UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Image Emotion Recognition using Region-based Multi-level Features

by

Tianrong Rao

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2019

Certificate of Authorship/Originality

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as a part of the requirements for other degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research and in the preparation of the thesis itself has been fully acknowledged. In addition, I certify that all information sources and literature used are quoted in the thesis.

Signature of Student: Production Note: Signature removed prior to publication.

Date: 2019.02.20

ABSTRACT

Image Emotion Recognition using Region-based Multi-level Features

by

Tianrong Rao

According to psychology studies, human emotion can be invoked by different kinds of visual stimuli. Recognizing human emotion automatically from visual contents has been studied for years. Emotion recognition is an essential component of human-computer interaction and has been involved in many applications, such as advertisement, entertainment, education, and accommodation system. Compared to other computer vision tasks, visual emotion recognition is more challenging as it involves analyzing abstract emotional states which are complexity and subjectivity. For complexity, emotion can be evoked by different kinds of visual content and the same kind of visual content may evoke various kinds of emotions. For subjectivity, people from different cultural background may have different kinds of emotions for the same kind of visual content. Automatic visual emotion recognition system consists of several tuned processing steps which are integrated into a pipeline. Previous methods often relay on hand-tuned features which can introduce strong assumptions about the properties of human emotion. However, the vague assumptions related to the abstract concept of emotion and learning the processing pipeline from limited data often narrows the generalization of the visual emotion recognition system.

Considering the two challenges on complexity and subjectivity as mentioned above, more information should be used for image-based emotion analysis. Features from different level including low-level visual features, such as color, shape, line and texture, mid-level image aesthetics and composition and high-level image semantic need to be taken into consideration. Local information extracted from emotion-related image regions can provide further support for image emotion classification.

In recent years, deep learning methods have achieved great success in many computer vision tasks. The state-of-art deep learning methods can achieve performances slightly under or even above human performances in some challenging tasks, such as facial recognition and object detection. The Convolutional Neural Networks applied in deep learning methods consist of hierarchical structures which can learn increasingly abstract concept from local to global view than hand-crafted features. This observation suggests exploring the application of CNN structure to image emotion classification. This thesis is based on three articles, which contribute to the field of image emotion classification.

The first article is an in-depth analysis of the impact of emotional regions in images for image emotion classification. In the model, multi-scale blocks are first extracted from the image to cover different emotional regions. Then, in order to bridge the gap between low-level visual features and high-level emotions, a mid-level representation, exploiting Probabilistic Latent Semantic Analysis (pLSA) is introduced to learn a set of mid-level representations as a set of latent topics from affective images. Finally, Multiple Instance Learning (MIL), based on the multi-scale blocks extracted from an image, is employed to reduce the need for exact labeling and analyze the image emotion. The experimental results demonstrate the effectiveness of emotional regions in image emotion classification.

However, one drawback of the method described in the first article is the hand-crafted using in this method is only valid for limited domains of affective image. The experimental results show that the performance of the method in abstracting paintings, whose emotion is mainly conveyed by low-level visual features, is not as well as in images that contain emotional content. CNN can automatically learn generalized deep features for various kinds of affective images. Therefore, in the second article, we analyze the different level of deep representations extracted using CNN from affective images. A comparison of CNN models with different modalities that exploit different level of deep representations shows the significant improvement of our proposed network fusing different level of deep representations for image emotion recognition. In addition to the proposed model, a Recurrent Neural Network

(RNN) with bi-direction Gated Recurrent Unit (GRU) can be added to deal with the correlations between different level of deep representations to further improve the performance of the proposed model.

The last article proposes a new framework based on Region-based CNN (RCNN) to integrate the different level of deep representations extracted from both global and local view. The framework consists of a Feature Pyramid Network (FPN) to extract and fuse different level of deep representations and a RCNN to detect emotional regions in the image. What's more, the framework also considers the label noise existing in the training dataset, an estimated emotion distribution derived from the reliability of emotion label of the image is used to improve the image emotion classification results. The integrated feature and new loss function considering label noise help the framework to achieve state-of-the-art performance for image emotion classification.

In summary, this thesis explores and develops a deep learning framework using region-based multi-level features for image emotion recognition, making significant steps towards the final goal of efficiently recognizing emotion from visual contents.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Min Xu for introducing me to the field of computer vision, for continuous support of my Ph.D study and research and for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. It's a great honor for me to have such excellent supervisor for my Ph.D study.

I also would like to appreciate my co-supervisor Dr. Xiaoyong Kong for providing me with continuous support throughout my Ph.D study and research.

I thank my fellow lab mates in Global Big Data Technical Center: Haimin Zhang, Lingxiang Wu, Yukun Yang, Ruiheng Zhang, Xiaoxu Li, Haodong Chang, Jiatong Li, Wanneng Wu, Lei Sang, Yunkun Yang, Sheng Wang and others that I cannot list them all for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the funs we have had.

I appreciate all the help from USTC alumni in Sydney, especially for Prof. Dong Xu, Prof. Dacheng Tao, Dr. Tongliang Liu, Huan Fu, Baosheng Yu, Zhe Chen and Yang He. Their help make me feel warm in Sydney as in my home.

I have met many great people during my 11 years University life. I would like to thank My undergraduate and master classmates, including Shuangwu Chen, Wei Wang, Lixiang Xu, Dao Xiang, Jiawei Jiang, Tianyi Zhao, Yu Liu, Chen Tang, Yue Li and Yicheng Zhang. I would give my special thanks to Prof. Qiang Ling, who is my master supervisor and Dr. Feng Li, who introduced me to UTS.

Most of all, I would like to thank my parents, for their unconditional support, both financially and emotionally throughout the whole PhD studying.

Finally, I would like to thank the China Scholarship Council (CSC) for funding my research.

List of Publications

This thesis is based on the following publications:

- **Chapter 4**

  Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, Ian Burnett, Multi-scale blocks based image emotion classification using multiple instance learning. *in* 'Proceedings of International Conference on Image Processing (**ICIP**)' (2016): 634-638

  Tianrong Rao, Min Xu, Huiying Liu, Generating affective maps for images, *in* 'Journal of Multimedia Tools and Applications' 77.13 (2018): 17247-17267.

- **Chapter 5**

  Tianrong Rao, Xiaoxu Li, Min Xu, Learning multi-level deep representations for image emotion classification, *submitted to* 'Neural Processing Letters' in October 30th 2018

  Xinge Zhu, Liang Li, Weigang Zhang, **Tianrong Rao**, Min Xu, Qingming Huang, Dong Xu, Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition, *in* 'Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)' (2017)

- **Chapter 6**

  Tianrong Rao, Xiaoxu Li, Haimin Zhang, Min Xu, Multi-level Region-based Convolutional Neural Network for Image Emotion Classification, *in* 'Journal of Neurocomputing' 333 (2019): 429-439

Other publication

- Huiying Liu, Min Xu, Jinqiao Wang, **Tianrong Rao**, Ian Burnett, Improving visual saliency computing with emotion intensity, *in* 'IEEE transactions on neural networks and learning systems' 27.6 (2016): 1201-1213

- Lingxiang Wu, Min Xu, Guibo Zhu, Jinqiao Wang, **Tianrong Rao**, Appearance features in Encoding Color Space for visual surveillance, *in* 'Journal of Neurocomputing' (2018)

Contents

List of Figures

List of Tables

Abbreviation

AIM - Attention based on Information Maximization

AUC - Area Under the Curve

BoVW - Bag of Visual Words

CBIR - Content Based Image Retrieval

CES - Categorical Emotion States

CNN - Convolutional Neural Network

DES - Dimensional Emotion Space

EMD - Earth Mover's Distance

FPN - Feature Pyramid Network

GCH - Color Histogram features for Global view

GBVS - Graph Based Vision Saliency

GRU - Gated Recurrent Unit

HSV - Hue, Saturation, Value

IoU - Intersection-over-Union

LCH - Color Histogram features for Local view

LMC - Linear Matrix Completion

LSTM - Long Short-Term Memory

MIL - Multiple Instance learning

MLP - Multi-Layer Perceptron

NLMC - Non-Linear Matrix Completion

pLSA - Probabilistic latent Semantic Analysis

R-CNN - Region-based CNN

RFA - Rate of Focused Attention

RNN - Recurrent Neural Network

SGD - Stochastic Gradient Descent

SIFT - Scale-Invariant Feature Transform

SLIC - Simple Linear Iterative Clustering

SUN - Saliency Using Natural statistics method

SVM - Support Vector Machine

WTA - Winner-Takes-All

Chapter 1

Introduction

This thesis by article deals with the problem of *image emotion classification* by exploring and developing various kinds of visual features. With the rapid development of mobile internet and digital devices such as mobile phones and digital cameras, a vast number of images are created and uploaded to the internet each day. Recording their daily life through images and sharing the images with their family, friends or even strangers via the internet has become an important social activity for more and more internet users. This also encourages the proliferation of many photo-based social networks, such as Instagram, Flickr *et al.* However, people encountered several issues when they want to manage the huge amount of images. First, internet users have their individual preference for the images on photo-based social networks. So, it is necessary for social networks to classify and select the images to suit the diverse demands of the users. Second, it is important to pick up the images that contain some uncomfortable contents such as bloody, violent and pornographic to protect the underage internet users. As a result, there is an urgent requirement and demand for intelligent image indexing and retrieval techniques. Currently, most real-life applications rely on textual information as the input basis for search and/or retrieval of relevant images. Unfortunately, the shear number of digital images being generated makes it impossible to manually label every image with textual information. Therefore, automatic analysis of image content has become an emerging research direction in recent years.

Content-based image retrieval (CBIR) systems support image search uses low-

level visual features such as colors, textures or shapes for image search to avoid the tedious job for manual tagging (Smeulders, Worring, Santini, Gupta and Jain, 2000). Human interprets an image through either digging its semantic meaning or understanding the image from an affective perspective. Recently, most existing research analyze image from a semantic perspective, such as research on scene detection (Itti, Koch and Niebur, 1998) and object detection (Guo, Zhao, Zhang and Chen, 2014). However, analyzing image at sentimental level is still away from humans' understanding of the image, due to the two major challenges, **Complexity** and **Subjectivity**, existing in image emotion analysis.

For complexity, image emotion is related to both the whole image from a global view (global feature) and the emotional regions in the image from a local view (local feature). In (Zhao et al., 2014*a*), the authors indicate that "the emotion recognition performance is not well for images whose emotions are dominated by emotional regions, which contain objects and concepts". For the images whose emotion are evoked by emotional regions, small changes may even evoke an opposite emotion, compared to the original image. Features extracted from such non-emotional regions may generate classification noise and the emotion difference cannot reliably be identified using the features extracted from the whole image. However, due to difficulty of collecting a well labeled dataset with emotional region, demonstrate the effectiveness of emotional regions and utilize local features extracted from these emotional regions is a challenging problem.

