Latino, and Not Hispanic or Latino. Participants also had various income and education levels.

5.3 Constructs of Roundness, Angularity, and Simplicity

To articulate a specific relationship between roundness, angularity, and simplicity with human emotion, this work proposes computational methods to map images to the scales of roundness, angularity, and simplicity as three new computational constructs. We detailed the three constructs in the following sections. In this section, we present the computational approaches to compute roundness, angularity, and simplicity in one number. Example results were shown in Figures 5.6, 5.7, and 5.8. Results showed that the three constructs were completely interpretable and could be used in other applications involving roundness, angularity, and simplicity of visual scenes.

5.3.1 Roundness

Roundness was defined as “a measure of how closely the shape of an object approached that of a circle.” [121]. To compute the roundness score of an image, we first segmented the image into regions, then traced their boundaries, and finally computed the goodness of fit to a circle for each region. The step-by-step procedure was:

1. The segmentation approach in [93] was adopted\(^5\). Suppose the segments are

\(^5\)Image segmentation is still a challenging and active research topic in computer vision. Given
Figure 5.3: Examples of images and their scores of roundness. The images with highest roundness scores are shown in the first row; images with medium ranges of roundness scores in the second row; and images with lowest roundness scores in the third row.

\[ S = \{S_1, S_2, \ldots, S_N\}, \] where the number of segments \( N \) was automatically determined by the algorithm. Let the set of boundary points of segment \( S_i \) be \( B_i = \{(x_j, y_j)\} \).

2. The Pratt Algorithm [94] was applied to find the circle \( C_i \) best fitted to \( B_i \). Denote the center of the circle by \((c_i, d_i)\) and radius by \( u_i \). The Pratt Algorithm was applied because of its capacity to fit incomplete circles, i.e., arcs of any degree.

3. For each segment, we defined the roundness disparity of \( S_i \) by \( r_i = \sigma(d(B_i, C_i)) \). Denoted by \( d(B_i, C_i) \) a set of distance between each point in \( B_i \) to \( C_i \), and denoted by \( \sigma \) the standard deviation of that set. The distance between a point \((x_i, y_i)\) and a circle \( C_i \) was computed by the absolute difference between the radius \( u_i \) and the Euclidean distance from the point to the center of the
tens of thousands of images, we assumed it was possible to find a sufficient number of images with obvious characteristics of roundness to analyze its relationship with perceived emotions.
4. The roundness disparity of an image $I$ was denoted by $r_I = \min_{i=1}^{N} r_i e^{-\lambda r_i / \max(v, h)}$. Denoted by $v$ the number of rows and $h$ the number of columns of the image $I$.

![Figure 5.3: Examples of images and their roundness scores. The images with highest roundness scores are shown in the first row; images with medium ranges of roundness scores in the second row; and images with lowest roundness scores in the third row.](image)

In the experiments, we set $\lambda = 0.5$ and normalized the roundness disparity values to $[0, 1]$. We quantified the roundness score as $1 - r_I$, so the closer $r_I$ was to 1 meant that the image was associated with an obvious round property and to 0 the opposite. We present examples of images and their roundness scores in Figure 5.3. The images with highest roundness scores are shown in the first row; images with medium ranges of roundness scores in the second row; and images with lowest roundness scores in the third row.

![Figure 5.4: Examples of images and their scores of angularity. The images with highest angularity scores are shown in the first row; images with medium ranges of roundness scores in the second row; and images with lowest roundness scores in the third row.](image)
5.3.2 Angularity

In the Merriam-Webster Dictionary, angularity is defined as “the quality of being angular”\(^6\), and angular is explained as being lean and having prominent bone structure. We also interviewed five subjects, including one undergraduate student, three graduate students, and one faculty member. The faculty member remarked that angular images in his mind referred to “sword-like” images. The college student said that tall buildings/architecture with angular shapes reflected his perception of angularity. The three graduate students gave examples such as streets, cubes, and tall and lean buildings. These clues motivated us to examine how similar object boundaries were to long ellipses. Similar to roundness, an image was segmented into regions, for each of which an angularity measure was computed. We approximated the quality of being lean and having prominent bone structure by the elongatedness of fitted ellipses. Specifically, the angularity score of an image was computed as following:

1. For each set \(B_i\), least-squares criterion was used to estimate the best fit to an ellipse \(E_i\). We denoted the center of the ellipse by \((c_i, d_i)\), semimajor axis by \(m_i\), semiminor axis by \(n_i\), and angle of the ellipse by \(e_i\).

2. For each image segment \(S_i\), denoted the angularity of a region \(i\) by \(a_i = m_i/n_i\). As our goal was to find lean ellipses, we omitted horizontal and vertical ellipses according to \(e_i\) and ellipses that were too small.

3. We computed angularity of the image \(I\), denoted by \(a_I = \max_{i}^N a_i\).

Angularity scores for images in the EmoSet were computed and normalized to \([0, 1]\). The closer \(a_I\) was to 1 meant that the image showed an obvious angular property. Examples of images and their angularity scores are presented in Figure 5.4.

5.3.3 Simplicity

According to [11], simplicity of an image is primarily depending on two objective factors: Minimalistic structures that are used in a given representation and the

---

\(^7\)http://www.mathworks.com/matlabcentral/fileexchange/3215-fit-ellipse.
simplest way of organizing these structures. Motivated by such concept, we used the number of segments in an image as an indication for its simplicity. We defined the simplicity score by $s_i = |S|$ and normalized the scores to [0, 1] for images in the EmoSet. We present examples of images and their scores of simplicity in Figure 5.5.

5.4 Findings

Analyses presented in this section were performed on the EmoSet utilizing the three constructs of roundness, angularity, and simplicity discussed in Section 5.3 with the purpose of (1) articulating the correlation between roundness, angularity, and simplicity and human emotion in natural setting photos and (2) articulating the capacity of single dimensional representations in emotion classification tasks, such as classifying the positivity and negativity of images and distinguishing images that evoke calmness from excitement.

Whereas psychological conventions treated roundness and angularity as opposite properties, some natural photographs showed neither of the properties. As the goal
of the study was to examine the capabilities of roundness, angularity, and simplicity in evoking human emotion, we targeted visual stimuli with at least a non-zero construct of roundness or angularity. We thus removed 12,158 images from the EmoSet because they were associated with zero constructs of both roundness and angularity, which resulted in 31,679 retained images.

5.4.1 Statistical Correlations

To examine the intrinsic relationship between the three constructs and evoked emotion, we computed correlations between one construct, such as simplicity, roundness, and angularity, and a dimension of the emotional response, such as valence, arousal, and dominance, and found all the correlations were statistically significant in terms of p-value, except the correlation between roundness and likeability. Results are shown in Figures 5.6, 5.7, and 5.8. The red number at the top left corner indicates the statistically significant correlations in terms of p-value.

The strongest correlation coefficient we found was between simplicity and valence, which is 0.11. The correlation coefficients between simplicity and arousal, dominance, and likeability were 0.09, 0.04, and 0.07, respectively. Similarly, for angularity, its correlation coefficients with valence, arousal, dominance, and likeability were 0.05, 0.03, 0.02, and 0.05, and for roundness, results were 0.01, 0.02, 0.02, and −0.01. Correlation coefficients between angularity, roundness and perceived emotion were smaller than simplicity, which implied that simplicity related more strongly with perceived emotion compared to angularity and roundness on an arbitrary photograph.

As shown in these results, the correlation coefficients are small numerically from a psychological perspective. We derive a p-value that was much smaller than 0.0001 is because the correlation was computed on a large collection of images (31,679). A small p-value then only showed that the computed correlation is trustable. Given the correlation coefficients are small, our conclusion is that for natural setting photos, roundness, angularity, and simplicity have rather weak relationship with emotion although in controlled psychological studies, they were found related to emotional response.

\[ \alpha = 0.05 \] for 95% confidence intervals.
5.4.2 Emotion Classification

To examine the capacity of the three constructs to classify the emotional responses to natural image stimuli, especially whether those single dimensional representations are equally as good as high-dimensional representations in emotion classification, we formulated two classification tasks, one is to distinguish positive emotions from negative ones, i.e., high and low valence, and the other is to distinguish calmness from excitement. We compared the classification performance of the proposed three constructs described in Section 5.3 with high-dimensional features, including color, texture, facial, and composition features presented in [6] and shape features introduced in Section 4.2. We detail the experimental settings and classification results as follows.

**Experimental Settings: Positivity of Emotions** The scale of valence is ranged from 1 to 9, where 1 referred to the lowest value in valence and 9 the highest. Images with a medium-range score, such as 5, showed neither positive emotions nor negative emotions. Following conventions in computer science that a gap may apply to facilitate classifier training, we adopted a gap of 1.87 to divide image collections into two groups, images arousing positive emotions (valence > 6.63) and
negative emotions (valence < 4.5). To adjust the classifier parameters and evaluate the trained classifier, we randomly divided the data into training, validation, and testing sets, where the number of images with positive and negative emotions were equal. Specifically, we randomly selected 70% of the data used for training, 10% for validation, and 20% for testing. This resulted in 12600 images in training, 1800 images for validation, and 3600 images for testing. The SVM classifier with RBM kernel was applied, as it is one of the best classifier training approaches in computer science. Among the 144 pairs of parameter candidates, the best performing $c$ and $g$ were selected given their performance on the validation dataset.

**Experimental Settings: Calmness of Emotions** To examine the ability of visual characteristics of complex scenes to evoke different dimensions of emotion, we classified the calmness of emotions, i.e., high and low arousal, following an approach similar with what was used with positivity of emotions. First, a gap was adopted of 2.7 to divide image collections into two groups, high arousal (arousal > 6.4) and low arousal (arousal < 3.7). Then, training, testing, and validation sets were generated, where 7000 images were used in training, 1000 images for validation, and 2000 images for testing. Finally, the SVM classifier was trained and the best set of parameters was selected according to their performance on the validation set.
Table 5.2: Classification Results of High Arousal vs. Low Arousal

<table>
<thead>
<tr>
<th>Features</th>
<th>Dimension</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>roundness</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>angularity</td>
<td>1</td>
<td>52.5</td>
</tr>
<tr>
<td>simplicity</td>
<td>1</td>
<td>56.15</td>
</tr>
<tr>
<td>3-constructs</td>
<td>3</td>
<td>56.15</td>
</tr>
<tr>
<td>color</td>
<td>70</td>
<td>73</td>
</tr>
<tr>
<td>color+3-con</td>
<td></td>
<td>58.7</td>
</tr>
<tr>
<td>C.T.C.F.+shape</td>
<td>332</td>
<td>59.75</td>
</tr>
<tr>
<td>C.T.C.F.+3-con</td>
<td>116</td>
<td>58.65</td>
</tr>
</tbody>
</table>

**Classification Results** Classification results are presented in Tables 4 and 5. “3-constructs” refers to the concatenation of the three constructs, and “color+3-con” denotes the concatenation of color features and the three constructs. “C.T.C.F.+3-con” refers to a concatenation of color, texture, composition, facial, and the three constructs. “C.T.C.F.+shape” refers to the concatenation of color, texture, composition, facial, and shape features.