What's more, image emotion is related to complex visual features from high-level to low-level. Low-level visual features, such as color, shape, line, and texture, were first used to classify image emotions (Kang, 2003; Wang and He, 2008; Aronoff, 2006; Hanjalic, 2006*a*). Joshi *et al.* (Joshi, Datta, Fedorovskaya, Luong, Wang, Li and Luo, 2011) indicated that image emotion is highly related to image aesthetics for artistic works. Based on their study, mid-level features that represent image

aesthetics, such as composition, visual balance, and emphasis, are applied for image emotion classification (Machajdik and Hanbury, 2010*b*; Zhao, Gao, Jiang, Yao, Chua and Sun, 2014*b*). Machajdik and Hanbury suggested that image emotion can be significantly influenced by the semantic content of the image (Machajdik and Hanbury, 2010*b*). They combined high-level image semantics from the global view with Itten's art theory on the relevance of colors (Itten and Van Haagen, 1962) to recognize image emotion. However, most of the works focus on a specific level of features, a comprehensively consideration of combining different level of features for image emotion classification is still need to be investigate.

For subjectivity, people from different cultural background may have various emotional reactions to the same image. A viewer may also have various kinds of emotional response for an image. Therefore, it is unable to collect the hard emotional label of an image. Instead, emotion category is labeled with the probability is widely applied in affective image datasets. In most affective image datasets, the emotional labels of images are usually collected using the rule of majority voting, which means the emotion that has most votes is selected as the emotional label of the image. Compared the images in the datasets for image semantic analysis, which have certain labels, the uncertainty labels in affective image datasets clearly improve the difficulty to build an accurate classifier for image emotion classification.

Therefore, a comprehensive consideration of the complexity and subjectivity of image emotion is needed in the image emotion classification framework. To deal with the complexity, this thesis first explore the effectiveness of local emotional regions. To deal with the weakly labeled dataset, affective map is proposed to automatically explore the emotional regions in an image using the emotion label of the image. The effectiveness of the emotional regions extracted from the affective map can also be demonstrated through combining with saliency map for salinecy detection. Then, a new CNN framework that can fuse different level of deep features are proposed to

explore the impact of multi-level features for image emotion recognition. Finally, integrating different levels of features from both global and local view to further improve the performance of image emotion classification. For subjectivity, the emotion distribution can be estimated through the probability of emotional labels using the similarity of different emotion types and then help to reduce the impact bring from the uncertainty. Our final goal is to develop an emotion recognition framework that can comprehensively consider both the complexity and subjectivity existing in image emotion analysis.

Each of the presented articles in this thesis is a step towards the final goal of building a framework to analyze image emotion using visual features from high-level to low-level for both global and local views and consider the datasets with uncertainty emotional labels. The first article explores the effect of emotional regions on image emotion classification. The second article studies different levels of visual features and applies CNN with different level for image emotion analysis. In the last article, both different levels of visual features and emotional regions are integrated into one framework to predict emotional labels for images. In addition, emotion distribution estimation and multi-task learning are introduced to deal with the emotional label with probability of images caused by the subjectivity.

This chapter provides background on the basic concepts this work is based on. Followed by Chapter 2, which contains an additional literature review and Chapter 3, which presents the contributions of this thesis. The three aforementioned articles are then presented in Chapters 4, 5 and 7. Finally, Chapters 7 and 8 discuss the results of the presented work as a whole and draw conclusions.

Chapter 2

Literature Review

Chapters 4 to 6 each provide an overview of related work. This chapter focuses on more general context and includes a review of more recent work.

2.1 Affective Modeling and Classification Methods

2.1.1 Affective Modelling

To analyze emotions from a given image, there are two widely used models: categorical emotion states (CES) and dimensional emotion space (DES). In CES methods, computational results are mapped directly to one of a few basic categories (Irie, Satou, Kojima, Yamasaki and Aizawa, 2010)(Machajdik and Hanbury, 2010*a*)(Soleymani, Larson, Pun and Hanjalic, 2014)(Zhao et al., 2014*b*), such as *anger*, *excitement*, *sadness*, *etc.* DES methods, which contain the 3-D valence-arousal-control emotion space (VAD), 3-D natural-temporal-energetic connotative space, 3-D activity-weight-heat emotion factors, and 2-D valence-arousal (VA) emotion space, have also been widely adopted (Benini, Canini and Leonardi, 2011)(Hanjalic and Xu, 2005)(Tarvainen, Sjoberg, Westman, Laaksonen and Oittinen, 2014)(Xu, Luo and Jin, 2008)(Zhang, Huang, Jiang, Gao and Tian, 2010). VAD is the most popular in DES, where valence measures the pleasantness of an emotion ranging from happy to unhappy, arousal represents the intensity of an emotion evoked by a stimulus ranging from excited to peaceful and dominance express the degree of control exerted by a stimulus ranging from controlled to in control (Warriner, Kuperman and Brysbaert, 2013).

DES methods provide a more predict and flexible description for emotions, while CES methods in emotion classification is easier for users to understand and label. To compare our result with previous work (Machajdik and Hanbury, 2010*a*)(Zhao et al., 2014*a*), CES method has been adopted to classify emotions into eight categories, including positive emotion *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotion *Anger*, *Disgust*, *Fear*, *Sadness*, defined in a rigorous psychological study (Mikels et al., 2005*a*).

2.1.2 Classification Methods

Early research on image emotion mainly used classic machine learning methods. Solli and lenz (Solli and Lenz, 2009) simply used colour based bag-of-words to classify emotions. In order to map low-level features to high-level emotions, different machine learning methods were used for emotion classification. Kang proposed to detect the emotion using Hidden Markov Models (HMM) with low-level features. Recently, Support Vector Machine (SVM) had also been applied by many researchers to classify emotions for the good generalization ability on limited samples (Wei-ning, Ying-lin and Sheng-ming, 2006)(Yanulevskaya, Van Gemert, Roth, Herbold, Sebe and Geusebroek, 2008)(Zhao et al., 2014*a*). In (Machajdik and Hanbury, 2010*a*), Machajdik and Hanbury compared different classification methods, such as Naive Bayes, SVM and Random Forest and found Naive Bayes classifier achieved the best performance with their extracted features based on element-of-art. Due to the multiple distinct features extracted from images, multi-view learning was also considered to a feasible methods for emotion classification (Zhang, Gönen, Yang and Oja, 2015). Besides, considering the individual preference, Bianchi proposed the Kansei Distributed Information Management Environment (K-DIME) system, which built an individual model for each user using a neural network (Bianchi-Berthouze, 2003). Jia *et al.* (Jia, Wu, Wang, Hu, Cai and Tang, 2012) adopted a partially-labeled

factor graph model (PFG) using the social correlation between images to learn and predict image emotions.

Considering the recent success from CNN-based approaches in many computer vision tasks, such as image classification (Krizhevsky et al., 2012), image segmentation (Long, Shelhamer and Darrell, 2015), object detection (Ren, He, Girshick and Sun, 2015) and scene recognition (Zhou, Lapedriza, Xiao, Torralba and Oliva, 2014), CNN based methods have also been employed in image emotion analysis. Peng *et al.* (Peng, Chen, Sadovnik and Gallagher, 2015) first attempted to apply the CNN model in (Krizhevsky et al., 2012). They fine-tuned the pre-trained convolutional neural network on ImageNet (Deng, Dong, Socher, Li, Li and Fei-Fei, 2009) and demonstrated that the CNN model outperforms previous methods rely on different levels of handcrafted features on the Emotion6 dataset. Pang *et al.* utilize deep multi-modal learning to analyze and retrieval affective images (Pang, Zhu and Ngo, 2015). You *et al.* (You, Luo, Jin and Yang, 2016) employed a progressive strategy to train a CNN model to detect image emotion on the large-scale dataset of web images. In (You, Jin and Luo, 2017), local emotional regions extracted using attention model were considered for sentiment analysis. A deep coupled adjective and noun neural networks are presented to discover the shared features of the same adjective/noun in (Wang, Fu, Xu and Mei, 2016). These works usually borrow the popular CNN models that are used for image classification and object detection for image emotion classification. However, these widely used CNN models can not effectively classify the images whose emotion are mainly evoked by low-level and mid-level features, i.e. abstract paintings and art photos. Therefore, a new CNN model that can specifically deal with image emotion need to be designed.

2.2 Emotion Features Extraction

Image emotion classification has been studied for several years (Kang, 2003). Early researches focus on extracting hand-crafted features from the whole image, such as colors, textures, lines, and shapes and applying machine learning algorithms to classify emotions. Yanulevskaya et al.(Yanulevskaya et al., 2008) propose to categorize emotions based on Gabor and Wiccest features. The authors consider emotions to be one of a few basic categories, such as happiness, sadness, etc. and use machine learning to differentiate various emotion evoking categories. Wang Weining et al.(Wei-ning et al., 2006) introduce an algorithm based on histograms designed to express the emotional impact of image color and sharpness descriptors. Solli and Lenz(Solli and Lenz, 2009) also detect emotions using color-based emotion-histogram derived for patches surrounding each interest point. Besides, shape features have also been used for emotion estimation(Lu, Suryanarayan, Adams Jr, Li, Newman and Wang, 2012). In (Machajdik and Hanbury, 2010*a*), Machajdik and Hanbury propose to extract emotional features from an image based on theoretical and empirical concepts from psychology and art theories. Not only focusing on low-level features, e.g. color and texture, they also consider mid-level features, e.g. composition and even high-level features, e.g. human faces and skin. Jia et al.(Jia et al., 2012) propose an affective analysis method based on the correlation features for images from the social network. Recently, Zhao et al.(Zhao et al., 2014*a*) investigated the concept of principles-of-art-based emotion features, which are the unified combination of representation features derived from different principles, including *balance*, *emphasis*, *harmony*, *variety*, *gradation*, and *movement* and its influence on image emotions. The experiments demonstrated that the use of new features improves the results of image emotion classification. In the following section, the different features will be introduced which are widely used in affective image analysis methods in detail.

2.2.1 Low-level Visual Features

Low-level visual features, such as color, texture, line, shape etc, are widely used in affective image analysis.

Color

Colors can be (and often are) effectively used by artists to induce emotional effects. However, mapping low-level color features to emotions is a complex task which must consider theories about the use of colors, cognitive models and involve cultural and anthropological backgrounds (Colombo, Del Bimbo and Pala, 1999). In other words, people from different cultures or backgrounds might perceive and interpret the same color pattern quite differently. To effectively utilize color features in affective image analysis, colors which occur in an image are firstly needed to measure. There are different effective methods to measure colors using the following features:

Saturation and Brightness statistics are computed because saturation and brightness have direct influence on emotions, especially positive emotions, such as *Amusement*, *Awe*, *Contentment*, *Excitement*, etc. Valdez and Mehrabian (Valdez and Mehrabian, 1994) investigated human emotional reactions to different saturation and brightness with the valence-arousal-dominance emotion model. They conduct experiments in a controlled environment, where 250 people were shown a series of single color patches and rated them on a standardized emotional scale (Pleasure-Arousal-Dominance) to describe how they feel about the color. The results of the analysis show a significant relationship between the brightness and saturation of color and their emotional impact.

Colorfulness are also computed, as rich colors can easily evoke different kinds of emotion. It is measured using the Earth Mover's Distance (EMD) between the histogram of an image and the histogram having a uniform color distribution (Datta,

Joshi, Li and Wang, 2006).

Itten contrasts are a powerful concept in art theory (Itten and Van Haagen, 1962). Itten studied the usage of color in art extensively and formalized concepts for combining colors to induce an emotional effect in the observer and to achieve a harmonious image using itten contrasts. He identified different types of contrast: contrast of saturation, contrast of light and dark, contrast of hue and contrast of warm and cold. Contrast of light and dark can be measured using the standard deviation over the Brightness membership functions of all regions weighted by their relative size, and the contrast of saturation can be defined in an analogue method. The vector based measurement of the hue spread can be used to measure the contrast of hue. The contrast of warm and cold is defined in (Corridoni, Del Bimbo and Pala, 1999). For each region r_i, three membership functions $w_t, t = 1, 2, 3$, which express the degree of cold ($t = 1$), neutral ($t = 2$) and warm ($t = 3$). The contrast of warm and cold between two regions can be compute as: $\frac{\sum_{t=1}^{3} w_t(r_1) w_t(r_2)}{\sqrt{\sum_{t=1}^{3}(w_t(r_1))^2 \sum_{t=1}^{3}(w_t(r_2))^2}}$.

Texture

Texture is used to describe the surface quality of the objects in an image. It is important for the emotional expression of an image. Artists usually use different textures, such as smooth and rough, to achieve a desired expression of emotions. Some of the commonly used features, which have been developed to describe texture, is chosen for affective image analysis.

Wavelet-based features uses the Daubechies wavelet transform to measure spatial smoothness/graininess in images (Daubechies, 1992). In HSV color space, a three-level wavelet transform is performed on all three channels, Hue H, Saturation S and Brightness V. Denoting the coefficients in level i or the wavelet transform of one channel of an image as w_i^h, w_i^v and w_i^d $(i = 1, 2, 3)$, the wavelet features can be

compute as:

$$f_i = \frac{\sum_{x,y} w_i^h(x,y) + \sum_{x,y} w_i^v(x,y) + \sum_{x,y} w_i^d(x,y)}{|w_i^h| + |w_i^v| + |w_i^d|} \tag{2.1}$$

This is computed for every level i and every channel (H, S, and V) of the image.

Tamura features are efficient to measure texture. Three of the Tamura features, i.e. coarseness, contrast, and directionality, have been used for affective image retrieval (Wu, Zhou and Wang, 2005).

Line

There are mainly two types of lines, emphasizing lines and de-emphasizing lines. Emphasizing lines, better known as contour lines, show and outline the edges or contours of an object. When artists stress contours or outlines in their work, the pieces are usually described as lines. Not all artists emphasize lines in their works. Some even try to hide the outline of objects in their works. De-emphasizing lines are used to describe works that do not stress the contours or outlines of objects.

Lines can be used to suggest movement in some direction. They are also used in certain ways to give people different feelings (Machajdik and Hanbury, 2010*a*). For example, horizontal lines are associated with a static horizon and communicate calmness, peacefulness, and relaxation; vertical lines are clear and direct and communicate dignity and eternity; slant lines, on the other hand, are unstable and communicate dynamism. Lines with many different directions present chaos, confusion or action. The longer, thicker and more dominant the line the stronger the induced psychological effect. Usually, the amounts and lengths of static and dynamic lines are calculated by Hough transform to describe lines (Machajdik and Hanbury, 2010*a*).

Shape

Shape features are also proven to be efficient for emotion classification. In (Aronoff, 2006), researchers found that the geometric properties of a visual display, such as diagonal, straight and round, carried the meaning of emotions, like *anger* and *happy*. Bar *et al.* (Bar and Neta, 2006) proposed that roundness can affect different emotions. They found that curved contours usually linked to positive feelings, while sharp transitions usually led to negative feelings.

2.2.2 Mid-level Visual Features

Mid-level visual features are derived from low-level visual features, such as features based on principles-of-the art including balance, emphasis, harmony, variety, gradation, movement, rhythm, and proportion (Hobbs, Salome and Vieth, 1995) and aesthetics (Joshi et al., 2011). Mid-level visual features are more related to emotions. Different combinations of these features can evoke different emotions. In the following section, the mid-level features will be reviewed in detail.

Balance

Balance, in others words, symmetry, is used to express the feeling of equilibrium or stability of the image. Balance is always used to set the dynamics of a composition by artists. Balance contains three types: symmetrical, asymmetrical and radial. Symmetrical balance means the two halves of an image are identical or very similar. It is visually stable for most of the people. Symmetrical balance can be characterized by an exact or nearly exact compositional design on both sides of the horizontal, vertical or any axis of the picture plane. Opposite to symmetry, asymmetry is visually unstable balance, which uses compositional elements that are offset from each other. Asymmetrical balance is the more dynamic due to its more complex design construction. Radial balance refers to balance within a circular shape or

object, which is usually central symmetry and attracts attention at the center of the composition (Collingwood, 1938).

In affective image analysis, radial balance can be treated as a special part of symmetrical balance and the asymmetrical balance is opposite to symmetrical balance. Therefore the symmetrical balance is used to measure the balance of the image.

Symmetrical balance include bilateral symmetry, rotational symmetry and radial symmetry (Loy and Zelinsky, 2003; Loy and Eklundh, 2006). Symmetry detection method based on matching symmetrical pairs of feature points is used to detect bilateral and rotational symmetry (Loy and Zelinsky, 2003). In the method, the location, orientation, and scale of each feature can be represented by a point vector in x, y coordinates. Every pair of feature points can be treated as a potential candidate for a symmetrical pair. For bilateral symmetry, a potential perpendicular symmetrical axis passing the mid-points of the connecting line between each pair of matching points. Compared to bilateral symmetry detection, which requires to develop special feature descriptors for measurement, detecting rotational symmetry is more simple by matching the features against each other through the central point. for a given pair of non-parallel feature point vectors, there exists a central point that one feature vector can be rotated to precisely align with another feature vector. The Hough transform can be used to find dominant symmetry axes or centers (Ballard, 1981). Each potential symmetrical pair casts a vote in Hough space weighted by their symmetry magnitude. The bilateral symmetry magnitude needs to consider the difference between the two feature points of a symmetrical pair, while the rotational symmetry magnitude should be unified. Finally, the dominant symmetries present in the image are determined by the overall symmetries, which are related to the number of all individual pairs in a voting space. The result is blurred with a Gaussian and the maxima are identified as dominant axes of bilateral symmetry or centers of rotational symmetry.

Emphasis

Emphasis is used to emphasize the discrepancy of different contents in an image. It can be described as the sudden and abrupt changes in an image. Emphasis is usually used to detect the parts of images that can attract and focus most viewers' attention. Sun's rate of focused attention (RFA) can be used to measure emphasis (Sun, Yao, Ji and Liu, 2009).

RFA was proposed to measure the attention rate of viewers when watching an image. It is defined as the attention focus on some predefined aesthetic templates or some statistical distributions according to the image's saliency map. A 3-dimensional RFA vector can be computed as:

$$RFA(i) = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H} Saliency(x,y)Mask_i(x,y)}{\sum_{x=1}^{W}\sum_{y=1}^{H} Saliency(x,y)} \tag{2.2}$$

where W and H denote the width and height of image I. $Saliency(x,y)$ and $Mask_i(x,y)$ are the saliency value and mask value at pixel (x,y), respectively. $i = 1,2,3$ represents different aesthetic templates.

Harmony

Harmony, also called as unity, can be defined as a way of combining similar content (such as line, shape, color, texture) in an image to stress their similarities. Artists achieve harmonious in an image by using repetition and gradual changes for the content in the image. Pieces that are in harmony give the work a sense of completion and have an overall uniform appearance (Zhao et al., 2014*a*).

The harmony intensity of each pixel of the image can be computed on its hue and gradient direction in a neighborhood. The circular hue or gradient direction are equally divided into eight parts, which are separated into two adjacent groups $c = i_1, i_2, ..., i_k | 0 \leq i_j \leq 7, j = 1,2,...,k$ and $I \setminus c$, where $i_{k+1} \equiv i_{k+1} \pmod 8), I =$

$0, 1, ..., 7$. The harmony intensity at pixel x, y can be computed as:

$$H(x,y) = \min_{c} e^{-|h_m(c)-h_m(I\setminus c)|} |i_m c - i_m(I \setminus c)| \tag{2.3}$$

where

$$\begin{aligned} h_m(c) &= \max_{i\in c} h_i(c) \\ i_m(c) &= \arg\max_{i\in c} h_i(c) \end{aligned} \tag{2.4}$$

where $h_i(c)$ is the hue or gradient direction in groups c.The harmony intensity of the whole image is the sum of all pixels' harmony intensity:

$$H = \sum_{(x,y)} H(x,y) \tag{2.5}$$

Variety

Artists use different colors to evoke different emotions for viewers. Variety is used to create complicated images by combining different colors. However, harmony and variety are not black-or-white types. Without variety, the visual interest in an image could be lost and hard to evoke people's emotions. On the other hand, without harmony, the image could become very complex and make viewers confused. The pixel amount of 11 basic colors (black, blue, brown, green, gray, orange, pink, purple, red, white, yellow) present in an image can be computed as variety using the algorithm proposed by Weijer *et al.* (Van de Weijer, Schmid and Verbeek, 2007).

Gradation

Gradation indicates a way of combining low-level features by using a series of gradual changes. For example, gradation may be a gradual change from one color to another. Pixel-wise value *windowed total variation* and *windowed inherent variation*

and their combination can be used to measure gradation for each pixel in an image (Xu, Yan, Xia and Jia, 2012).

The *windowed total variation* for pixel $p(x, y)$ in image I is defined as:

$$D_x(p) = \sum_q g_{p,q}|(\partial_x I)_q|, D_y(p) = \sum_q g_{p,q}|(\partial_y I)_q| \tag{2.6}$$

where $q \in R(p)$, $R(p)$ is a rectangular region centered at p, $D_x(p)$ and $D_y(p)$ are windowed total variations in the x and y directions for pixel p, which count the absolute spatial difference within the window for pixel $R(q)$. $g(p, q)$ is a weighting function:

$$g_{p,q} = e^{-\frac{(x_p - x_q)^2 + (y_p - y_q)^2}{2\sigma^2}} \tag{2.7}$$

where σ controls the spatial scale of the window.

The *windowed inherent variation* for pixel $p(x, y)$ in image I is defined as

$$L_x(p) = |\sum_q g_{p,q}(\partial_x I)_q|, L_y(p) = |\sum_q g_{p,q}(\partial_y I)_q| \tag{2.8}$$

Other than $D_x(p)$ and $D_y(p)$, $L_x(p)$ and $L_y(p)$ evaluate the overall spatial variation, without incorporating modules.

The *relative total variation* (RTV), which can be used to measure an image's relative gradation, can be computed using *windowed total variation* and *windowed inherent variation*

$$RG = \sum_p RTV(p) = \sum_p (\frac{D_x(p)}{L_x(p) + \varepsilon} + \frac{D_y(p)}{L_y(p) + \varepsilon}) \tag{2.9}$$

Movement

Movement is used to produce the looking and feeling of action. Usually, the viewers' eyes move throughout the images following the guidance of movement. Movement is achieved through placement of different components of an image so that the eye follows a certain path, like the curve of a line, the contours of shapes, or the repetition of certain colors, textures, or shapes (Sun, Yao and Ji, 2012).