To examine whether the three predictors are equally good as the high-dimensional features, we combined the three constructs with color features and used concatenated features to classify the positivity of emotions. As shown in Table 4, the 73 dimensional “color+3-con” feature improved upon the “color” feature slightly. Similarly, we concatenated the three constructs with color, texture, composition, and facial features. As shown in the Table, we found that the 116 dimensional features “C.T.C.F.+3-con” improved the classification performance achieved by “C.T.C.F.+shape”, the 332 dimensional features. We also noticed that the best classification results were achieved by “C.T.C.F.+3-con”, which clearly showed that single dimensional representations proposed in this work outperformed the high-dimensional shape features discussed in Section 4.2, in distinguishing images that arouse positive or negative emotions.

Similarly, as shown in Table 5, the 73 dimensional “color+3-con” feature improved upon the “color” feature slightly in distinguishing calmness from excitement. The 116 dimensional features “C.T.C.F.+3-con” achieved slightly lower but competitive classification accuracy than “C.T.C.F.+shape”. With results shown in Tables 4 and 5, we found that (1) the three constructs showed consistent capacity in classifying both dimensions of emotions, (2) the three constructs achieved comparable or equally as good as the high-dimensional shape features discussed in Section 4.2 in emotion classification.
5.5 Summary

In this chapter, we investigated roundness, angularity, and simplicity of complex scenes that evoked human emotion on natural setting photos. The three new constructs mapped the visual content to the scales of roundness, angularity, and simplicity, which was shown as completely interpretable and could be used in other applications involving roundness, angularity, and simplicity of visual scenes. We built a large collection of ecologically valid stimuli, i.e., EmoSet, to examine the correlation between roundness, angularity, and simplicity, and human emotion on natural setting photos. Through correlation analyses, we found that roundness, angularity, and simplicity have rather weak relationship with emotion. This finding was opposed to the one found in controlled psychological studies. In addition, we evaluate whether the three single dimensional representations are equally as good as high-dimensional representations in emotion classification by formulating two classification tasks on the EmoSet. Classification results demonstrated the capacity of the three constructs in classifying both dimensions of emotion. Interestingly, when combining with color features, the three constructs outperformed classification accuracy on distinguishing the positivity to a set of 200 texture, composition, facial, and shape features.

As future work, the proposed approach could be easily applied to examine other visual characteristics that evoke human emotion in complex scenes. We expect that our efforts may contribute to research regarding visual characteristics of complex scenes and human emotion from perspectives of visual arts, psychology, and computer science.
RAPID: RAting PIctorial Aesthetics using Deep Learning

Figure 6.1: Global views and local views of an image. Global views are represented by normalized inputs: center-crop, warp, and padding (shown in the top row). Local views are represented by randomly-cropped inputs from the original high-resolution image (examples shown).
6.1 Introduction

Patterns in aesthetically-pleasing photographs often indicate photographers’ visual preferences. Among those patterns, composition [95] and visual balance [44] are important factors [11]. They are reflected in the global view (e.g., top row in Figure 6.1) and the local view (e.g., bottom row in Figure 6.1). Popular composition principles include the rule of thirds, diagonal lines, and golden ratio [1], while visual balance is affected by position, form, size, tone, color, brightness, contrast, and proximity to the fulcrum [44]. Some of these patterns are not well-defined or even abstract, making it difficult to calculate those features for assessing image aesthetic quality. Given a large collection of training images, the holistic and fine-grained descriptors could be learned through deep learning approaches to rate or to categorize the aesthetic quality of images.

This chapter presents novel deep neural network approaches to automatically predict image aesthetics by leveraging fine-grained high resolution details and holistic information from images. We intend to automatically learn effective aesthetics features from images directly. We are motivated by the feature learning power of deep convolutional neural networks [30], where feature learning is unified with classifier training using RGB images, and we propose to learn effective aesthetics features using convolutional neural networks. However, applying classic architecture to our task is not straightforward. Image aesthetics depends on a combination of local cues (e.g., sharpness and noise levels) and global visual cues (e.g., the rule of thirds), as shown in Figure 6.1. To learn aesthetics-relevant representations of an image, we generated two heterogeneous inputs to represent its global cues and local cues respectively, as shown in Figure 6.1. Meanwhile, we developed a double-column neural network architecture that takes parallel inputs from the two columns to support network training on heterogeneous inputs, which extends the method in [30]. In the proposed double-column neural network, one column takes a global view of the image and the other column takes a local view of the image. The two columns are aggregated after some layers of transformations and mapped to the label layer. In addition, we developed regularized double-column convolutional neural network to incorporate additional information, such as style or semantics, to improve the accuracy of image aesthetics assessment.
6.2 Deep Convolutional Neural Networks

6.2.1 Convolutional Neural Network

Deep convolutional neural network [30] takes inputs of fixed aspect ratio and size. However, an input image can be of arbitrary size and aspect ratio. To normalize image sizes, we propose three different transformations: center-crop \( g_c \), warp \( g_w \), and padding \( g_p \), which reflect the global view \( I_g \) of an image \( I \). \( g_c \) isotropically resizes original images by normalizing their shorter sides to a fixed length \( s \). Center-crop normalizes the input to generate a \( s \times s \times 3 \) input. \( g_c \) was adopted in a recent image classification work [30]. \( g_w \) anisotropically resizes (or warps) the original image into a normalized input with a fixed size \( s \times s \times 3 \). \( g_p \) resizes the original image by normalizing the longer side of the image to a fixed length \( s \) and padding border pixels with zeros to generate a normalized input of a fixed size \( s \times s \times 3 \). For each image \( I \) and each type of transformation, we generate an \( s \times s \times 3 \) input \( I_g^j \) with the transformation \( g_j \), where \( j \in \{ c, w, p \} \). As resizing inputs can cause harmful information loss (i.e., the high-resolution local views) for aesthetic assessment, we also use randomly sampled fixed size (at \( s \times s \times 3 \)) crops with the transformation \( l_r \). Here we use \( g \) to denote global transformations and \( l \) to denote local transformations. This results in normalized inputs \( \{ I_r^g \} \) \( (r \) is an index of normalized inputs for each random cropping\), which preserve the local views of an image with details from the original high-resolution image. We used these normalized inputs \( I_t \in \{ I_c^g, I_w^g, I_p^g, I_r^g \} \) for CNN training. In this work, we set \( s \) to 256, thus the size of \( I_t \) is \( 256 \times 256 \times 3 \). To alleviate overfitting in network training, for each normalized input \( I_t \), we extracted a random \( 224 \times 224 \times 3 \) patch \( I_p \) or its horizontal reflection to be the input patch of our network.

We present an example for the four transformations, \( g_w, g_c, g_p, \) and \( l_r \), in Figure 6.1. As shown, the global view of an image is maintained via the transformations of \( g_c, g_w, \) and \( g_p \). Among the three global views, \( I_w^g \) and \( I_p^g \) maintain the relative spatial layout among elements in the original image. \( I_w^g \) and \( I_p^g \) follow rule of thirds whereas the \( I_c^g \) does not. In the bottom row of the figure, the local views of an original image are represented by randomly-cropped patches \( \{ I_r^g \} \). These patches depict the local details in the original resolution of the image.

The architecture of the SCNN used for aesthetic quality assessment is shown
Figure 6.2: Single-column convolutional neural network for aesthetic quality rating and categorization. We have four convolutional layers and two fully-connected layers. The first and second convolutional layers are followed by max-pooling layers and normalization layers. The input patch of the size $224 \times 224 \times 3$ is randomly cropped from the normalized input of the size $256 \times 256 \times 3$ as done in [30].

in Figure 6.2. It has a total of four convolutional layers. The first and the second convolutional layers are followed by max-pooling layers and normalization layers. The first convolutional layer filters the $224 \times 224 \times 3$ patch with 64 kernels of the size $11 \times 11 \times 3$ with a stride of 2 pixels. The second convolutional layer filters the output of the first convolutional layer with 64 kernels of the size $5 \times 5 \times 64$. Each of the third and forth convolutional layers has 64 kernels of the size $3 \times 3 \times 64$, and the two fully-connected layers have 1000 and 256 neurons respectively.

Suppose for the input patch $I_p$ of the $i$-th image, we have the feature representation $x_i$ extracted from layer fc256 (the outcome of the convolutional layers and the fc1000 layers), and the label $y_i \in \mathcal{C}$. The training of the last layer is done by maximizing the following log likelihood function:

$$l(W) = \sum_{i=1}^{N} \sum_{c \in \mathcal{C}} \mathbb{I}(y_i = c) \log p(y_i = c \mid x_i, w_c),$$  \hspace{1cm} (6.1)$$

where $N$ is the number of images, $W = \{w_c\}_{c \in \mathcal{C}}$ is the set of model parameters, and $\mathbb{I}(x) = 1$ iff $x$ is true and vice versa. The probability $p(y_i = c \mid x_i, w_c)$ is expressed as

$$p(y_i = c \mid x_i, w_c) = \frac{\exp (w_c^T x_i)}{\sum_{c' \in \mathcal{C}} \exp (w_{c'}^T x_i)}. \hspace{1cm} (6.2)$$

The aesthetic quality categorization task can be defined as a binary classification
problem where each input patch is associated with an aesthetic label \( c \in \mathcal{C} = \{0, 1\} \). In Section 6.2.3, we explain a SCNN for image style categorization, which can be considered a multi-class classification task.

As indicated by the previous study [30], the architecture of the deep neural network may critically affect the performance. Our experiments suggest that the general guideline for training a good-performing network is to first allow sufficient learning power of the network by using sufficient number of neurons. Meanwhile, we adjust the number of convolutional layers and the fully-connected layers to support the feature learning and classifier training. In particular, we extensively evaluate the network trained with different numbers of convolutional layers and fully-connected layers, and with or without normalization layers. Candidate architectures are shown in Table 6.1. To determine the optimal architecture for our task, we conduct experiments on candidate architectures and pick the one with the highest performance, as shown in Figure 6.2.

With the selected architecture, we train SCNN with four different types of inputs \((I^c_g, I^w_g, I^p_g, I^r_l)\) using the AVA dataset [23]. During training, we handle the overfitting problem by adopting dropout and shuffling the training data in each epoch. Specifically, we found that \( l_r \) serves as an effective data augmentation approach which alleviates overfitting. Because \( I^r_l \) is generated by random cropping, an image contributes to the network training with different inputs when a different patch is used.

We experimentally evaluate the performance of these inputs with SCNN. Results will be presented in Section 6.3. \( I^w_g \) performs the best among the three global input variations \((I^c_g, I^w_g, I^p_g)\). \( I^r_l \) yields an even better results compared with \( I^w_g \). Hence, we use \( I^r_l \) and \( I^w_g \) as the two inputs to train the proposed double-column network. In our experiments, we fix the dropout rate as 0.5 and initiate the learning rate with 0.001. Given a test image, we compute its normalized input and followed by generating the input patch, with which we calculate the probability of the input patch being assigned to each aesthetic category. We repeat this process for 50 times, average those results, and pick the class with the highest probability.
6.2.2 Double-Column Convolutional Neural Network

For each image, its global or local information may be lost when transformed to a normalized input using $g_c$, $g_w$, $g_p$, or $l_r$. Representing an image through multiple inputs can somewhat alleviate this problem. As a first attempt, we generate one input to depict the global view of an image and another to represent its local view. We propose a novel double-column convolutional neural network (DCNN) to support automatic feature learning with heterogeneous inputs, i.e., a global-view input and a local-view input. We present the architecture of the DCNN in Figure 6.3. As shown in the figure, networks in different columns are independent in convolutional layers and the first two fully-connected layers. The inputs of the two columns are $I_g^w$ and $I_r^l$. We take the two $256 \times 1$ vectors from each of the fc256 layer and jointly train the weights of the final fully-connected layer. We avoid the interaction between two columns in convolutional layers because they are in different spatial scales. During training, the error is back propagated to the networks in each column respectively with stochastic gradient descent. With the proposed architecture, we can also automatically discover both the global and the local features of an image from the fc1000 layers and fc256 layers.