Based on Super Gaussian Component analysis, Sun *et al.* obtained a response map by filtering the original image and adopted the winner-takes-all (WTA) principle to select and locate the simulated fixation point and estimate a saliency map (Sun et al., 2012). The distribution of eye scan path can be obtained using Sun's method.

Aesthetics

Image aesthetics are highly related to image emotions. Early researchers have already discovered that people have a positive emotional response to a wider concept of beauty (Joshi et al., 2011). The aesthetics behaviors, such as certain objects and human behavior, are inherently expressive of some emotional states. Composition features, which are always used in image aesthetics assessment, have much potential in exploiting the spatial relations between different parts of images, which can evoke different emotional responses for humans (Mao, Dong, Huang and Zhan, 2014).

2.2.3 High-level Visual Features

High-level visual features are usually the semantic content in images, which can evoke the greatest emotional response for viewers. Figure 2.1 shows of the emotional impact of semantic content in images. The two images have no significant discrepancy in colors and textures. The warm tones of colors on both images bring viewers a pleasant feeling. Differentiating these two images through other visual features,

such as texture, shape, composition, and aesthetics, is also very difficult. However, for most viewers, they would have violent emotional reactions, *e.g. fear* and *awe*, at first glance for the image of brown bear, but peaceful feeling to the teddy bear, which looks quite content. The example clearly indicates the effectiveness of the semantic content of an image for image emotion analysis. However, bridging semantic content and image emotion is still very challenging, due the the vague relationship between them. For example, a knife on blood may evoke fear, while a knife on a birthday cake may evoke even opposite emotions like contentment. In fact, it is impossible to precisely model the intricate relationship between semantic content and image emotion. Existing methods mainly utilize several specific semantic content to further improve the emotion classification accuracy.

Traditional image semantic analysis utilize some sophisticated designed high-level handcrafted features, like SIFT and GIST. However, compered to the deep learning features for image semantic analysis, which are very popular in recent years, transitional high-level handcrafted features are just valid for some specific images and cannot achieve acceptable performance in large scale images. Combining low-level and mid-level features with high-level deep learning features is a good way to improve the result of image emotion classification. Therefore, the thesis will introduce deep neural network in next section. Considering the analysis of the semantic content of images is also a big research area and will go beyond the scope of this work, only one widely used high-level image semantics, facial expression, for image emotion analysis will be introduced in this section.

Facial expression

Human face in an image always draws viewers major attention. The facial expression on human face can significantly evoke observers' emotion. Therefore, recognizing the emotional facial expressions on human faces in images can effectively

Figure 2.1 : Sample images show the impact of content for image emotion. The two images are formally similar, but have totally opposite emotional impact.

support to improve the performance of image emotion analysis. However, efficient algorithms that can precisely recognize the emotional expression of every human face in an image are still not yet fully mature (Lopes, de Aguiar, De Souza and Oliveira-Santos, 2017). Even so, the facial expressions on prominent faces (if there are any) in the images can still be detected and utilized to improve the image emotion classification results.

2.3 Convolutional Neural Networks

Convolutional Neural Network (CNN) (LeCun, Bottou, Bengio and Haffner, 1998) is a type of neural networks that has achieved outstanding breakthroughs in many computer vision tasks (Krizhevsky et al., 2012; He, Zhang, Ren and Sun, 2016). CNN is built based on the observation that the cells in the visual cortex only respond to stimuli in sub-regions of the visual field (Hubel and Wiesel, 1962). The whole

visual consists of these sub-regions, each of which are responded to a specific cell detecting a particular type of visual stimulus. Follow this observation, the convolutional layers in CNN use filters with shared weights to detect the same local pattern all over the input image. Therefore, compared to traditional Neural Networks with fully connections in every layer, like Multi-Layer Perceptron (MLP), CNN usually has sparse connections. A one-dimensional convolution can be mathematically defined as:

$$s(t) = (x * w)(t) = \int x(\tau)w(t-\tau)d\tau \tag{2.10}$$

in which x is the input, w is a filter and s is the output. t can be defined as time for temporal inputs. For images, they can be treated as discrete two-dimensional inputs representing pixel value in a spatial coordinate. Therefore, for images, convolution can be formulated as:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{2.11}$$

in which, I is the input image, K is the two-dimensional filter and S is the output feature map. Usually, the region of non-zero elements of the feature map is much smaller than the size of the input image. The filter is slid across the input image and the responses of each location are computed respectively. For the image that has multiple channels, the filter is applied per channel and final results are the sum of all channels.

In a CNN, the convolutional layer contains various filters which are applied separately to the input. The number of the output feature maps corresponds to the number of filters in the different convolutional layer. Figure 2.2 shows the convolutional layer with multi-channel input on the left. The output feature maps can be treated as multi-channel inputs and fed into another convolutional layer.

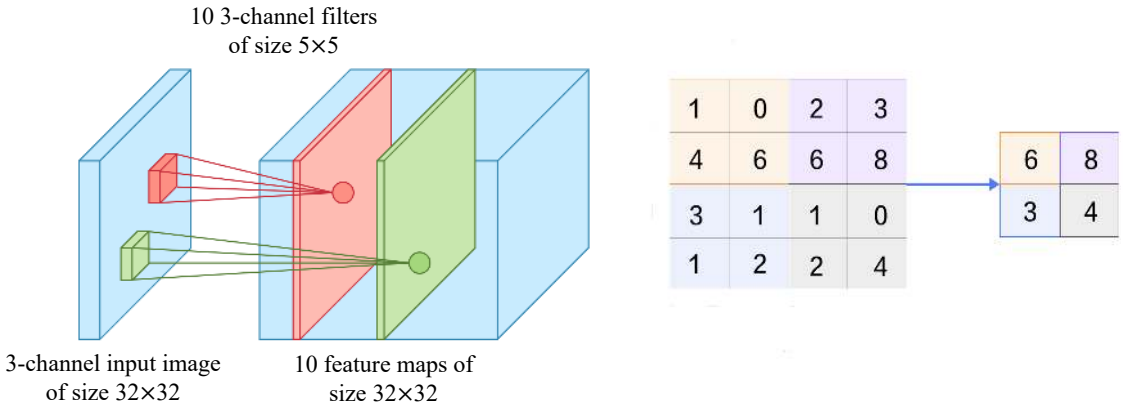

Figure 2.2 : Left: A convolutional layer that takes a three-channel (RGB) image as input and applies a filter bank of size $10 \times 3 \times 5 \times 5$ yielding 10 feature maps of size 32×32. Right: Two-by-two maxpooling with non-overlapping pooling regions.

Therefore, a CNN that contains multiple convolutional layers can be built in this way.

Commonly, some convolutional layers in a CNN are followed by a pooling layer. Pooling layers summarize local regions of a feature map by replacing them with statistics such as the average or maximum value of that region. Figure 2.2 right shows an example of max-pooling operation. Obviously, the size of the resulting feature map can become smaller after a pooling layer. What's more, pooling operation adds *invariance* to small translations which benefit to recognition tasks. However, applying pooling operation will lose location information and decrease the resolution of the input image, which will reduce the performance of detection tasks.

2.3.1 Region-based CNN

Region-based CNN (R-CNN) (Girshick, Donahue, Darrell and Malik, 2014) is used to discover specific local regions in an image, which generates region proposals on CNN framework to localize and classify objects in images. Then, by introducing supervised pre-training for an auxiliary and domain-specific fine-tuning, the

object detection performance is significantly improved (Girshick, Donahue, Darrell and Malik, 2016). Girshick (Girshick, 2015) further develops the R-CNN model to faster-RCNN model to reduce the training time and computing consumption while improving the object detection accuracy. Ren *et al.* (Ren et al., 2015) combine the Region Proposal Network(RPN) with CNN architecture to share full-image convolutional features and predict object bounds and objectness scores simultaneously.

2.4 Affective Image Datasets

The hand-crafted visual features have only been proven to be effective on several small datasets, whose images are selected from a few specific domains.

Artphoto and **Abstract** datasets are proposed in (Machajdik and Hanbury, 2010*a*). In the ArtPhoto dataset, 806 photos are selected from some art sharing sites by using the names of emotion categories as the search terms. The artists, who take the photos and upload them to the websites, determine emotion categories of the photos. The artists try to evoke a certain emotion for the viewers of the photo through the conscious manipulation of the emotional objects, lighting, colors, etc. In the Abstract dataset, each image is assigned to one of the eight aforementioned emotion categories. This dataset consists of 228 abstract paintings. , the images in the Abstract dataset represent the emotions through overall color and texture, instead of some emotional objects. In this dataset, each painting was voted by 14 different people to decide its emotion category. The emotion category with the most votes was selected as the emotion category of that image. These two dataset are usually used to detect the relationships between image emotion and low-level/mid-level visual features.

IAPS-Subset is proposed in (Mikels, Fredrickson, Larkin, Lindberg, Maglio and Reuter-Lorenz, 2005*b*). The *International Affective Picture System* (IAPS) is a standard stimulus image set which has been widely used in affective image classification.

IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, landscapes and others (Lang, Bradley and Cuthbert, 2008*a*). Among all IAPS images, Mikels *et al.* selected 395 images and mapped arousal and valence values of these images to the above mentioned eight discrete emotion categories. EmotionROI is an emotion prediction benchmark proposed by Peng *et al.* (Peng, Sadovnik, Gallagher and Chen, 2016). 1,980 images are collected from Flickr dataset with emotion categories. 15 AMT workers are employed to label the emotional regions within the images that can evoke emotions. Therefore, the ground truth by assuming the influence of each pixel on evoked emotions is proportional to the number of drawn rectangles covering that pixel. The IAPS-Subset reveals the relationships between images senmaitcs and image emotion.

Emotion6 dataset (Peng et al., 2016) contains of 1,980 images collected from Flickr by using the emotion keywords and synonyms as search terms. For each emotion category, there are 330 images. AMT workers were invited to label the emotional regions that can evoke the Ekman's 6 emotions. Each image was labeled by 15 subjects. This dataset is used to detected the emotional regions in affective images.

Considering the limitations of these small-scale datasets, You *et al.* proposed a large scale dataset (You et al., 2016). To collect this dataset, 90,000 noisy labeled images are first downloaded from Instagram and Flickr by using the names of emotion categories as the keywords for searching. Then, the downloaded images were submitted to Amazon Mechanical Turk (AMT) for further labeling. Finally, 23,308 well-labeled images were collected for emotion recognition. The large scale dataset contains many different typs of affective images, which can be used to evaluate the generalization ability of the image emotion classification methods.

Flickr dataset is also a large scale dataset which consists of 301,903 images based on the Ekman six emotion model (Yang, Jia, Zhang, Wu, Chen, Li, Xing and Tang, 2014). Each of the six emotion is linked to a word list, which is manually defined on WordNet and HowNet The emotion category whose word list has the most same words as the adjective words of an images tags and comments is assigned to the image. However, compared to manual labeled dataset, the Flickr dataset is just a weakly labeled dataset and contains more noisy labels.

To solve the problem of subjectivity existing in emotional labels, two large scale image datasets for descrete emotion distribution are established (Yang, Sun and Sun, 2017). One is **FlickrLDL** dataset, which contains 10,700 images downloaded from Flickr, and are labeled by 11 viewers using Mikels emotion model. Another is **TwitterLDL**, which contains 10,045 images labeled by 8 viewers after collecting by searching various sentiment key words from Twitter and duplication removal. In both datasets, the ground truth emotion distribution for each image is obtained by integrating the votes from the workers. The discrete emotion distribution reveals the different emotion response of people towards one affective image, and can use to explore the imapct of subjectivity on image emotion.

Chapter 3

Overview of Research Goals and Contributions

Early research for image emotion analysis mainly relies on designing specific hand-crafted visual features extracted from whole images that only available for several kinds of affective images. Recently, computer vision research with deep learning is developing rapidly. Deep learning methods have yielded the state-of-the-art performance on many computer vision tasks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016). Benefiting from these breakthroughs, researchers began to apply deep learning for Image emotion recognition. Compared to hand-crafted visual features, deep learning can be used to learning representations of images automatically. Features learned by CNN have been shown to generalize well to large-scale image data, which contains various kinds of affective images. However, considering the complexity and subjectivity existing in image emotion analysis compared to other semantic level computer vision tasks, traditional single column CNN structure cannot achieve state-of-the-art performance in this task (Peng et al., 2016).

The research objective is to work towards a deep learning framework that can integrate different level of features from both global and local view for image emotion recognition. The work contributes to this endeavor by studying visual features and learning approaches that can be useful for image emotion analysis.

Three step is needed to achieve the research objective. First, considering collecting a labeled large scale dataset for emotional regions is almost an impossible work, due to the complexity and subjectivity of the affective images, demonstrate the effectiveness of emotional regions for image emotion analysis is a key problem

for our work. The first part of the work focuses on studying the impact of the local emotional region for image emotion classification. Second, though achieving huge success in image semantic analysis, existing deep neural networks lack the ability for utilizing low-level and mid-level features. However, existing research has revealed that image emotion can be evoked through different level of visual features. Therefore, developing a CNN based framework that can utilize the different level of deep representations is also important for the work. Finally, a deep learning framework hat can integrate different level of features from both global and local view for image emotion recognition can be designed based on the former works. Comprehensively considering these complex features helps to achieve high performance on image emotion classification and resolve the problem of subjectivity in affective images.

The initial study in Chapter 4 is based on the conference paper "Multi-scale blocks based image emotion classification using multiple instance learning" (Rao, Xu, Liu, Wang and Burnett, 2016) and journal papers "Generating affective maps for images" (Rao, Xu and Liu, 2017). It proposes to generate an 'affective map', with which image emotion analysis can be performed at a region-level. The purpose of affective map is to represent emotion intensity of each pixel (the probability of a pixel belonging to an emotion type) by a scalar quantity and to guide the selections of emotional regions. With affective maps, we can easily identify the emotional regions within an image. The probability of an image belonging to a certain emotion category can be further computed through the affective map by considering the impact of emotional regions for image emotion.

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

- Through generating an affective map, emotions of an image can be identified at a pixel-level. With pixel-level details, both the dominant emotion conveyed

by a whole image and region-based emotions can be explored.

- For an image, our method can generate multiple affective maps in regarding different emotion categories. This ensures the consistency with human perception on an image, which considers the subjectivity existing in image emotion.
- Affective maps enlarge the application field of affective image analysis and enable many popular applications, such as image emotion classification an visual saliency detection.
- Experimental results demonstrate the effectiveness of emotional regions from local view for image emotion analysis

The second study is based on "Learning multi-level deep representations for image emotion classification" (Rao, Xu and Xu, 2016) and "Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition" (Zhu, Li, Zhang, Rao, Xu, Huang and Xu, 2017) presents deep learning approaches for image emotion recognition. Based on the large-scale dataset (You et al., 2016), deep learning based framework shows its excellent performance and generalization on different kinds of affective images. In the article, we propose a new CNN based method that combines different levels of deep representations, such as image semantics, image aesthetics and low-level visual features, from both global and local views for image emotion classification. By combining different levels of deep representations, our method can effectively extract emotional information from images. Experimental results demonstrate that our method outperforms the state-of-the-art methods using deep features or hand-crafted features on both Internet images and abstract paintings.

The third article "Multi-level Region-based Convolutional Neural Network for Image Emotion Classification" focuses on integrating the different level of deep representations from both local and global view for image emotion classification. The contributions of this article are summarized as follows:

- We employ a feature pyramid network(FPN) to extract multi-scale deep feature maps that related to image emotion. The multi-scale deep feature maps extracted from different convolutional layers can combine high-level semantic features with low-level deep features, and thus significantly improve the performance of emotion region detection.

- We build a region-based CNN model that can effectively extract local emotional information from the emotional regions of the image. Ignoring the noisy information generating from non-emotional regions can significantly improve the emotion classification performance.

- Image emotion labeling is a highly subjective task and the uncertain emotion labels will degrade the classification accuracy. Thus, we modify the loss function to consider the emotion class probability, rather than a hard class label, into image emotion classification to overcome the subjectivity in emotion analysis.

Chapter 4

Region-based Affective Image Analysis

Previous works mainly focus on analysis image emotion from a global view, while ignore the local information that can help image emotion classification. Due to the vogue definition of emotion, it is impossible to collect local emotion regions with exact boundaries. Therefore, demonstrating the effectiveness of local emotion information for image emotion analysis and applying it becomes a hard task.

Previous research demonstrates that visual attention can be driven by emotions, which reveals that image emotion can be useful and can be investigated for visual attention related applications (Liu, Xu, Wang, Rao and Burnett, 2016).

Existing methods analyze image emotion at an image-level (Machajdik and Hanbury, 2010*a*; Zhao et al., 2014*a*), without considering the difference between different regions within an image, which places problems/challenges in affective image analysis:

- Analyzing image at an image-level might reduce the accuracy of affective image classification since different regions within an image can evoke different kinds of emotions and same features may evoke different kinds of emotions in different regions. These emotions may be different from or even be opposite to the dominant emotion of the image.

- Analyzing image emotion at an image-level is unable to provide detailed information about the regions within the image. This narrows the application field of affective image analysis since many applications, such as saliency detection

and object detection, need to access pixel-level information.

Dealing with the challenges mentioned above, this work proposes to generate an 'affective map', with which image analysis can be performed at a region-level. The purpose of affective map is to represent emotion intensity of each pixel (the probability of a pixel belonging to an emotion type) by a scalar quantity and to guide the selections of emotional regions. Fig. 4.1 shows an image example and its affective maps. Eight different affective maps are generated for the image in regards eight different emotion categories. With affective maps, the emotional regions within an image can be easily identified. The probability of an image belonging to a certain emotion category can be further computed through the affective map by considering the impact of emotional regions for image emotion.

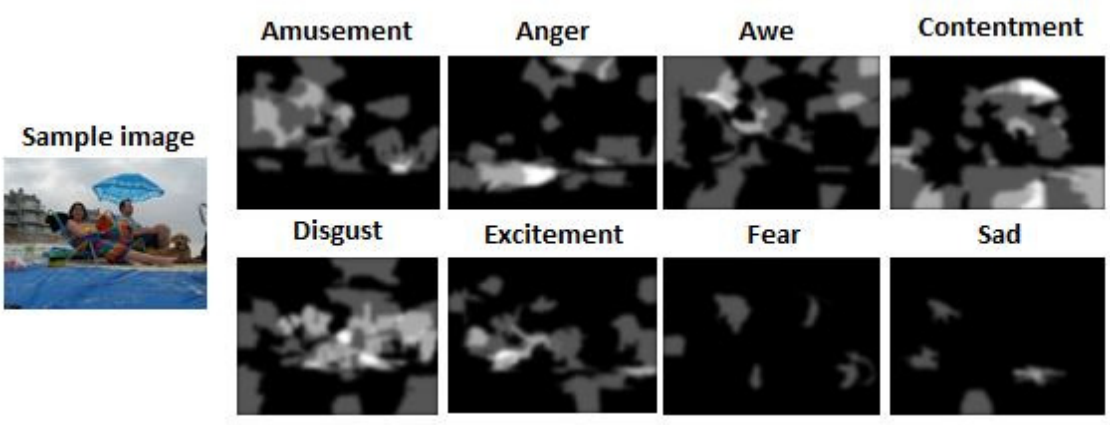

Figure 4.1 : Affective maps for different emotion categories. In this research, eight basic emotion categories which are defined in (Mikels et al., 2005a) is applied

In order to demonstate the effectiveness of the proposed affective map, in this research, affective map is applied for two important applications: affective image classification and visual saliency computing.

The major contributions of this work are summarized as follows:

- Through generating an affective map, emotions of an image can be identified at a pixel-level. With pixel-level details, both the dominant emotion conveyed by a whole image and region-based emotions can be explored.

- To people having different cultural backgrounds, an image is able to evoke different types of emotions (Zhang et al., 2015). For an image, the proposed method can generate multiple affective maps in regarding different emotion categories. This ensures the consistency with human perception on an image, compared to the existing methods of affective image analysis.

- Affective maps enlarge the application field of affective image analysis and enable many popular applications.

    - As one of the applications of affective map, affective image classification is tested on three popular datasets. Experimental results demonstrate that analyzing image at region-level significantly improve classification results.

    - This research also experimentally proves that affective map can be easily combined with traditional saliency detection methods. Compared with previous research (Liu, Xu, Wang, Rao and Burnett, 2016), this work integrates Simple Linear Iterative Clustering (SLIC) to further improve the performance of saliency detection.

4.1 Affective Map Generation

Estimating emotion intensity at a region-level faces some difficulties. First, within an image, there exist different emotional regions of different size and different shape. It is unable to cover different emotional regions using image blocks of the same size. Second, there exists an "affective gap", which can be defined as "the lack of coincidence between the measurable signal properties, commonly referred to

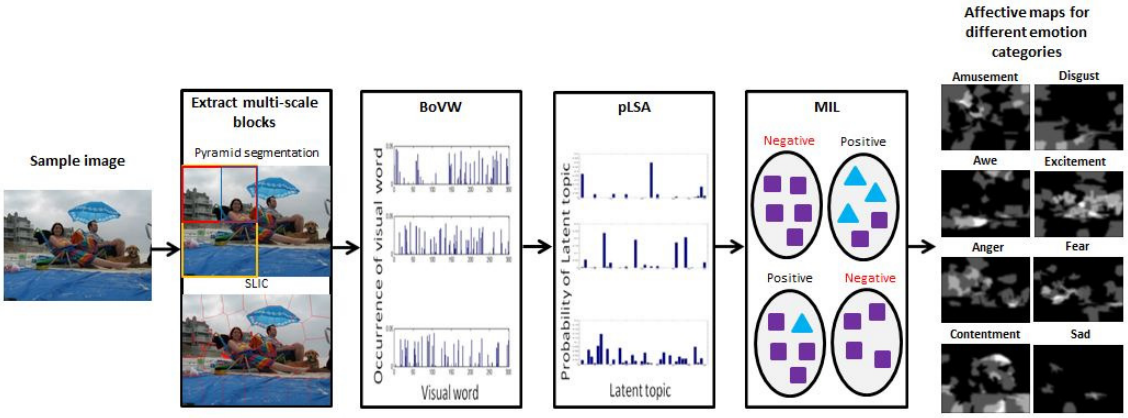

Figure 4.2 : An overview of the proposed method. Blocks of the image at multiple scales is firstly extracted. Each block is represented with the BoVW method. Then pLSA is employed to estimate the topic distribution of each block. Finally, MIL is performed to learn an emotion classifier.

as features, and the expected affective state in which the user will be found following perception of the image " (Hanjalic, 2006*b*) in affective image analysis. In other words, there is no one-to-one match between low-level features and high-level emotions. Finally, datasets, which are wildly used for affective image analysis, are weakly labeled, meaning that the images in these datasets are labeled with emotion category for a whole image instead of labeled with the boundaries of emotional regions within the image. This makes collection of a training set of reliable emotional regions difficult.