The proposed network architecture could easily be expanded to multi-column convolutional networks by incorporating more types of normalized inputs. DCNN allows different architectures in individual networks, which may facilitate the param-
Figure 6.4: Regularized double-column convolutional neural network (RDCNN). The style attributes $x_s$ are generated through pre-trained Style-SCNN and we leveraged the style attributes to regularize the training process of RDCNN. The dashed line indicates that the parameters of the style column is fixed during RDCNN training. While training the RDCNN, we only fine-tuned the parameters in the aesthetic column and the learning process is supervised by the aesthetic label.

6.2.3 Regularized Double-Column Convolutional Neural Network

The discrete aesthetic labels, i.e., high quality and low quality, provide weak supervision to make the network converge properly due to the large intra-class variation. This motivates us to exploit extra labels from the training images to help identify their aesthetic characteristics. We propose to leverage style attributes, such as complementary colors, macro, motion blur, rule of thirds, shallow depth-of-field (DOF), to help determine the aesthetic quality of images because they are regarded as highly relevant attributes [23].

There are two natural ways to formulate the problem. The first is to leverage the idea of multi-task learning [100], which jointly construct feature representation and minimize the classification error for both labels. Assuming we have aesthetic quality labels $\{y_{ai}\}$ and style labels $\{y_{si}\}$ for all training images, an optimization
problem is formulated as following:
\[
\max_{\mathbf{X}, \mathbf{w}_a, \mathbf{w}_s} \sum_{i=1}^{N} \left( \sum_{c \in C_A} \mathbb{I}(y_{ai} = c) \log p(y_{ai} | \mathbf{x}_i, \mathbf{w}_{ac}) + \sum_{c \in C_S} \mathbb{I}(y_{si} = c) \log p(y_{si} | \mathbf{x}_i, \mathbf{w}_{sc}) \right) ,
\]
(6.3)
where \( \mathbf{X} \) is the features of all training images, \( C_A \) is the label set for aesthetic quality, \( C_S \) is the label set for style, and \( \mathbf{W}_a = \{w_{ac} \}_{c \in C_A} \) and \( \mathbf{W}_s = \{w_{sc} \}_{c \in C_S} \) are the model parameters. It is more difficult to obtain images with style attributes. In the AVA benchmark, among 230,000 image with aesthetic labels only 14,000 of them have style labels. As a result, we cannot jointly perform aesthetics categorization and style classification with a single neural network due to many missing labels.

Alternatively, we can use ideas from inductive transfer learning [101], where we target minimizing the classification error with one label, whereas we construct feature representations with both labels. As we only have a subset of images with style labels, we first train a style classifier with them. We then extract style attributes for all training images, and apply those attributes to regularize the feature learning and classifier training for aesthetic quality categorization.

To learn style attributes for 230,000 training images, we first train a style classifier by performing the training procedure discussed in Section 6.2.1 on 11,000 labeled training images (Style-SCNN). We adopted the same architecture as shown in Figure 6.2. The only difference is that we reduced the number of filters in the first and fourth convolutional layers to a half due to the reduced number of training images. With Style-SCNN, we maximize the log likelihood function in Equation 6.1 where \( C \) is the set of style labels in the AVA dataset. We experimentally select the best architectures (to be shown in Table 6.4) and inputs \((I^c_g, I^w_g, I^p_g, I^r_l)\). The details are described in Section 6.3. Given an image, we apply the learned weights and extract the features from the fc256 layer as its style attribute.

To facilitate the network training with style attributes of images, we propose a regularized double-column convolutional neural network (RDCNN) with the architecture shown in Figure 6.4. Two normalized inputs of the aesthetic column are \( I^w_g \) and \( I^r_l \), same as in DCNN (Section 6.2.2). The input of the style column is
The training of RDCNN is done by solving the following optimization problem:

$$\max_{\mathbf{x}_a, \mathbf{W}_a} \sum_{i=1}^{N} \sum_{c \in C_a} \mathbb{I}(y_{ai} = c) \log p(y_{ai} \mid \mathbf{x}_{ai}, \mathbf{x}_{si}, \mathbf{w}_{ac}),$$

(6.4)

where $\mathbf{x}_{si}$ are the style attributes of the $i$-th training image, $\mathbf{x}_{ai}$ are the features to be learned. Note that the maximization does not involve style attributes $\mathbf{x}_s$. In each learning iteration, we only fine-tuned the parameters in the aesthetic column and the learning process is supervised by the aesthetic label. The parameters of the style column are fixed and the style attributes $\mathbf{x}_{is}$ essentially serve as a regularizer for training the aesthetic column.

6.3 Evaluation

We evaluated the proposed methods for image aesthetics on the AVA\textsuperscript{1} dataset and the IAD dataset. On the AVA dataset, we divided training images into two categories, i.e., low-quality and high-quality images, according to criteria presented in [23], and we learned style attributes using the AVA style dataset\textsuperscript{2}.

We first present the performance of SCNN using different network architectures and taking $I_I^*$ as the input. We selected the best performing architecture and evaluate SCNN using different types of inputs. Next, we present image aesthetics prediction results produced by DCNN, and qualitatively analyzed the advantage of the double-column architecture over a single-column one. We evaluated the network adaptation approach for content-based image aesthetics on the eight representative image categories [23] (portrait, animal, stilllife, fooddrink, architecture, floral, cityscape, and landscape), and compared the network adaptation results (including both the SCNN and DCNN) with the state-of-the-art aesthetics accuracy in each of the eight categories. Further, we show the accuracy of trained style classifier

\textsuperscript{1}The AVA dataset includes 250,000 images, each of which is associated with an aesthetics score, averaged by about more than 200 ratings. The scale of the aesthetics score is from 1 to 10. We took the same experimental settings as in [23]. Particularly, we took the same division of training and testing data as in [23], i.e., 230,000 images for training and 20,000 for testing.

\textsuperscript{2}The AVA style dataset includes 11,000 images for training and 2,500 images for testing. Each of the images in the training set is associated with one of the 14 style labels, i.e., complementary colors, duotones, HDR, image grain, light on white, long exposure, macro, motion blur, negative images, rule of thirds, shallow DOF, silhouettes, soft focus, and vanishing point. Each of images in the test dataset is associated with one or multiple style labels as the ground truth.
Table 6.1: Accuracy of Different SCNN Architectures

<table>
<thead>
<tr>
<th></th>
<th>Arch 1</th>
<th>Arch 2</th>
<th>Arch 3</th>
<th>Arch 4</th>
<th>Arch 5</th>
<th>Arch 6</th>
<th>Arch 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1(64)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>pool1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>rnorm1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>conv2(64)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>pool2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>rnorm2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>conv3(64)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>conv4(64)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>conv5(64)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>conv6(64)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fc1K</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fc256</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fc2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Accuracy</td>
<td>71.20%</td>
<td>60.25%</td>
<td>62.68%</td>
<td>65.14%</td>
<td>70.52%</td>
<td>62.49%</td>
<td>70.93%</td>
</tr>
</tbody>
</table>

and aesthetic categorization results generated by RDCNN with style attributes or semantic attributes incorporated. Moreover, we introduce a 1.5 million image dataset (IAD) with aesthetics scores, including images derived from DPChallenge\(^3\) and PHOTO.NET\(^4\). We further boost the aesthetics assessment accuracy, presented in Section 6.3.6, by training SCNN and DCNN on the IAD dataset. Finally, we discuss the computational efficiency of the proposed approaches on the AVA dataset and the IAD dataset.

### 6.3.1 SCNN Results

We first examine the performance of SCNN on different network architectures and present overall accuracy of image aesthetics using seven different architectures listed in Table 6.1. The selected layer for each architecture is labeled with a check mark. To fairly compare the performance of network architectures, we took the same normalized input, \(I^r\), as training examples, and we set \(\delta = 0\). As shown in the Table, the highest accuracy was achieved by the Arch 1. We thus fixed the network architecture to Arch 1 in the following experiments.

We evaluate the performance of SCNN with various normalized inputs as training inputs, and we set \(\delta = 0\). As shown in the Table, the highest accuracy was achieved by the Arch 1. We thus fixed the network architecture to Arch 1 in the following experiments.

\[^4\]http://photo.net.
Table 6.2: Accuracy of Aesthetic Quality Categorization with Different Inputs

<table>
<thead>
<tr>
<th>δ</th>
<th>$I^c_I$</th>
<th>$I^w_g$</th>
<th>$I^c_g$</th>
<th>$I^p_g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>71.20%</td>
<td>67.79%</td>
<td>65.48%</td>
<td>60.43%</td>
</tr>
<tr>
<td>1</td>
<td>68.63%</td>
<td>68.11%</td>
<td>69.67%</td>
<td>70.50%</td>
</tr>
</tbody>
</table>

Table 6.3: Accuracy of Aesthetic Quality Categorization with Different Methods

<table>
<thead>
<tr>
<th>δ</th>
<th>[23]</th>
<th>SCNN</th>
<th>AVG_SCNN</th>
<th>DCNN</th>
<th>RDCNN_style</th>
<th>RDCNN_semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>66.7%</td>
<td>71.20%</td>
<td>69.91%</td>
<td>73.25%</td>
<td>74.46%</td>
<td>75.42%</td>
</tr>
<tr>
<td>1</td>
<td>67%</td>
<td>68.63%</td>
<td>71.26%</td>
<td>73.05%</td>
<td>73.70%</td>
<td>74.2%</td>
</tr>
</tbody>
</table>

examples, i.e., $I^c_g$, $I^w_g$, $I^p_g$, and $I^c_I$. We trained deep networks with both $\delta = 0$ and $\delta = 1$ for each input type, and presented the overall accuracy in Table 6.2. We observed from the Table that the best performance was achieved by $I^c_I$, which indicates that $l_r$ is an effective data augmentation strategy to capture the fine-grained details pertinent to image aesthetics. We also noticed that $I^w_g$ produces the highest accuracy among the three inputs for capturing the global view of images. We present the best performance of SCNN using Arch 1 and $I^c_I$ as the training input in Table 6.3. As shown, our performance is better than the previous study [23] for both $\delta = 0$ and $\delta = 1$.

6.3.2 DCNN Results

As we have shown that Arch 1 performs the best among all attempted architectures in Section 6.3.1. In DCNN training and testing, we adopted the SCNN architecture Arch 1 for both columns. We took three inputs combinations to train the proposed double-column network, i.e., $I^c_I$ and $I^c_g$, $I^c_I$ and $I^p_g$, and $I^c_I$ and $I^w_g$. Let $\delta = 0$, we empirically evaluated results achieved by the three variations. The combination of $I^c_I$ and $I^c_g$ achieves 71.8% accuracy, and $I^c_I$ and $I^p_g$ achieves 72.27% accuracy. The combination of $I^c_I$ and $I^w_g$ performs the best among the three, achieving 73.25% accuracy. Thus, we used the two inputs of $I^c_I$ and $I^w_g$ in DCNN training and evaluation.