To address the issues mentioned above, an affective map generation method shown in Fig. 4.2 is proposed. To generate an affective map, multi-scale blocks are firstly extracted to cover different emotional regions in an image. Then, in order to bridge the gap between low-level features and high-level emotions, a mid-level representation is introduced, exploiting Probabilistic Latent Semantic Analysis (pLSA) to learn a set of mid-level representations as a set of latent topics from

affective images. Finally, Multiple Instance Learning (MIL) is employed, based on the multi-scale blocks extracted from an image, to reduce the need for exact labeling and analyze the affective images at region-level.

In this section, the major steps of affective map generation are introduced in detail.

4.1.1 Multi-scale Block Extraction

Due to the different size of emotional regions, multi-scale blocks are extracted in an image to cover them. In order to precisely segment image blocks to cover emotional regions, pyramid segmentation (Antonisse, 1982) used in previous work (Liu, Xu, Wang, Rao and Burnett, 2016) is replaced with Simple Linear Iterative Clustering (SLIC) (Achanta, Shaji, Smith, Lucchi, Fua and Susstrunk, 2012).

Pyramid segmentation is a simple but widely used image segmentation method. With pyramid segmentation, the number of blocks required to cover emotional regions can be minimized. In the proposed method, a five-level pyramid segmentation is used to extract multi-scale blocks as this avoids the extracted blocks being too small to reasonably cover emotional regions. However, the image blocks extracted using pyramid segmentation are all rectangles without considering the shape difference between emotional regions.

Compared to Pyramid segmentation, SLIC is very popular for many computer vision tasks in recent years. SLIC segments images by clustering pixels based on their color similarity and distance in CIELAB color space. Compared to other state-of-the-art methods, the image blocks extracted using SLIC are approximately equally sized, which are easily collected for each level and the exact number of image blocks can be fixed before. The image blocks extracted through SLIC adhere to the boundaries of emotional regions better than the image blocks extracted through pyramid segmentation and cover the emotional regions more precisely. To get multi-

scale blocks for affective image classification, multi-level SLIC is applied. Since the segmentation result is not very good for SLIC , the number of image blocks is less than 20, an image is divided into 20, 50, 100 and 200 blocks for 4 level.

Experiments have been set to compare the performance of the two image segmentation methods.

4.1.2 BoVW Description

In this work, the emotion type of an image is classified based on multi-scale blocks extracted from the image. SIFT features are used as a basic feature and a Bag-of-Visual-Words (BoVW) approach is adopted to represent image blocks (Fei-Fei and Perona, 2005). SIFT feature is one of the most commonly used features for computer vision, which is based on the detection and description of locally interesting regions (Siersdorfer, Minack, Deng and Hare, 2010). The k-means clustering algorithm is used to find the cluster of SIFT descriptors. The centroid of the clusters can be used as visual words to represent the image blocks. In the BoVW approach, let $\mathbf{W} = \{w_j\}, j = 1, 2, ..., M$ be a set of visual words, where M is the total number of visual words. Then, for a set of image blocks $\mathbf{I} = \{I_i\}, i = 1, 2, ..., N$, the i^{th} image block can be represented as $I_i = \{x_{ij}\}; i = 1, 2, ..., N, j = 1, 2, ..., M$, where x_j indicates the occurrence of word w_j, N is the total number of the image blocks.

4.1.3 pLSA Representation

For each image block, latent topics using pLSA is further extracted, which has been verified to be effective for image scene classification (Bosch, Zisserman and Muñoz, 2006)(Fei-Fei and Perona, 2005). The pLSA model associates an unobserved class variable (latent topic) $\mathbf{Z} = \{z_k\}, k = 1, 2, ..., K$ with the visual words $\mathbf{W}$ and image blocks $\mathbf{I}$, where K is the number of latent topics. The probability of the

occurrence of the visual word w_j in a particular image block I_i, is

$$P\left(I_i, w_j\right) = P\left(I_i\right) P\left(w_j|I_i\right) = P\left(I_i\right) \sum_{k=1}^{K} P\left(w_j|z_k\right) P\left(z_k|I_i\right) \tag{4.1}$$

the log likelihood is,

$$\begin{aligned} \log L &= \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} \log P(I_i, w_j) \\ &= \sum_{i=1}^{N} \|\mathbf{x}_i\|_1 \log P(I_i) + \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} \log \sum_{k=1}^{K} P(w_j|z_k) P(z_k|I_i) \end{aligned} \tag{4.2}$$

Here, $\|\mathbf{x}_i\|_1$ is the L1-norm of $\mathbf{x}_i$, meaning the total number of visual words in block I_i. To get $P(w_j|z_k)$ and $P(z_k|I_i)$, Expectation-Maximization is employed to maximize the likelihood.

$$\begin{aligned} L\left(\theta\right) &= \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} \log \sum_{k=1}^{K} P(w_j|z_k) P(z_k|I_i) \\ &\geq \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} \sum_{k=1}^{K} Q\left(z_k|I_i, w_j\right) \log P\left(w_j|z_k\right) P\left(z_k|I_i\right) \end{aligned} \tag{4.3}$$

In (3), $Q\left(z_k|I_i, w_j\right)$ is the posterior of $\mathbf{z}$ given the parameter (I, w).

In E-step:

$$Q\left(z_k|I_i, w_j\right) = \frac{P\left(w_j|z_k\right) P\left(z_k|I_i\right)}{\sum_{k=1}^{K} P\left(w_j|z_k\right) P\left(z_k|I_i\right)} \tag{4.4}$$

In M-step:

$$\begin{aligned} &E\left(x_{ij} \log P\left(w_j|z_k\right) P\left(z_k|I_i\right)\right) = \\ &\sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} \sum_{k=1}^{K} Q\left(z_k|I_i, w_j\right) \log P\left(w_j|z_k\right) P\left(z_k|I_i\right) \end{aligned} \tag{4.5}$$

Then $P(w_j|z_k)$ can be get:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} x_{ij} Q\left(z_k|I_i, w_j\right)}{\sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij} Q\left(z_k|I_i, w_j\right)} \tag{4.6}$$

and $P(z_k|I_i)$:

$$P(z_k|I_i) = \frac{\sum_{j=1}^{M} x_{ij} Q\left(z_k|I_i, w_j\right)}{\sum_{j=1}^{M}\sum_{k=1}^{K} x_{ij} Q\left(z_k|I_i, w_j\right)} \tag{4.7}$$

Using (6) and (7), the topic distribution of each block of the image can be estimated.

4.1.4 MIL Estimation

In many cases, the emotion conveyed by an image is primarily contained in a block rather than the whole image. However, the available labelling information is only for the whole image not for any image block. Labelling each individual block is an extremely difficult task. MIL is a powerful method to deal with this kind of weakly labeled data, as MIL allows the label of some instances in a bag to be different with the label of the bag (Maron and Ratan, 1998). For example, in the proposed method, the block in an image may have a different emotion category with the image.

Thus, the q^{th} image $X_q = \{I_{1q}, I_{2q}, ..., I_{Nq}\}$ is defined as a bag of N instances, in which the instances $I_{1q}, I_{2q}, ..., I_{iq}$ are the blocks extracted from X_q. The bag label $y_q \in \{0, 1\}$ indicates if the bag is positive ($y_q = 1$) or negative ($y_q = 0$), i.e., if the bag belongs to a certain emotion category or not. The bag label is decided by the instance labels $y_{1q}, y_{2q}, ..., y_{Nq}$:

$$y_q = \max\{y_{1q}, y_{2q}, ..., y_{Nq}\} \tag{4.8}$$

In other words, the bag label is considered positive if the bag contains at least one positive instance. The instance label, which is not known during training, indicates if the instance belongs to a certain emotion category or not. Using Bayes' rule, the log likelihood is maximized between bags and bag labels:

$$\log L = \sum_{q} \big(\log P(y_q|X_q) \big) \tag{4.9}$$

Since the instance labels are unknown during training, the log likelihood is defined for bags, not instances. So, $P(y_q|X_q)$, the probability of a bag being positive needs to be expressed in terms of its instances. the Noisy-OR (NOR) model is used as follows:

$$P(y_q|X_q) = 1 - \prod_{i=1}^{N} \big(1 - P(y_{iq}|I_{iq})\big) \tag{4.10}$$

From the above equation, it is easily to obverse that the probability of a bag being positive will be high if any instance in the bag has a high probability. The instances probability $P(y_q|I_{iq})$ can be calculated using Bayesian prediction. An iterative method is adapted to train the classifiers. At the initializing stage, all instances are classified by a naive Bayes classifier and positive instances are selected. Then, the selected instances are classified by a naive Bayes classifier with the classification result in the last step as a prior and positive instance are chosen again. This process is repeated until no change happens when choosing the positive samples or the likelihood between bags and bag labels doesn't increase.

4.1.5 Affective Map Generation

By using the trained classifiers, the probability of an image block extracted in section 4.1.1 belonging to a particular emotion category can be get. Using the result, the probability of a pixel belonging to the emotion category can be calculate. For

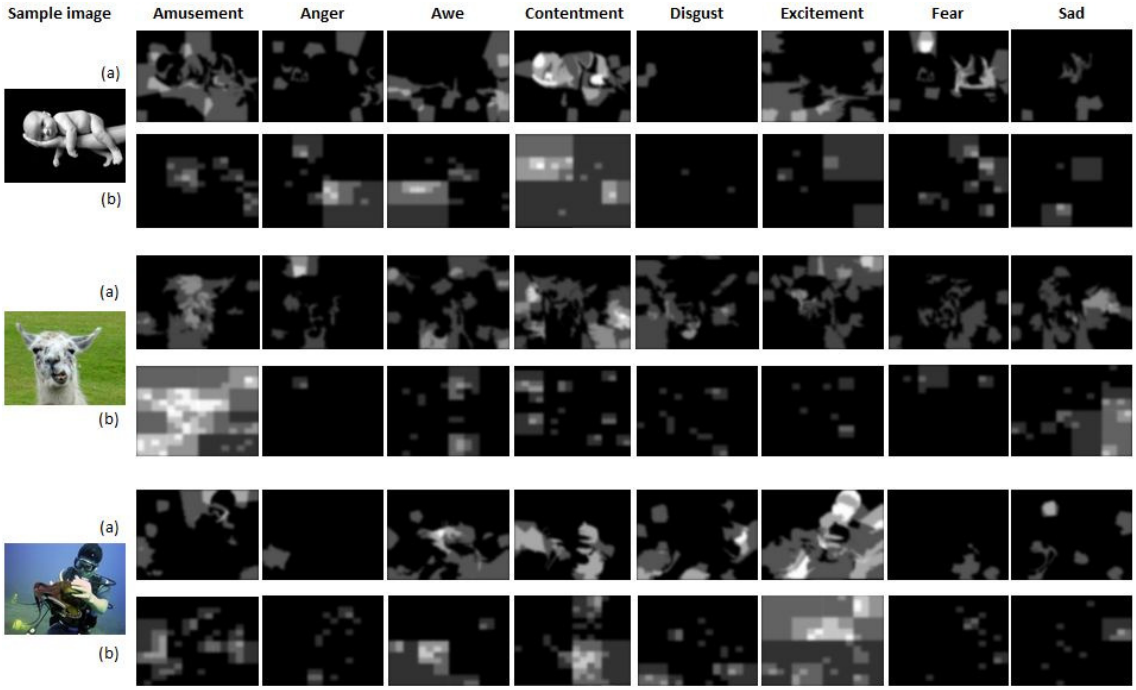

Figure 4.3 : Samples of affective maps based on SLIC (a) and affective maps based on pyramid segmentation (b) for 8 emotion categories.

each pixel in the image, the probability of belonging to the emotion category is the average of all the image blocks containing this pixel.

$$P(i,j,c) = \frac{\sum_{(i,j)\in I_{iq}} P(I_{iq},c)}{\sum_{(i,j)\in I_{iq}} 1} \tag{4.11}$$

Here (i,j) is the pixel in the image, c is the emotion category, I_{iq} is the block extracted from the image and P is the affective map for emotion category c. In this paper, images are classified into eight emotional categories, i.e. *Amusement*, *Anger*, *Awe*, *Contentment*, *Disgust*, *Excitement*, *Fear* and *Sad*.

Fig. 4.3 shows samples of affective map based on SLIC and affective map based on pyramid segmentation for eight emotion categories.

4.2 Affective Image Classification

Through affective map, the probabilities of an image belong to an emotion category can be calculated.

$$P(c) = \frac{\sum_i P(I_{iq}, c)}{\sum_i 1} \tag{4.12}$$

Here, $P(c)$ is the probability of the image belonging to an emotion category c.

The emotion category that has the highest probability is decided as the emotion category of the image.

4.2.1 Experimental Setup

Due to the unbalanced number of images per emotion category in table 4.3, the evaluation method needs to be considered carefully in order to obtain valid results. In the experiments, a 'one category against all' classification strategy is ued (Machajdik and Hanbury, 2010*a*). For a certain category, a classifier is trained and later used to determine whether an image sample is belonging to that category or not. The image samples of each category are separated into a training set and a test set using K-fold Cross Validation (K=5). In order to compare results over an unbalanced data distribution, *true positive rate per class* averaged over the positive and negative classes is calculated instead of the correct rate over all samples, which is similar to the one used in (Machajdik and Hanbury, 2010*a*)(Zhao et al., 2014*a*). The experiments are run for each category 10 times and apply the average *true positive rate* for a comparison. the proposed method is compared with two state-of-the-art methods:

- Machajdik *et al.* (Machajdik and Hanbury, 2010*a*). This method use the low-level visual features and mid-level features inspired by psychology and art theory which are extracted from the whole image for emotion classification.

- Zhao *et al.*(Zhao et al., 2014*a*). This method use principles-of-art-based emotion features extracted from the whole image, which are the unified combination of representation features derived from different principles, including *balance*, *emphasis*, *harmony*, *variety*, *gradation*, and *movement* for emotion classification.

4.2.2 Datasets

Three popular datasets are used to evaluate the performance of these methods.

IAPS dataset. The *International Affective Picture System* (IAPS) is a standard stimulus image set which has been widely used in affective image classification. IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, landscapes, scenes of poverty, pollution, mutilation, illness, accidents, insects, snakes, attack scenes and others (Lang, Bradley and Cuthbert, 2008*b*). Among all IAPS images, Mikels *et al.* (Mikels et al., 2005*a*) selected 395 images and mapped calculated arousal and valence values of these images to the above mentioned eight discrete emotion categories. As a result, an image might be able to link to more than one emotion categories since some emotion categories might have overlapping regions in the arousal-valence space. Table 4.1 and 4.2 show the percentage of images with multiple emotion labels.

Artistic dataset (Art Photo). In Art Photo dataset, 806 photos are selected from some art sharing sites using the names in emotion categories as search terms (Machajdik and Hanbury, 2010*a*). The artists, who take the photos and upload them to websites, determine emotion categories of the photos. The artists try to evoke a certain emotion for the viewers of the photo through the conscious manipulation of the emotional objects, lighting, colors, etc. In this dataset, each image is dedicated to a single emotion category.

Abstract dataset (Abstract). This dataset consists of 280 abstract paint-

Table 4.1 : The percentage of images in one emotion category that can evoke another emotion in IAPS dataset(positive emotions)

Evoke / Category	Amusement	Awe	Contentment	Excitement
Amusement%	100	10.8	70.3	40.5
Awe%	7.4	100	42.6	38.9
Contentment%	41.3	36.5	100	47.6
Excitement%	27.2	39.1	54.5	100

Table 4.2 : The percentage of images in one emotion category that can evoke another emotion in IAPS dataset(negative emotions)

Evoke / Category	Anger	Disgust	Fear	Sad
Anger%	100	87.5	25	75
Disgust%	9.5	100	35	23.0
Fear%	4.8	61.9	100	7.1
Sad%	9.7	27.4	4.8	100

ings. Unlike images in IAPS dataset and Art Photo dataset, images in the Abstract dataset represent emotion through overall color and texture, instead of some emotional objects (Machajdik and Hanbury, 2010*a*). In this dataset, the ground truth was obtained through a web-survey where the images were peer rated by the participants, who could select the best fitting emotional category from the eight emotion categories mentioned above. The total 280 images in the dataset were rated 14 times by different people. For each image, the emotion category with the most votes was selected as the emotion category of that image. The image would be removed from

Figure 4.4 : *Sad* images in IAPS(a), Art Photo(b) and Abstract(c). It can easily obvious that emotion is evoked through different ways in the three datasets.

the dataset if it has more than one identical highest votes. As a result, 228 images were used for affective image classification. Same as Art Photo dataset, each image in Abstract dataset only has one emotion label.

Fig. 4.4 shows three example images with emotion category of *sad* selected from three datasets. It can be found that images from different datasets evoke the emotion in different ways. For the images in IAPS (a), the emotion of an image is usually conveyed by objects and regions within that image. Images in Art Photo (b) evoke the emotion through the conscious combination of emotional objects and overall visual characteristics, such as color, shape, lighting. Images in Abstract (c) represent emotions through overall abstract visual characteristics instead of emotional objects. Table 4.3 lists the image distributions for different emotion categories of the above 3 datasets.

4.2.3 Parameter Tuning

Two parameters need to be tuned in the experiments: 1) M in Equation 4.2 indicating the number of words in BoVW approach and 2) K in Equation 2 indicating the number of topics in pLSA representation. The changes of these two parameters

Table 4.3 : The number of the images per emotional categories in three datasets

Dataset	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sad	Sum
IAPS	37	8	54	63	74	55	42	63	395
Art photo	101	77	102	70	70	105	115	166	806
Abstract	25	3	15	63	18	36	36	32	228
Combined	163	88	171	196	162	196	193	260	1429

may lead to the changes of the affective map. Experiments were carried out to select these two parameters for the best performance. As shown in Fig. 4.5, the different number of words (range from 100 to 500) is tested combined with the different number of topics (range from 10 to 50) in order to select an optimized combination of the parameters mentioned above. Images with emotion *amusement* from Art Photo dataset were used for parameter tuning, because of two reasons: 1) in between IAPS and Abstract, Art Photo images focus on using not only objects/regions but also overall visual characteristics to evoke emotions. 2) The emotion category of *amusement* has a moderate number of images for experiments.

From Fig. 4.5, it can be found that for the same number of words M, affective image classification results improve when the number of topics K increases. However, the improvement becomes very little when K exceeds 30. Furthermore, if K is fixed , affective image classification results hardly improve when M exceeds 300. Meanwhile, the increase of M and K largely increase the computing time. As a tradeoff, 300 for M and 30 for K is chosen in the experiment.

To find out the proper number of level for image segmentation in the proposed methods. Experiments of different levels on emotion category *amusement* is also conducted in IAPS dataset. From Fig. 4.6, it can be easily found out that when the number of levels is less than 5 for pyramid segmentation and 4 for SLIC, the emotion classification results are improved when the number of levels is increased.

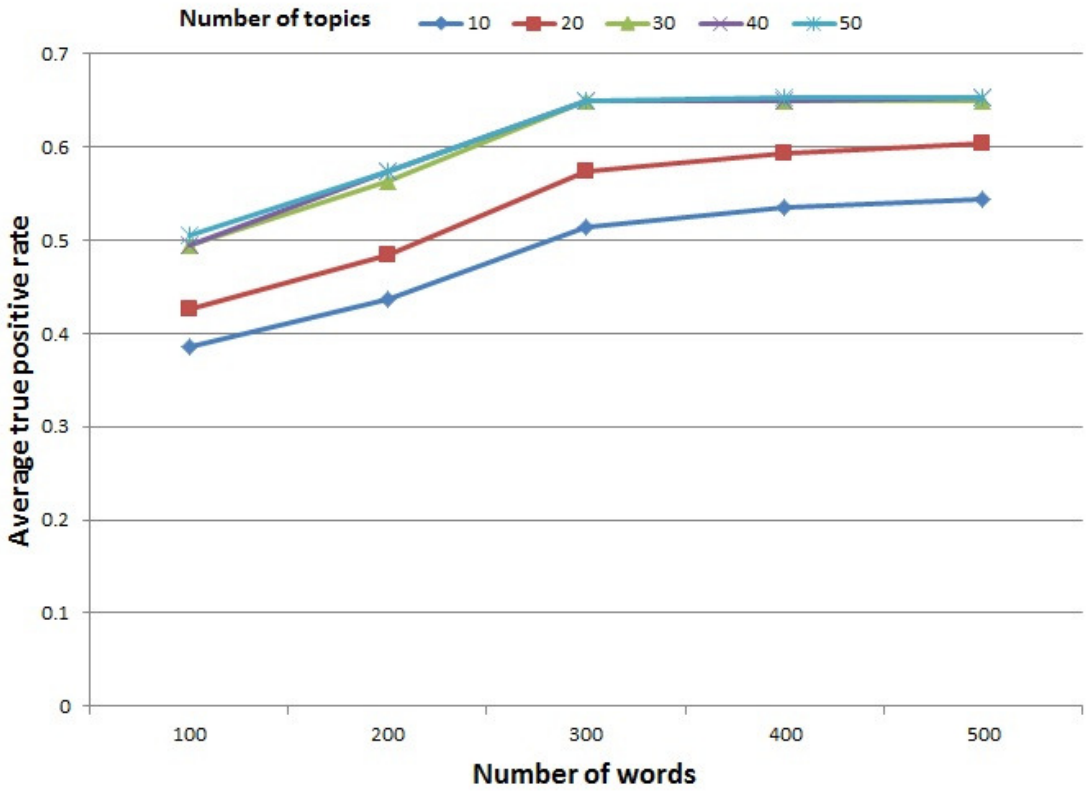

Figure 4.5 : Affective image classification results for different number of words combined with different number of topics

This demonstrates that the emotional information extracted from image regions can help to classify the emotion of the whole image. As the image blocks extracted from coarse to fine segmentation of the image, the performance is also improved. It is obvious that the image blocks extracted from SLIC can cover the emotional regions more exactly. However, when the number of image blocks in a level is increased to 400 for SLIC and 1024 for pyramid segmentation, the size of each image blocks is too small to cover the emotional region. Using the emotion information from these invalid image blocks will reduce the performance of the proposed method. Therefore, the best number of level is 4 for SLIC and 5 for pyramid segmentation, separately.