In Figure 6.5, we visualize the filters of the first convolutional layer in the trained DCNN. The first 64 filters are from the local column (using the input $I^c_g$), and the last 64 filters are from the global column (using the input $I^w_g$). For comparison,
Figure 6.5: Filter visualization of DCNN for image aesthetics. In particular, 128 convolutional kernels of the size $11 \times 11 \times 3$ learned by the first convolutional layer. The first 64 are from the local view column (with the input $I_{rl}$) and the last 64 are from the global view column (with the input $I_{wg}$).

Figure 6.6: 64 convolutional kernels of the size $5 \times 5 \times 3$ learned by the first convolutional layer of CNN for object classification on the CIFAR dataset.

we showed filters trained in the object recognition task on CIFAR dataset\(^5\) in Figure 6.6. Interestingly, we found that the filters learned with image aesthetic labels are free from radical intensity changes and look smoother and cleaner. Such observation indicates that differences between low-aesthetic and high-aesthetic cues primarily lie in the amount of texture and complexity of the entire image. Such intuitions could also be observed from example images, presented in Figure 6.7. In general, the images ranked the highest in aesthetics are smoother than those ranked the lowest. This finding substantiates the significance of simplicity and complexity features recently developed for analyzing perceived emotions [16].

To further show the power of DCNN, we quantitatively compared its performance with that of the SCNN and [23]. We show in Table 6.3 that DCNN outperforms SCNN for both $\delta = 0$ and $\delta = 1$, and significantly outperforms the earlier study. We further demonstrate the effectiveness of joint training strategy adopted in DCNN by comparing DCNN with AVG_SCNN, which averaged the two SCNN results taking $I^w_\delta$ and $I^r_\delta$ as inputs. We present the comparisons in Table 6.3, and the results show that DCNN performs better than the AVG_SCNN with both $\delta = 0$ and $\delta = 1$.

To analyze the advantage of the double-column architecture, we visualize test images correctly classified by DCNN and misclassified by SCNN. Examples are

Figure 6.7: Images ranked the highest and the lowest in aesthetics generated by DCNN. Differences between low-aesthetic images and high-aesthetic images heavily lie in the amount of textures and complexity of the whole image.

presented in Figure 6.8, where images in the first row are misclassified by SCNN with the input $I_l$, and images in the second row are misclassified with the input $I_w$. The label annotated on each image indicates the ground-truth aesthetic quality.
We found that images misclassified by SCNN with the input $I_r$ mostly dominated by an object, which is because the input $I_r$ fails to consider the global information in an image. Similarly, images misclassified by SCNN with the input $I_g$ usually contain fine-grained details in their local views. The result implies that both global view and fine-grained details help improve the aesthetics prediction accuracy as long as the information is properly leveraged.

As discussed in Section 6.2.2, a natural extension of DCNN is to use multiple columns in CNN training, such as a quad-column CNN. Let $\delta = 0$, we attempted to use the four inputs $I_r$, $I_c$, $I_p$, and $I_g$ to train a quad-column CNN. Compared with the DCNN, training a quad-column network architecture requires GPU with a larger memory. We adopted the SCNN architecture, the Arch 1, for all the four columns, and it turns out that a GPU with 5G memory (such as Nvidia Tesla M2070/M2090 GPU) is no longer applicable for network training with a mini-batch size of 128. Meanwhile, optimizing parameters in quad-column CNN is more difficult than a double-column CNN because an individual column may require a different learning rate and training duration. In practice, we initialized each of the four columns with SCNN trained using $I_r$, $I_c$, $I_p$, and $I_g$, respectively. We then fine-tuned the last layer of the quad-column neural network. Compared with DCNN, a quad-column CNN achieved a slightly higher accuracy of 73.38%.
6.3.3 Content-based Image Aesthetics

To demonstrate the effectiveness of network adaptation for content-based image aesthetics, we took the eight most popular semantic tags as used in [23]. We used the same training and testing image collection with [23], roughly $2.5K$ for training and $2.5K$ for testing in each of the categories.

Few images in the AVA dataset have been removed from the Website, so we might have slightly smaller number of test images compared with [23]. Specifically, the number of test images in each of the eight categories is: portrait(2488), animal(2484), stilllife(2491), fooddrink(2493), architecture(2495), floral(2495), cityscape(2494), and landscape(2490).
In each of the eight categories, we systematically compared the proposed network adaptation approach (denoted by “adapt”) built upon the SCNN (with the input $I^r$ and $I^w$) and the DCNN with two baseline approaches (“cats” and “generic”) and a state-of-the-art approach [23]. “cats” refers to the approach that trains the network using merely the categorized images (i.e., roughly 2.5K in each category), and “generic” refers to the approach that trains the network using the AVA training set (i.e., including images of arbitrary semantic categories). As presented in the Figure 6.9 (a), (b), and (c), the proposed network training approach significantly outperforms the state of the art [23] (except the category of stilllife). In particular, the “generic” produces higher accuracy in general than the “cats” for SCNN with the input $I^w$ and the DCNN, and “general” performs similar with “cats” for SCNN with the input $I^r$. This indicates the effectiveness of $g_r$. For the SCNN with both the inputs and the DCNN, “adapt” shows better performance in most of the categories than “cats” and “generic”.

In our experiments, the adaptation of the SCNN with the input $I^r$ takes about 50 epochs and the adaptation of SCNN with the input $I^w$ takes about 10 epochs. The fact that SCNN adaptation with the input $I^w$ takes less epochs is because the random selection of patches $l^r$ results in more training examples than the warping operation $g^w$. As shown in the Figure 6.9, the SCNN with the input $I^w$ performs better than the SCNN with the input $I^r$. This indicates once an image is associated with an obvious semantic meaning, then the global view is more important than the local view in terms of assessing image aesthetics. Moreover, for the “generic” and the “adapt”, the DCNN outperforms the SCNN with the input $I^w$ and $I^r$, which indicates that both the global view and the local view contribute to the aesthetic quality categorization of content-specific images. The DCNN does not show improvement in the “cats” due to the limited number of training examples. In addition, by combination with image classification [30], this content based method can produce overall aesthetics predictions given uncategorized images.

We also observed that the SCNN with the input $I^w$ performs better than the SCNN with the input $I^r$, as shown in Figure 6.9. This result indicates that once an image is associated with an obvious semantic meaning, then the global view is more important than the local view in terms of assessing image aesthetics. Moreover, for both the “generic” and the “adapt”, the DCNN outperforms the SCNN with

---

7 We refer to the best performance of content-based image aesthetics in [23].
the input $I_g^w$ and $I_l^r$, which indicates that both the global view and the local view contribute to the aesthetic quality categorization of content-specific images. The DCNN does not show improvement in the “cats” due to the limited number of training examples. In combination with image classification [30], this content-based method can produce overall aesthetics predictions given uncategorized images.

We investigated the reason why the performance of SCNN is worse than [23] in the “stilllife” category, while in the other 7 categories, the SCNN performs better than [23]. We observed that unlike images in the other 7 categories, images in the stilllife category do not share obvious visual similarity. As shown in Figure 6.10 (left), images in this category involve objects and scenes of various semantic categories that we may encounter in our everyday life, such as animals, fruits, houses, and people. Moreover, we found that images in the stilllife category tend to be associated with certain artistic styles, as shown in Figure 6.10 (right). This fact makes the problem of assessing the aesthetics of stilllife images more challenging than the other 7 categories using the deep network training approach. Due to the diverse content of images in this category and the limited number of training data (2491), compared to [23], the performance of SCNN in this category is worse in general. An exception, according to Figure 6.9 (b), is the results produced by “adapt” with the inputs of $I_g^w$, which slightly outperforms [23].
Table 6.4: Accuracy of Different Network Architectures for Style Classification

<table>
<thead>
<tr>
<th></th>
<th>Arch 1</th>
<th>Arch 2</th>
<th>Arch 3</th>
<th>Arch 4</th>
<th>Arch 5</th>
<th>Arch 6</th>
<th>Arch 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1(32)</td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>pool1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>rnorm1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>conv2(64)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>pool2</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>rnorm2</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>conv3(64)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>conv4(32)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>conv5(32)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>conv6(32)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>fc1K</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>fc256</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>fc2</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>MAP</td>
<td>56.81%</td>
<td>52.39%</td>
<td>53.19%</td>
<td>54.13%</td>
<td>53.94%</td>
<td>53.22%</td>
<td>47.44%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>59.89%</td>
<td>54.33%</td>
<td>55.19%</td>
<td>55.77%</td>
<td>56.0%</td>
<td>57.25%</td>
<td>52.16%</td>
</tr>
</tbody>
</table>

Table 6.5: Accuracy of Style Classification with Different Inputs

<table>
<thead>
<tr>
<th></th>
<th>$I_f$</th>
<th>$I_w$</th>
<th>$I_q$</th>
<th>$I_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>56.93%</td>
<td>44.52%</td>
<td>45.74%</td>
<td>41.78%</td>
</tr>
<tr>
<td>MAP</td>
<td>56.81%</td>
<td>47.01%</td>
<td>48.14%</td>
<td>44.07%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>59.89%</td>
<td>48.08%</td>
<td>48.85%</td>
<td>46.79%</td>
</tr>
</tbody>
</table>

### 6.3.4 Categorization with Style Attributes

We evaluate the performance of RDCNN in two steps to show the effectiveness of using style attributes in helping image aesthetics prediction. We first evaluate the style classifier, and then we evaluate the aesthetics prediction accuracy achieved by RDCNN.

The style classification performance achieved by SCNN was compared with the performance reported in [23]. The Average Precision (AP) and mean Average Precision (mAP) were used as the evaluation metrics. We trained and evaluated the SCNN on the same collection of training and testing images as presented in [23]. We conducted similar experiments as we have presented in Section 6.3.1 to select a best-performed network architecture. We fixed the architecture and compare performance of SCNN using the four input types. As shown in Table 6.4, the best
mAP we achieved is 56.81% which outperforms the accuracy of 53.85% reported in [23]. The best performance is produced by Arch 1, shown in Table 6.4, using the $I_r$ as the input. Results produced by other inputs are presented in Table 6.5. We visualize the filters learned by the first convolutional layer of SCNN for image style classification in Figure 6.11.

![Figure 6.11](image)

Figure 6.11: 32 convolutional kernels of the size $11 \times 11 \times 3$ learned by the first convolutional layer of Style-SCNN for style classification.

The Average Precision (AP) and Mean Average Precision (MAP) are also calculated. The best MAP we achieved is 56.81% which outperforms the accuracy of 53.85% reported in [23].

To demonstrate the effectiveness of style attributes, the results produced by RDCNN were compared with ones produced by DCNN for both $\delta = 0$ and $\delta = 1$. The results, shown in Table 6.3, reveal that RDCNN outperforms DCNN. We qualitatively analyzed the results produced by RDCNN and found that examples correctly classified by RDCNN$_{\text{style}}$ are mostly associated with obvious stylistic

![Figure 6.12](image)

Figure 6.12: Test images correctly classified by RDCNN$_{\text{style}}$ and misclassified by DCNN. The label on each image indicates the ground truth aesthetic quality of images.
characteristics, such as rule-of-thirds, HDR, black and white, long exposure, complementary colors, vanishing point, and soft focus. Examples are presented in Figure 6.12. The observation implies that style attributes help image aesthetics prediction in cases when images are associated with obvious styles.

6.3.5 Categorization with Semantic Attributes

We evaluate the RDCNN\textsubscript{semantic} approach in the same way with RDCNN\textsubscript{style}. In our experiments, we first fine-tuned the ImageNet to the image aesthetics problem by adding an fc256 layer and replacing the label layer. We then trained regularized RDCNN\textsubscript{semantic} as introduced in Section 6.2.3. The categorization results with semantic attributes are shown in Table 6.3, where RDCNN\textsubscript{semantic} improves the accuracy produced by DCNN. By comparing the best aesthetic quality categorization accuracy with and without semantic attributes, we demonstrate the effectiveness of using semantic attributes in determining the aesthetics of images.

We conducted qualitative analysis to show the advantage of the RDCNN architecture using semantic attributes. We present the examples in Figure 6.13,
Figure 6.14: The Mean Score Distributions of images collected from DPChallenge and PHOTO.NET. Left: The mean score distributions of the 1.2 million PHOTO.NET images; Right: The mean score distributions of the 300K DPChallenge images.

where we show representative test images that have been correctly classified by RDCNN\textsubscript{semantic} but misclassified by DCNN. Our observation is that the classification accuracy improves with images that contain obvious objects.

### 6.3.6 The IAD dataset

We introduce a new large-scale image aesthetics dataset (IAD), containing 1.5 million images, to explore the impact of a larger-scale training dataset to the proposed approach in terms of classification accuracy. Among the images in the IAD dataset, 300K images were derived from the DPChallenge\textsuperscript{8}, and 1.2 million were derived from the PHOTO.NET\textsuperscript{9}. The score distributions of the two sub-collections are presented in Figure 6.14.

To generate a training dataset with two categories (low aesthetics and high aesthetics), we divided the images crawled from the PHOTO.NET based on their mean score 4.88 (i.e., images with score higher than 4.88 are labeled as high aesthetics, and images with score lower than 4.88 are labeled as low aesthetics.) For images crawled from DPChallenge, we followed [23] and labeled the images as

---

\textsuperscript{8}We crawled all the images on the DPChallenge uploaded upon April 2014.

\textsuperscript{9}We included all the images on the PHOTO.NET uploaded upon April 2014 that have been associated with more than 5 aesthetic labels.
high aesthetics when the score is larger than 5 and otherwise labeled the images as low aesthetics. We handled the images crawled from the two sources separately because the score scales in the PHOTO.NET (1 – 7) and the DPChallenge are different (1 – 10). This results in 747K and 696K training images in the categories of high aesthetics and low aesthetics respectively.

We trained the SCNN and DCNN on the IAD dataset using the same architecture as introduced in [30]. We first trained the SCNN with the input $I_r$, and we evaluated the network on the AVA test set and achieved 73.21% accuracy, about 2% higher than the SCNN trained on AVA dataset. We then trained the SCNN with $I_w$ as the input, and the network achieved 73.65% accuracy, 5% percent higher than the SCNN trained on the AVA dataset. We initialized the two columns of DCNN with the SCNN with inputs of the $I_r$ and $I_w$, and the DCNN achieved an accuracy of 74.6%, compared to 73.25% using only the AVA training set. Even though the accuracy is not as high as RDCNN using semantic attributes, the results indicate that by increasing the size of training data that associated with only aesthetics labels, the prediction accuracy could also be improved.

We attempted an alternative strategy to build the training dataset, that is, using the top 20% rated images as positive samples and the bottom 20% images as negative samples. The accuracy produced by $I_w$ and $I_r$ are 72.65% and 72.11%, respectively, and the accuracy of DCNN is 72.9%. The results are worse than using the entire IAD dataset for training with the mean value of 4.88 as the boundary for positive and negative training examples. The results may caused by two reasons. First, the AVA test set contains images with ratings in the middle. By cutting off images with medium scores from training, the prediction accuracy on medium-range test images may be affected. Second, utilizing the the top 20% and bottom 20% rated images reduces the number of training data. To fit the new dataset, network architectures have to be carefully adjusted in order to achieve good prediction results, which is non-trivial. We will take it as our future work and discuss the variations of problem formulation in image aesthetics and their corresponding performance.

In addition to the large scale, another advantage of the IAD dataset is that a sub-collection of images in the dataset is associated with camera parameters, such as Aperture/FNumber, ISO/ISOSpeedRatings, Shutter/ExposureTime, and Lens/FocalLength. While we did not use these information in this work, we believe
such information may facilitate future studies and help users to take aesthetically appealing photos.

Qualitative analyses were performed to demonstrate the advantage of the RDCNN architecture using semantic attributes. We present the examples in Figure 6.13, where we show representative test images that have been correctly classified by RDCNN_{semantic} but misclassified by DCNN. Our observation is that the classification accuracy improves on images that contain obvious objects.

### 6.3.7 Implementation Details and Computational Efficiency

All the networks presented in this work were implemented using ConvNet\(^{10}\), which supports multi-column inputs for a fully-connected layer. We used the logistic regression cost layer in all network trainings. We initialized the weights and biases learning rate of convolutional and fully-connected layers as 0.001 and 0.002, respectively. Both the weight momentum and the bias momentum were set to 0.9, and the dropout rate was 0.5 on all fully-connected layers. The detailed network architectures are presented in Sections 6.3.1 and 6.3.6.

On the AVA dataset, it takes 2 days to train SCNN for a certain input type, and about 3 days for DCNN. Training SCNN for style classification takes roughly a day, and 3-4 days for RDCNN training. With Nvidia Tesla M2070/M2090 GPU, it took about 50 minutes, 80 minutes, and 100 minutes for SCNN, DCNN, and RDCNN, to compute predictions of 2,048 images (each with 50 views) respectively. On the IAD dataset, training SCNN for a specific input type takes about 4 days, and training DCNN takes about 1 day using two SCNN as tow columns for initialization. Classifying 2048 images (each with 50 views) took about 60 minutes, 80 minutes for SCNN and DCNN, respectively, with Nvidia Tesla K40 GPU\(^{11}\).

### 6.4 Summary

This work studied deep neural network training approaches for image aesthetics assessment. In particular, we introduced a double-column deep convolutional neural

---

\(^{10}\)https://code.google.com/p/cuda-convnet/.

\(^{11}\)The Nvidia Tesla K40 GPU is faster than Nvidia Tesla M2070/M2090 GPU in testing because the SCNN and DCNN trained on the IAD dataset has much larger capacity than the one trained on the AVA dataset.
network approach to assess image aesthetics. Using such novel architecture, we learned aesthetic-related features automatically and unified the feature learning and classifier training. The proposed double-column architecture captures both the global and local views of an image for judging its aesthetic quality. We further developed a network adaptation strategy to apply the trained double-column network training approach for content-based image aesthetics. In addition, image style and semantic attributes were leveraged respectively to boost performance. Experimental results showed that our approaches produced significantly higher accuracy than earlier-reported results on the AVA test set, one of the largest existing benchmark with rich aesthetic ratings. Moreover, we introduced a 1.5 million IAD dataset for image aesthetics and improved the aesthetic assessment accuracy on the AVA test set. This result showed that the performance of image aesthetics could be further improved given a larger-scale training dataset.

One limitation of our work is that we have not yet been able to explain what we have learned exactly from the proposed network training approach and why those features help improve the performance. Visualizing trained neuron networks is one of the most active and significant research problems among recent deep learning studies. We would like to treat it as our future work.
Chapter 7

Conclusions and Future Work

This chapter concludes the dissertation. We summarize the major contributions and findings of this dissertation on automatic emotion recognition and image aesthetics assessment in Section 7.1. We present limitations of our studies and discuss future potentials in Sections 7.2 and 7.3, respectively.

7.1 Summary

The major contribution of this dissertation is the identification of effective visual features to automatically assess image emotions and aesthetics. The approaches for image emotion recognition developed in this dissertation advance existing methods in solving the following research questions: (1) distinguishing emotional images from neutral images; (2) determining whether images evoke positive or negative emotions (i.e., high valence vs. low valence); (3) identifying images evoke strong or weak emotions (i.e., high arousal vs. low arousal); and (4) recognizing specific types of emotions. The deep learning approach developed in this dissertation for image aesthetics assessment achieved the state-of-the-art performance on the AVA and IAD dataset\(^1\).

To advance the performance of automatic emotion recognition, we have developed a collection of shape features to encode the concepts of roundness, angularity, and simplicity to automatically recognize human emotion, given visual stimuli. We

\(^1\)The IAD dataset was proposed and built in this dissertation. The detailed description was in Section 6.3.6.
demonstrate the emotion recognition performance of proposed shape features on
the IAPS dataset by comparing with the state-of-art approaches.

To articulate what specific visual characteristics of those scenes might be
associated with a certain emotion, we have developed computational methods to
map images to the scales of roundness, angularity, and simplicity as three new
constructs. The relationship between the three constructs and human emotions were
examined on a large collection of ecologically valid image stimuli (i.e., the EmoSet),
where the EmoSet marks the first time such a large collection of ecologically valid
image stimuli was collected.

To generate effective features that are able to improve the performance of
existing visual features on image aesthetics assessment, we have explored the deep
neural network approaches and developed a double-column deep convolutional
neural network approach for aesthetic quality rating and categorization. The
aesthetics features, including features representing the global view and the local
view, were learned automatically by the proposed approach, which unified the
feature learning and classifier training. This is the first time that deep learning
approach was introduced for image aesthetics. The approach developed in this
dissertation achieved the state-of-the-art performance on the AVA and the IAD
datasets.

7.2 Limitations

The area of understanding emotions and aesthetics aroused by images is still in its
infancy. We itemize the main limitations of our solutions:

1. Emotion/Aesthetics Representation
At the current stage, we represent emotion using dimensional or categorical ap-
proaches, and we formulate the problem of image aesthetics utilizing a binary
classification framework. This representation might not be correct because different
people may hold different emotional and aesthetics feelings given the same visual
stimuli. Meanwhile, we need to assess the role of context or other semantic infor-

mation.

2. Psychologically Inspired Features
Although the proposed shape features, the constructs of roundness, angularity, and complexity, and CNN features encode high-level semantic meanings, the performance of using those visual features for emotion prediction is still far from human performance.

3. Classifiers
The current classifier does not take into consideration of the reliability of emotion or aesthetic labels. We discuss possible future work to address these issues in the next section.

4. Evaluation
The evaluation of an emotion or aesthetics prediction algorithm is highly challenging and difficult due to their subjective nature. Still, the developed large-scale emotion and aesthetics datasets provide an initial step towards better evaluation metrics to identify the strengths and weaknesses of image affect prediction algorithms.

7.3 Future Work

We may advance the computational modeling of image aesthetics and emotions in the following aspects:

1. Problem Formulation
A distribution could be used to represent emotions evoked by natural scenes to allow an image to evoke a mixed collections of emotions. For example, an image may arouse sadness in one group of population, whereas evoke happiness in another group of population.

2. Image Representation
We have obtained good progress in discovering useful image representation to predict emotion and aesthetics, such as roundness, angularity, simplicity, and complexity for emotion prediction and holistic and local characteristics for image aesthetics assessment. We need to design or learn more effective, high-level visual features to boost performance.

3. High-Level Classifiers
A classifier or regressor could be developed, considering the imperfect image affect labels, to perform prediction given visual features. Alternatively, we can develop a
classifier to leverage multi-modal information. We have made progress on designing classifiers that take heterogeneous inputs, and we need to continue our efforts to enable multi-modal inputs for emotion prediction and image aesthetics assessment. For instance, voice could also be an indicator of people’s emotion and aesthetic feeling when aroused by a given pictorial scene.

4. Evaluation
Currently, most emotion prediction algorithms are evaluated using accuracy-related metrics. Diverse metrics could be developed to evaluate an emotion prediction algorithm or an aesthetics prediction algorithm from different perspectives. For instance, some algorithms may be good at predicting the consensus of aroused emotions, whereas others may accurately predict the diversity of aroused emotions given a pictorial scene.

5. Accuracy with Very Large Dataset
We have built a comparatively large dataset (40K) for emotion prediction and a 1.5 million dataset for image aesthetics assessment. We need to build even larger dataset to capture more variations of pictorial scenes that evoke emotion and aesthetics.

6. Applications
We are interested in developing real-world applications of emotion prediction and image aesthetics assessment. Rather than attaching to existing services, such as image search engine, image editing tools, and personal photos management, more interesting standalone applications deserve explorations. For instance, emotion recognition tools could be used to analyze and summarize movie clips. They could also be used to select the highlights from movie clips or personal albums.
Appendix

Image-Specific Prior Adaptation for Denoising

Image priors are essential to many image restoration applications, including denoising, deblurring, and inpainting. Existing methods use either priors from the given image (internal) or priors from a separate collection of images (external). We find through statistical analyses that unifying the internal and external patch priors may yield a better patch prior. We propose a novel prior learning algorithm that combines the strength of both internal and external priors. In particular, we first learn a generic Gaussian Mixture Model from a collection of training images and then adapt the model to the given image by simultaneously adding additional components and refining the component parameters. We apply this image-specific prior to image denoising. Experimental results show that our approach yields better or competitive denoising results in terms of both the peak signal-to-noise ratio and structural similarity.

1 Introduction

Noise is a fundamental problem in measuring light. No matter how good the sensors are, there is noise in images, especially in low-light conditions. Image denoising is the problem of reducing undesired noise in images. It has been studied extensively over the last half century because of its practical importance. The problem is mathematically ill-posed and image priors are used to regularize it such that meaningful solutions exist.

There are two kinds of image priors one can use: priors learned from the given
image and priors learned from a separate set of images. We follow the common naming convention and refer to the former as internal priors and the latter as external priors (or generic priors). Smoothness and piecewise smoothness are probably the simplest form of internal priors. They led to many successes of PDE-based denoising algorithms in the nineties. Recently, more interesting internal priors, such as patch self-similarity, have been proposed. They have led to methods such as BM3D [103] which is still regarded as one of the state-of-the-art methods in image denoising. The obvious limitation of internal priors is that external images are completely ignored. For instance, for any given image, one can find similar images (at least in terms of parts) that contain significantly less noise and use these similar images to do a better job of denoising. It was a breakthrough in image denoising in going beyond internal priors to external priors. Allowing external images opens a wide range of possible priors. Popular priors that one learns from external images include sparse representations and nonlinear regression functions between clean and noisy patch pairs. One of the top-performed image denoising algorithms is based on external images [104]. There is one problem with external priors: It is well known that images form a heavy-tail distribution. No matter how large the external image set is, some images will fall in the heavy tail, i.e., will not be well-modeled by the learned priors. We believe this problem prevents methods that use external images from significantly outperforming those that do not.

There exists a natural question which is if there is a way to combine internal and external priors. It turns out that we are not the first to ask this question. For instance, [105] and [106] explored similar problems. They discovered that internal and external priors are good for different types of image patches and proposed methods to combine them. Strictly speaking, both [105] and [106] focus on combining internal and external denoising algorithms rather than internal and external priors. Instead, in this work, we focus on priors rather than specific denoising algorithms. It is our belief that in addition to mathematical elegance, priors are more fundamental than specific algorithms. One can use different denoising algorithms with the same priors. The main idea in this work is that we can learn generic priors from external images and adapt them to a specific image using internal priors learned from that image. We overcome the heavy-tail problem in external priors by adapting generic priors to a specific image.
1.1 Related Work

State-of-the-art image denoising approaches leverage various types of patch priors for regularizing the ill-posed nature of the problem. In general, the priors can be categorized into internal patch priors and external patch priors.

Internal patch priors refer to patch statistics derived from the image itself. Typical examples include patch self-similarity [103,107], sparsity prior [108], structural similarity [109], and patch recurrence across image scales [110]. The local patch self-similarity has been quite successful for denoising due to its effectiveness and simplicity. Non-Local Means [107] denoises each pixel of an image based on the weighted average of central pixels of similar patches. BM3D [103] exploits both patch self-similarity and 3D transform domain collaborative filtering, which achieves state-of-the-art performance in denoising. Mairal et al. [108] further advanced the patch similarity idea with sparse coding priors. To use patch recurrence across image scales, Zontak et al. [110] attempted to find clean versions of noisy patches by searching similar patches across multiple image scales. To leverage structural similarity, Dong et al. [110] examined the structural information of the entire image and considered the joint sparsity for noise removal.

External patch priors refer to patch statistics or denoising operators learned from external image sets, such as statistical distribution of image patches in natural images [111], sparse representation [112], global statistical prior [113], and regression functions from noisy patches to clean patches [104,114,115]. Roth and Black [113] used Markov Random Field to learn generic image priors. In [112], a dictionary learning-based method is introduced for compact patch representation, whereas in [111], a Gaussian Mixture Model (GMM) is learned from external patch databases and used as a prior for denoising. More recent work favors mapping functions learned through neural networks. Burger et al. [104] employed a plain multi-layer perceptron to learn the mapping functions between pairs of noisy patches and noise-free patches. Cho [114] applied the Boltzmann machines to map noisy images to clean images. In the same vein, Xie et al. [115] combined sparse coding and deep networks pre-trained with denoising auto-encoder for the training scheme to learn external priors.

Although internal and external priors have been widely used in previous work, little effort has been devoted to combining them in a principled way. In this work,
we propose a novel, unified prior model combining internal and external priors. Most recent works, such as Mosseri et al. [106] and Burger et al. [105], are related to ours in that they also leverage both internal and external priors. Mosseri et al. [106] developed a Patch-SNR measure to decide whether a noisy patch is denoised using internal priors or external priors. Burger et al. [105] attempted to learn a non-linear regression function that can map two denoising results, one with the internal prior and the other with the external prior, to produce a better denoising result. Unlike these methods where the denoising is conducted separately with either the internal or external priors on image patches, we introduce a unified prior which is fundamentally different from those methods. The prior is general and thus can be applied to any image restoration application. Specifically, we derive the prior in the context of denoising. During the offline phase, a generic patch prior is learned from an external set of natural images; and during the online phase, we adapt the generic patch distribution to the test image by analyzing patch statistics in the test image. Our work is also related to EPLL [111] in that both use GMM to represent patch distributions. The difference is that our method adapts the generic prior to the test image while EPLL relies on the fixed prior model, additionally, our method does not require the reconstruction term and computationally expensive iterations as in EPLL.

The approach proposed by Wang and Morel [116] is very close to ours; they also adapt the generic prior, which is a mixture model, to the test images. The difference is that their approach conducts an “in-place” modification to the prior (i.e., components of the mixture model are shifted and deformed according to image-specific information) while ours augments the prior model by image-specific component addition (see Section 3.1 for details). Lebrun et al. [117] proposes a patch-based denoising method where a noisy patch is restored by nearby similar patches assuming from a Gaussian model.

1.2 Contributions

The main contributions of this work are the following: 1. We propose a unified algorithm that learns a generic prior from external images and adapts the prior to a specific image. This is fundamentally different from combining different denoising algorithms as done in [105, 106]. 2. We show that by adapting a generic prior into
an image-specific one, we do not need global image reconstruction¹ as done in [111]. This significantly speeds up the algorithm.

2 Modeling Patch Statistics using Gaussian Mixture Model

This section introduces learning patch prior using Gaussian Mixture Model. We learn the GMM model using an external patch database (external patch statistics) and using patches in one image (internal patch statistics), respectively. We empirically analyze the internal and external GMM patch priors, and then discuss ideas of leveraging both priors into a unified natural image patch prior.

2.1 Gaussian Mixture Model

We learn finite Gaussian Mixture Model (GMM) over natural image patches \( \{x_i\} \) as patch priors. Using the GMM model, the log likelihood of a given patch \( x_i \) is:

\[
p(x_i) = \sum_{k=1}^{K} \pi_k N(x_i | \mu_k, \Sigma_k),
\]

where \( x_i \) is a D-dimensional vector, \( K \) is the total number of mixture components chosen for the GMM, \( N \) is the total number of patches in the training set, and the GMM model is parameterized by mean vectors \( \{\mu_k\} \), covariance matrices \( \{\Sigma_k\} \), and mixture weights of mixture components \( \{\pi_k\} \). We collectively represent these parameters by \( \Theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^{K} \), and \( \Theta \) is learned using Expectation Maximization algorithm (EM).

In the E-Step, we calculate the posterior probability for the component \( k \) as:

\[
Pr(k|x_i, \Theta) = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_i | \mu_k, \Sigma_k)},
\]

\[
n_k = \sum_{i=1}^{N} Pr(k|x_i, \Theta).
\]

¹The “global image reconstruction” refers to using a cost function on overlapping patches extracted from the noisy image for image denoise, rather than merely applying denoising operators on each individual noisy patch, e.g., the algorithm presented in [111], Section 3.1.
In the **M-Step**, we update the model parameters as follows:

\[
\pi_k = \frac{n_k}{N}, \quad (4)
\]

\[
\mu_k = \frac{\sum_{i=1}^{N} \pi_k x_i}{\sum_{i=1}^{N} \pi_k}, \quad (5)
\]

\[
\Sigma_k = \frac{\sum_{i=1}^{N} \Pr(k|x_i, \Theta)x_i x_i^t}{n_k}. \quad (6)
\]

We iterate over the E-Step and M-Step until convergence.

### 2.2 Building External and Internal GMM

#### 2.2.1 External/Generic GMM

We built a generic, external GMM (denoted as GMM\(_{\text{ext}}\) with parameters \(\Theta_{\text{ext}}\)) following the settings in [111]. We used the same image collection as used in [111], i.e., the BSD training dataset\(^2\) (200 images in total). We densely sampled 50K 8 × 8 zero-mean patches for GMM model training as discussed in Section 2.1\(^3\). We set \(K = 200\) for the external GMM model learning, and all images are converted into gray scale in this work.

#### 2.2.2 Internal GMM

We took the BSD test set (100 images) for the analysis purpose, and none of the images in the BSD test set was used for external GMM training. We built an internal GMM model of each individual image in the BSD test set. For each image, we extracted all 8 × 8 overlapping patches, generated zero-mean patches, and trained the GMM model as discussed in Section 2.1. We let all internal GMM models (denoted as \{GMM\(_{\text{int}}\}\) with parameters \{\Theta\(_{\text{int}}\)\}) to have the same number of components with the external GMM model, i.e., \(K = 200\).

\(^2\)[http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/].

\(^3\)We represent a patch by a \(D-1\) vector, and zero-mean patches refers to the mean-subtracted patches.
2.3 Internal and External Statistics

For each image in the BSD test set, we extracted all $8 \times 8$ overlapping patches. For each patch $x$, we computed $P_{\text{ext}} = \max_k Pr(k|x, \Theta_{\text{ext}})$ and $P_{\text{int}} = \max_k Pr(k|x, \Theta_{\text{int}})$ presented in Equation 3. For some patches that have low probabilities in the GMM$_{\text{ext}}$ (i.e., these patches are not very frequent across all images) but higher probabilities in GMM$_{\text{int}}$ (i.e., many self-similar examples), we might be able to model them better. There are some patches that cannot be modeled well with either GMM$_{\text{ext}}$ or GMM$_{\text{int}}$ (i.e., not self-repeating patterns). We present example images in Figure 1, where each pixel refers to a patch’s center pixel ($(4, 4)$ in a $8 \times 8$ patch). We annotate the pixel as blue if the corresponding patch has a low probability in GMM$_{\text{ext}}$, and otherwise we annotate the pixel in red. These results indicate that we may get a better patch prior by unifying the GMM$_{\text{int}}$ and GMM$_{\text{ext}}$ priors.

3 Denoising with Image-Specific Prior

Based on our analysis of the patch statistics in Section 2, we propose a novel method to learn image-specific prior by unifying internal and external GMM image patch priors. Figure 2 gives an overview of our approach. We first train a generic GMM model on a collection of patches randomly sampled from a set of clean natural images (as presented in Section 2.2), which serves as our generic GMM patch prior. The training is off-line and only needs to be performed once. Given a noisy image,
Figure 2: Approach overview. We first trained a generic GMM patch prior using a collection of natural image patches as introduced in Section 2. We represent the generic GMM model in magenta. Given a noisy image, we conduct a two-step adaptation to generate an image-specific patch prior. In the component addition, new Gaussian components (in yellow) are constructed for individual or image-specific patches that the generic GMM prior fails to model. In component adaptation, GMM adaptation is performed to better fit the distribution of both generic and image-specific patches of the given image. Equipped with the image-specific prior, we can conduct more effective denoising on the input.

we conduct a two-step adaptation to generate an image-specific patch prior inspired by the Kolmogorov-Smirnoff (KS) Test. The first step constructs new Gaussian components for individual or image-specific patches that the generic GMM prior fails to model, and the second step performs GMM adaptation to better fit the distribution of both generic patches and image-specific patches. Equipped with the image-specific prior, we can conduct more effective denoising on the input. In the following, we first detail the two steps for learning an image-specific prior for a clean image, and then we describe how to adapt the procedure to noisy inputs for denoising.

3.1 Image-Specific Components Addition

A generic GMM may not be able to model every image patch well, considering the large space of natural image patches (e.g., $8 \times 8$ patches). The image-specific patches of a given image will be outliers in the view of a generic GMM prior,
Algorithm 1: Component Addition

**Input:** Generic GMM model with parameters of $\Theta_g$, a coarse image-specific GMM model $\Theta_x$.

1. $\Theta_{output} = \Theta_g$, $K_{mod} = K_g$
2. for $\Theta_k \in \Theta_x$, $k \in [1, K_y]$
3. flag = true
4. for $\Theta_k \in \Theta_g$, $k \in [1, K_g]$
5. if $D_{KL}(\Theta_k, \Theta_g) < \delta$
6. flag = false; break;
7. end if
8. end for
9. if flag == true
10. $\Theta_{output}^{K_{mod}+1} = \Theta_k$
11. $K_{mod} = K_{mod} + 1$
12. end if
13. end for

and thus we need to construct new components to model these outlier patches. This process is essentially similar to the online clustering or online GMM in the statistics literature. We build the new Gaussian components and add them through online GMM. Specifically, given an image represented by a collection of overlapping patches $Y = \{y_i\}_{i=1}^M$, where $M$ is the total number of patches in the image. We first generate an image-specific (internal) GMM model as presented in Section 2.2, where we learn parameters $\Theta_y = \{\mu_{in}^k, \Sigma_{in}^k, \pi_{in}^k\}_{k=1}^{K_y}$, where $K_y$ is the total number of components. We let $K_y = 200$, the same number of Gaussian components as in the generic GMM. To accelerate the process, we first do a k-means clustering on the internal patches and do one EM iteration to generate a coarse image-specific GMM model. Given the generic GMM model $\Theta_g = \{\mu_g^k, \Sigma_g^k, \pi_g^k\}_{k=1}^{K_g}$, we discover representative components and add them to the generic GMM prior following Algorithm 1, where $D_{KL}(\Theta_k, \Theta_g)$ is the Kullback-Leibler (KL) divergence$^4$ defined as follows:

$^4$While the Kullback-Leibler (KL) divergence is not symmetrical, the same generic GMM is used for all images. In our implementation, we computed the KL divergence on pairs of an internal GMM component and an external GMM component, where we fixed the order of internal and external GMM components in each of the pairs. We thus avoid the problem caused by the unsymmetrical KL divergence.
\begin{equation}
D_{KL}(\Theta_y^k, \Theta_y^k) \\
= \int [\log(\Theta_y^k(t)) - \log(\Theta_y^k(t))] \Theta_y^k(t) dt \\
= \frac{1}{2} \left[ \log \left( \frac{|\Sigma_y^k|}{|\Sigma_y|} \right) - d + tr((\Sigma_y^k)^{-1}\Sigma_y) + ((\mu_y^k - \mu_g^k)^T(\Sigma_y^k)^{-1}(\mu_y^k - \mu_g^k)) \right],
\end{equation}

where $d$ is the dimension of the patch vector. $\delta$ is the threshold for detecting new Gaussian components: if the divergence between the current component and the nearest generic component is larger than $\delta$, a new component is added. We empirically determine the value of $\delta$ for denoising in the experimental section. Note that the image-specific GMM $\Theta_y$ is coarsely constructed for efficiency. We discuss how to refine the new GMM to fit the patch distribution in the following subsections.

### 3.2 Gaussian Mixture Model Adaptation

Equipped with the added new Gaussian components capturing the image-specific patches, we are ready to adapt our complete image specific GMM to the patch distribution of the given image. To avoid notation clutter, we use $\Theta_0 = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ ($K$ is the number of mixtures) to denote the current GMM after component addition. Given the patches extracted from the input image $Y = \{y_i\}_{i=1}^M$, we want to estimate the image-specific GMM parameters $\hat{\Theta} = \{\hat{\mu}_k, \hat{\Sigma}_k, \hat{\pi}_k\}_{k=1}^K$, which can be solved via maximizing a posteriori (MAP) estimation:

\begin{equation}
\hat{\Theta}_{MAP} = \arg \max_{\Theta} p(\Theta|Y) \\
= \arg \max_{\Theta} f(Y|\Theta)g(\Theta),
\end{equation}

where $g(\Theta)$ denotes the prior information of the unknown GMM parameters.

The adaptive GMM model is derived by following the GMM adaptation algorithm proposed in [119]. In the **E-Step**, we calculate the posterior probability for the component $k$ as:

\begin{equation}
\Pr(k|y_i, \Theta) = \frac{\pi_k N(y_i|\mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k N(y_i|\mu_k, \Sigma_k)},
\end{equation}

where $\pi_k$ is the prior probability of component $k$, $N(y_i|\mu_k, \Sigma_k)$ is the probability density function of a Gaussian distribution with mean $\mu_k$ and covariance $\Sigma_k$, and $\sum_{k=1}^K \pi_k N(y_i|\mu_k, \Sigma_k)$ is the normalizing constant.
\[ n_k = \sum_{i=1}^{N} Pr(k|y_i, \Theta) . \] (10)

In the **M-Step**, we update the model parameters as follows:

\[
\hat{\pi}_k = \left[ \alpha^\pi n_k / N + (1 - \alpha^\pi) \pi_k \right] \gamma ,
\]

(11)

\[
\hat{\mu}_k = \alpha^\mu \frac{\sum_{i=1}^{N} \pi_k y_i}{\sum_{i=1}^{N} \pi_k} + (1 - \alpha^\mu) \mu_k ,
\]

(12)

\[
\hat{\Sigma}_k = \alpha^\Sigma \frac{\sum_{i=1}^{N} Pr(k|y_i, \Theta) y_i y_i^T}{n_k} + (1 - \alpha^\Sigma) \Sigma_k ,
\]

(13)

where \( \gamma \) is the scale factor to ensure \( \sum_{k=1}^{K} \hat{\pi}_k = 1 \). We iterate over the E-Step and M-Step until convergence. In practice, we find that it is good enough to just run one iteration of EM by setting the parameter adaptation rates \( \alpha^\pi \), \( \alpha^\mu \) and \( \alpha^\Sigma \) to be 1, which means we fully trust the new input patches. We detail the parameter selection of \( \alpha^\pi \), \( \alpha^\mu \) and \( \alpha^\Sigma \) in the experimental results.

### 3.3 Learning Image-Specific Prior for Denoising

The above image-specific prior learning assumes clean input images. To learn an image-specific prior from noisy images for denoising, we have two options for the model construction:

1. Assuming zero mean for the noise, the model parameters affected are mainly the covariance matrices. Also assuming a diagonal covariance matrix \( \Sigma_n \) for the noise, we can estimate the coarse image-specific GMM parameters with \( \Sigma_{in}^k - \Sigma_n \) for component addition\(^5\). For GMM adaptation, we can similarly subtract \( \Sigma_n \) from the estimation in Equation 13. We found that this scheme works well for small noise cases (noise with low variance).

2. We first apply the generic GMM prior for denoising to get an initial result, from which we can learn an image-specific prior, which is more accurate for the final denoising. In practice, we find this scheme works well for strong noise cases.

\(^5\)Through out this chapter, we assume the noise is additive and zero-mean and the variance of signal is far larger than the variance of noise to apply \( \Sigma_{in}^k - \Sigma_n \). Practically, to avoid singularity of covariance matrix in GMM computation, we add very small value (e.g., 1e-5) to its diagonal elements.
After we have learned an image-specific prior for the noisy input image, we first extracted overlapping noisy patches from the input image. For each noisy patch $y$, the Bayes Least Square solution of its denoised version $\hat{y}$ is $\hat{y} = E[x|y] = \sum_{k=1}^{K}[Pr(k|y)E[x|y, \Theta_k]]$, where $\Theta_k$ represents the parameters of $k$-th component of the mixture. Considering the large number of overlapping patches in an image (more than 60K patches in a 256 x 256 image), this solution is costly if we utilize all Gaussian components to denoise each patch. In practice, we took an approximate solution as presented in [111] and found that this simple solution worked well in general. We discuss this approximation in the experimental section.

Specifically, for each noisy patch $y$, we select the Gaussian component with the maximum posterior probability:

$$k^* = \arg\max_k Pr(k|y, \Theta).$$

(14)

We then apply the Wiener filter for noise removal as discussed in [111]:

$$\hat{y} = (\hat{\Sigma}_k + \Sigma_n)^{-1}(\hat{\Sigma}_k y + \Sigma_n \hat{\mu}_k),$$

(15)

where $\Sigma_n = \sigma^2 I$. With all overlapping denoised $\hat{y}$, we generate the final denoised image by averaging the overlapping pixels.

### 3.4 Discussions

There are alternative ways of learning an image-specific patch prior. One could be to adapt the generic GMM to the given image without constructing any new components (adaptive GMM), and the other is to directly learn a GMM based on the internal examples only (internal GMM). Compared with these two methods, our method is superior in terms of speed and potentially performance:

1. As we mentioned before, the image-specific patches are outliers in the view of the generic GMM. As a result, adapting the generic GMM takes more iterations in order to model those outlier patches well.

2. Building an internal GMM based on the internal examples alone is prone to overfitting to noise. It is also very slow for the algorithm to converge to a reasonably good model for denoising. Note that in our component
addition step, we also train a coarse image-specific GMM for finding new components based on fast kmeans for efficiency. The coarse image-specific GMM is different from the internal GMM, which needs to be well trained and thus is much slower.

4 Evaluation

4.1 Synthetic Noisy Images

We evaluate the effectiveness of our approach by comparing the image-specific prior generated by the proposed approach with priors generated by internal GMM and adaptive GMM (discussed in Section 3.3) as well as state-of-the-art denoising approaches such as BM3D [103], EPLL [111], MLP [104], NLB [117], SPLE [116], combined internal and external method (CBIE) [106]<sup>6</sup>, and combined BM3D and MLP (CBBM) [105]<sup>7</sup>. Our approach achieves better or competitive image denoising results in terms of PSNR (i.e., Peak Signal-to-Noise Ratio) and SSIM [120] (i.e., Structural Similarity). In the following, we refer to our approach as “OL-GMM”, internal GMM as “ING-GMM”, and adaptive GMM as “ADA-GMM”. We also use the global reconstruction term as in [111] to refine the performance of OL-GMM, and we denoted it as “OL-GMM-RT”. We made a single EM iteration for OL-GMM and ADA-GMM. In OL-GMM-RT, we iterate once using the reconstruction term defined in [111] with the OL-GMM result.

To evaluate our performance quantitatively, we corrupted clean images with the white Gaussian noise and used the same noisy test images to generate the results of all the methods for fair comparison. All our experiments reported in this chapter are conducted on gray-scale images.

The training patches used in our experiments were sampled from the Berkeley Segmentation Dataset (BSD) (training images)<sup>8</sup>. The performance of our approach is evaluated on the four benchmark datasets including standard test images [103], Berkeley Segmentation Dataset (BSD) (testing images)<sup>8</sup>, Pascal VOC2007 dataset<sup>9</sup>,

<sup>6</sup>The results of [106] are based on our implementation of [106]. NLM [107] is used for both internal denoising and external denoising.

<sup>7</sup>The results of BM3D [103], EPLL [111], MLP [104], NLB [117], SPLE [116], and CBBM [105] are based on authors’ source codes released on the Web.

<sup>8</sup>http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsd5/. 

Table 1: Average PSNR on standard images with different $\alpha$

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 5$</td>
<td>38.07</td>
<td>38.08</td>
<td>38.09</td>
<td>38.12</td>
<td><strong>38.16</strong></td>
</tr>
<tr>
<td>$\sigma = 10$</td>
<td><strong>34.37</strong></td>
<td><strong>34.37</strong></td>
<td><strong>34.37</strong></td>
<td><strong>34.37</strong></td>
<td><strong>34.37</strong></td>
</tr>
<tr>
<td>$\sigma = 20$</td>
<td>30.9</td>
<td>30.92</td>
<td>30.94</td>
<td>30.94</td>
<td><strong>30.95</strong></td>
</tr>
<tr>
<td>$\sigma = 30$</td>
<td>28.88</td>
<td>28.9</td>
<td>28.93</td>
<td>28.94</td>
<td><strong>28.95</strong></td>
</tr>
</tbody>
</table>

and McGill dataset\(^{10}\).

### 4.2 Evaluation on The Standard Test Images

We empirically determined the value of $\delta$ for image denoising when noise is of different variance. We evaluated average PSNR on the standard images as $\delta$ changes, and we present results in Figure 3, where the x-axis denotes $\delta$, and the y-axis refers to average PSNR. As shown in the figure, the PSNR drops when $\delta$ is larger than 10 for all the four noise levels. In particular, we found that setting $\delta$ to 1 provided the highest PSNR for denoising when the $\sigma$ is 5 or 10, and setting $\delta$ to 10 provided the highest PSNR for denoising when $\sigma$ is 20 or 30. We thus set $\delta$ as 1 for all the remaining evaluation when $\sigma$ is 5 or 10 and set $\delta$ as 10 when $\sigma$ is 20 or 30.

\(^{10}\)http://pirsquared.org/research/mcgilldb/.
Similarly, we empirically determined the value of $\alpha$ (i.e., $\alpha^\pi$, $\alpha^\mu$ and $\alpha^\Sigma$). We evaluated denoising results in terms of PSNR on standard images by setting $\alpha$ as 0.2, 0.4, 0.6, 0.8, and 1, respectively. In all these experiments, we ran one iteration of EM for GMM adaptation, and present results in Table 1. As shown in the table, for all the four noise levels, the highest PSNR is achieved when $\alpha$ is 1. We thus set $\alpha$ as 1 for all the remaining evaluation.

We present our evaluation results (average PSNR and SSIM) on the standard test images in Figure 4, with different noise levels ($\sigma = 5, 10, 20, 30$). The test image dataset includes 14 images: pepper, house, boat, couple, man, lena, fingerprint (denoted as “fprint”), barbara, montage, monarch, hill, straw, cameraman (denoted
Figure 5: Denoising performance on standard images. The proposed online GMM approach (OL-GMM) (in green) was compared with EPLL [111] (in blue) and BM3D [103] (in red) for $\sigma = 5, 10, 20, 30$. The figure is better viewed in color.

As shown in Figure 4, compared with image priors generated with ADA-GMM and INT-GMM, OL-GMM produces significantly better denoising performance at all noise levels. Also, the OL-GMM approach achieves competitive results to both [106] and [105] that leverage internal and external patch priors, respectively. As shown in Figure 4, OL-GMM outperforms all other competing methods at $\sigma = 5$ in terms of average performance on standard images. In particular, our method outperforms BM3D by 0.15 db in average PSNR, which is a remarkable improvement considering that BM3D is still the state of the art for image denoising with low variance (small noise) cases.

By comparing OL-GMM with OL-GMM-RT, we identify that with an image-specific prior, we do not need global image reconstruction as in [111]. Compared with EPLL [111], we achieved a better denoising result with a single EM iteration while the former requires multiple iterations. For large $\sigma$, our approach can be further improved if we use complimentary internal self-similarity-based methods such as BM3D to initialize our internal GMM model. We leave this fusion method as our future work.

Individual results of $\sigma = 5, 10, 20, 30$ are presented in Figure 5, where the proposed OL-GMM approach was compared with EPLL [111] and BM3D [103]. The advantage of our method at small noise levels can be further validated from
the comparisons for individual test images, shown in Figure 5, which demonstrate that the proposed OL-GMM approach achieves the best result on 12 out of the 14 benchmark images. Small noise levels ($\sigma < 5$) has practical importance since real-world images are mostly corrupted by such noises.

To examine the accuracy of the adopted approximated solution in Section 3.3, we empirically evaluated denoising results in terms of PSNR using 1 to 10 largest GMM posterior probability components, respectively, and present the results in Figure 6. As shown in the figure, when $\sigma$ is 5, the highest average PSNR is achieved when the two largest-posterior-probability components are used; and when $\sigma$ is 10, 20, or 30, the highest average PSNR is achieved by only using the component with the maximum posterior probability. These results demonstrate the accuracy of the approximated solution of only using the component with the maximum posterior probability for noising. The phenomenon that the average PSNR decreases when using more GMM components may be caused by the imperfect statistical patch model (i.e., the GMM model).

With Matlab implementation using Intel(R) Xeon(R) CPU X5550 @ 2.67GHz, it took about 2 minutes to denoise a $256 \times 256$ image (all overlapping patches were used in denoising). The computational bottlenecks lie in two parts: 1) clustering internal patch and generating a coarse image-specific GMM model and 2) one
4.3 Evaluation on VOC2007, McGill, and BSD-Test

To further verify the effectiveness of the proposed approach, we compared our results produced by OL-GMM to those produced by BM3D [103] and EPLL [111] on the BSD-test, Pascal VOC2007, and McGill datasets. We used all of the 100 images in the BSD-test set, and we randomly sampled 100 images from both the Pascal VOC2007 set and the McGill set. We present the average PSNR and SSIM results on different noise levels ($\sigma = 5, 10, 15, 20, 25, 30$). The comparison of denoising performance is presented in Figure 7. As shown in the figure, with the image-specific prior, we achieve competitive denoising results on all noise levels. In particular, for small noise levels (i.e., $\sigma = 5$ and $\sigma = 10$), the OL-GMM consistently generate better denoising results than do the EPLL [111] approach on the test images in all three datasets, and it approaches BM3D in terms of quality.
Figure 8: Examples of added components and associated denoised patches. The first row shows the covariance matrix of the components, and the second row presents the patches being assigned to the components.

4.4 Qualitatively Analysis of Image-Specific Patches

In addition to performing quantitative analyses, we also qualitatively analyze patches that are referred to as “image-specific patches”. We take the image of “barbara” as an example, and visualize examples of added components given the noisy image ($\sigma = 5$). As shown in Figure 8, frequently recurring patches in the test image tend to form new components, which indicates that those patches are the referred “image-specific patches”. Those patches have higher density among patches in the same image than patches sampled from a diverse collection of images have. For instance, the texture of stripe frequently recurred in the “barbara” image, and patches with those patterns may not recur frequently in natural image patches. As shown in Figure 8, when we conduct denoising on the corrupted “barbara” image, patches with stripe texture tend to form new Gaussian components, and the new Gaussian components are used to denoise the image. Patches with low-contrast patterns also tend to be “image-specific patches” in case they appear frequently in the test image, as shown in the third component in Figure 8.

We found that “image-specific” patches are spatially clustered together, i.e., “image-specific patches” tend to cluster in some region of the image and forms new Gaussian components. By forming individual GMM components for these “image-specific patches”, we adapt the generic GMM prior to an image-specific prior, which leverages the local patch similarity for image denoising.
Figure 9: Denoising results on real images. On each row, the left column shows the original noisy image, and the right column shows the denoised results using “online-gmm”.

4.5 Real Images

To examine the performance of the proposed approach on real noisy images, we present some denoising results in Figure 9.
5 Summary

We presented a unified algorithm for learning an image-specific patch prior and applied the prior to the image denoising problem. Rather than combining internal and external denoising results, we unified the two types of priors in a principled way in order to use more fully the internal and external information. We demonstrated the effectiveness of our approach by comparing with image-specific priors generated by alternative approaches. With the novel prior, we achieved a better or competitive denoising performance in terms of the peak signal-to-noise ratio and structural similarity.

As a future work, the proposed image-specific patch prior could be used for other restoration tasks, such as image deblurring and image super-resolution. Since our method provides an improved prior over EPLL, it can be applied to deblurring by following the deblurring algorithm in Zoran and Weiss, ICCV2011 [19]. The image specific prior can be first learned from any existing deblurring algorithm and then the blurred input image can be restored based on the learned prior, with which the restoration result could be iteratively improved. For super-resolution, we can learn the image-specific prior from the low-resolution input and then use it to constrain re-constructed patches in high resolution in an MAP framework.
Bibliography


Vita

Xin Lu

Xin Lu received her B.E. and B.A. degree in Electronic Information and Engineering and English, the M.E. Degree in Signal and Information Processing, all from Tianjin University, China, in 2008 and 2010 respectively. She received her Ph.D. degree in Spring 2016 from the College of Information Sciences and Technology, The Pennsylvania State University, University Park. She worked as a research intern at Microsoft Research Asia in 2009-2010 and at Adobe Systems, Inc. in the summers of 2012, 2013, and 2014. She started working at Adobe Systems Inc. as a Research Scientist since Aug, 2015. Her research interests include computer vision and multimedia analysis, deep learning, image processing, Web search and mining, and user-generated content analysis.