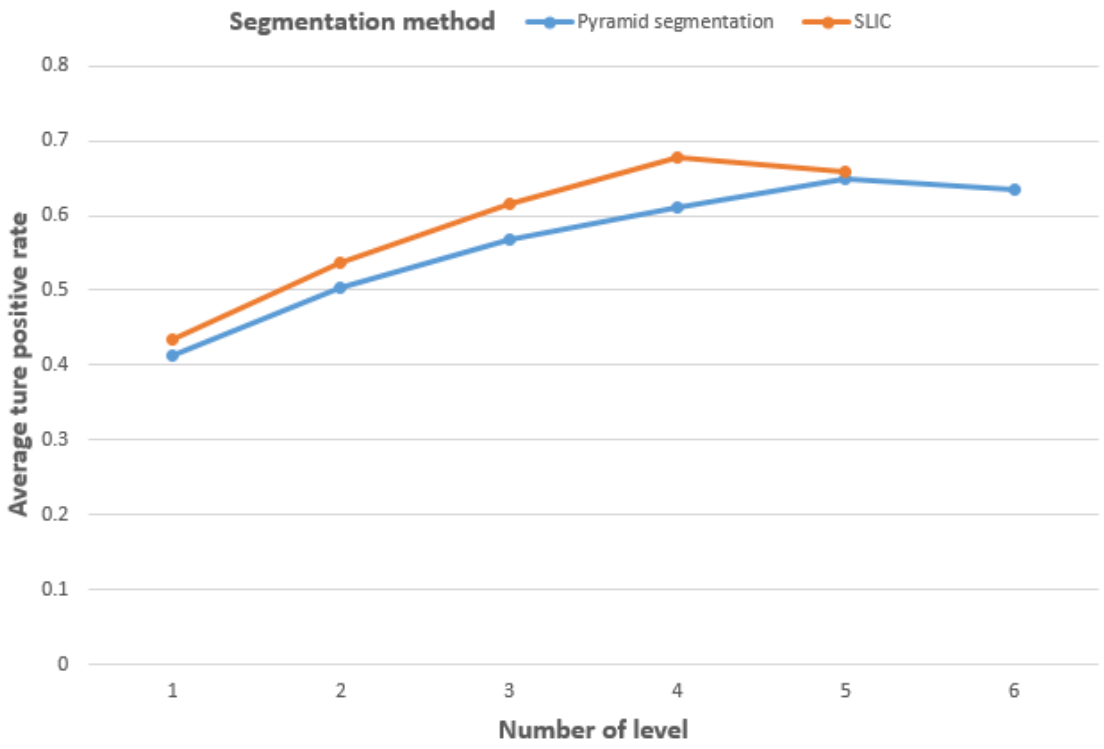

Figure 4.6 : Affective image classification results for different levels using two image segmentation methods

4.2.4 Results and Discussions

Fig 4.7, 4.8 and 4.9 show the comparison of the affective image classification results. Four different results are compared: the proposed method with SLIC, the proposed method with pyramid segmentation, Machajdik *et al.* (Machajdik and Hanbury, 2010*a*) and Zhao *et al.* (Zhao et al., 2014*a*). From the results, it can be found out that the proposed method (with pyramid segmentation and with SLIC) outperforms the other two methods and improves the classification accuracy for most of the emotion types. The classification results of the proposed method with SLIC is 5.1% higher than the state-of-the-art methods. The average improvements for the three datasets are 6.7% for the IAPS, 5.2% for the Art photo and 2.7% for the Abstract. The improvements of the classification results can be attributed

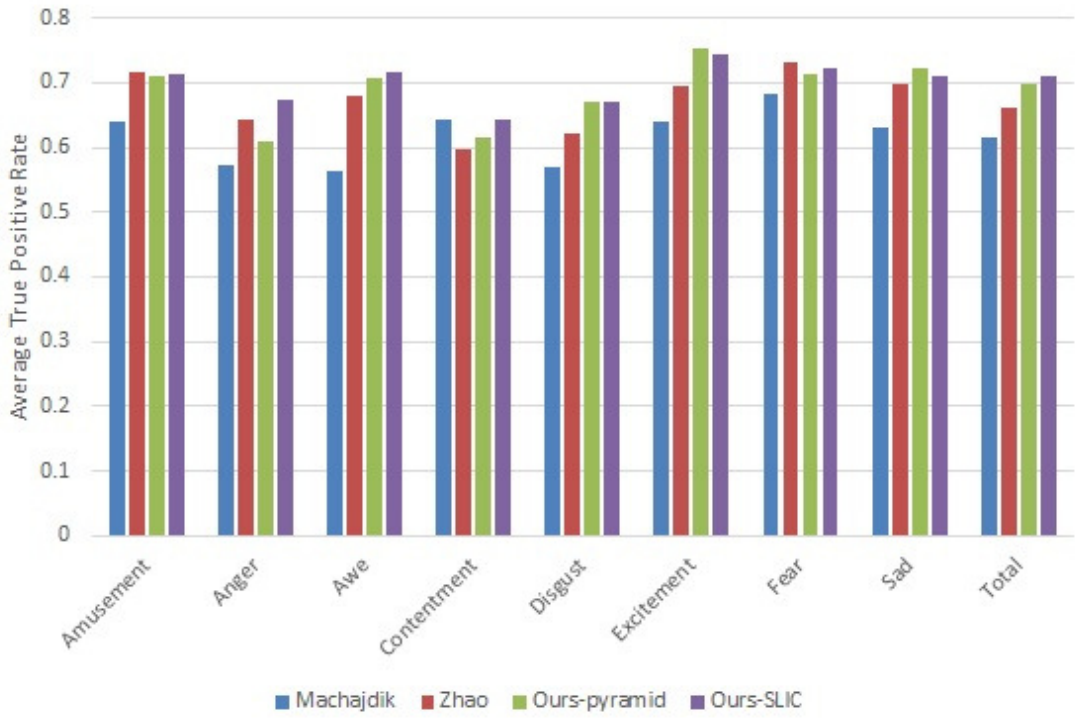

Figure 4.7 : Classification performance on the Art Photo for the proposed method with pyramid segmentation and SLIC compared to Machajdik *et al.* (Machajdik and Hanbury, 2010*a*) and Zhao *et al.*(Zhao et al., 2014*a*)

to analyzing affective images at region-level. Different regions in an image may evoke various emotions, which might be different from the dominant emotion of that image. Compared to existing methods which extract features from a whole image, region-level analysis can reduce the interference among different regions.

However, the improvements of using the proposed method on different datasets are unbalanced. The improvement on IAPS dataset is higher than that on the Abstract dataset. Images in the IAPS dataset usually evoke emotions through objects and regions, while in the Abstract dataset, emotions are represented by overall abstract visual characteristics instead of objects and regions. The unbalanced improvements demonstrate that the proposed method is good at dealing with the images whose emotions are mainly conveyed by emotional object and/or regions.

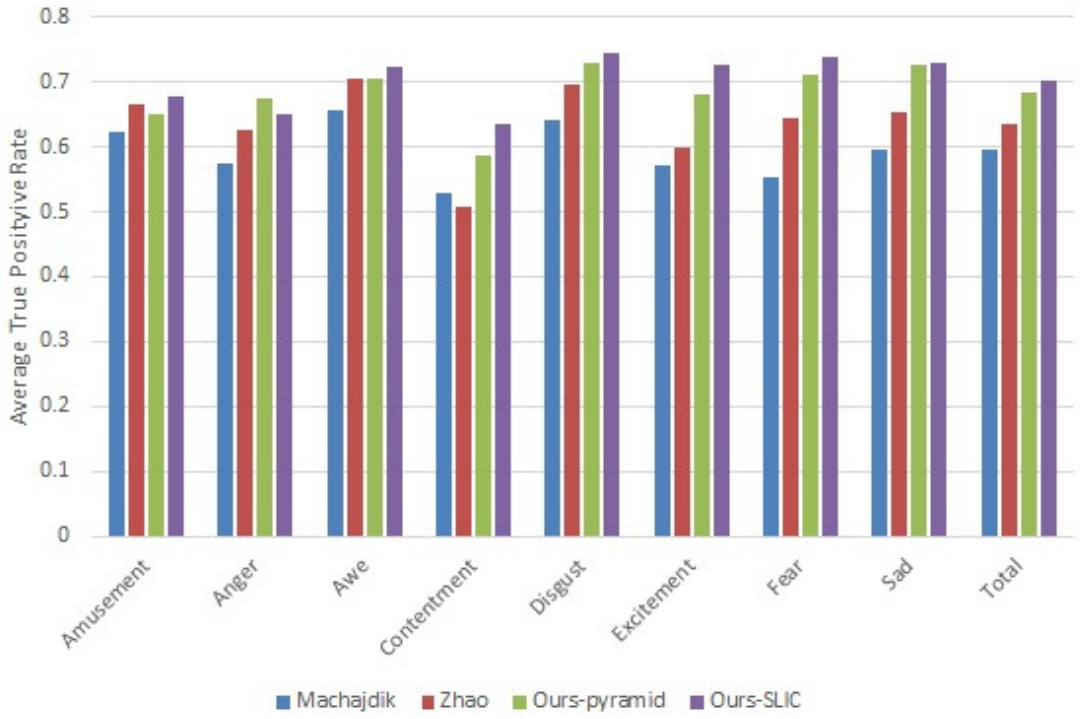

Figure 4.8 : Classification performance on the IAPS for the proposed method with pyramid segmentation and SLIC compared to Machajdik *et al.* (Machajdik and Hanbury, 2010*a*) and Zhao *et al.* (Zhao et al., 2014*a*)

Experimental results also indicate that SLIC provides more precise segmentation results for emotional regions than pyramid segmentation, therefore, improves the performance of affective image classification.

Since emotion is a very subjective concept, people with different cultural background may have different feelings to an image. In some datasets, some images are labeled with multiple emotion categories. Affective maps of an image can provide the probability of a pixel belonging to eight different emotion categories, from which the emotion distributions of eight emotions in that image can be calculated. Fig. 4.10 shows emotion distributions of eight emotion categories in three image examples

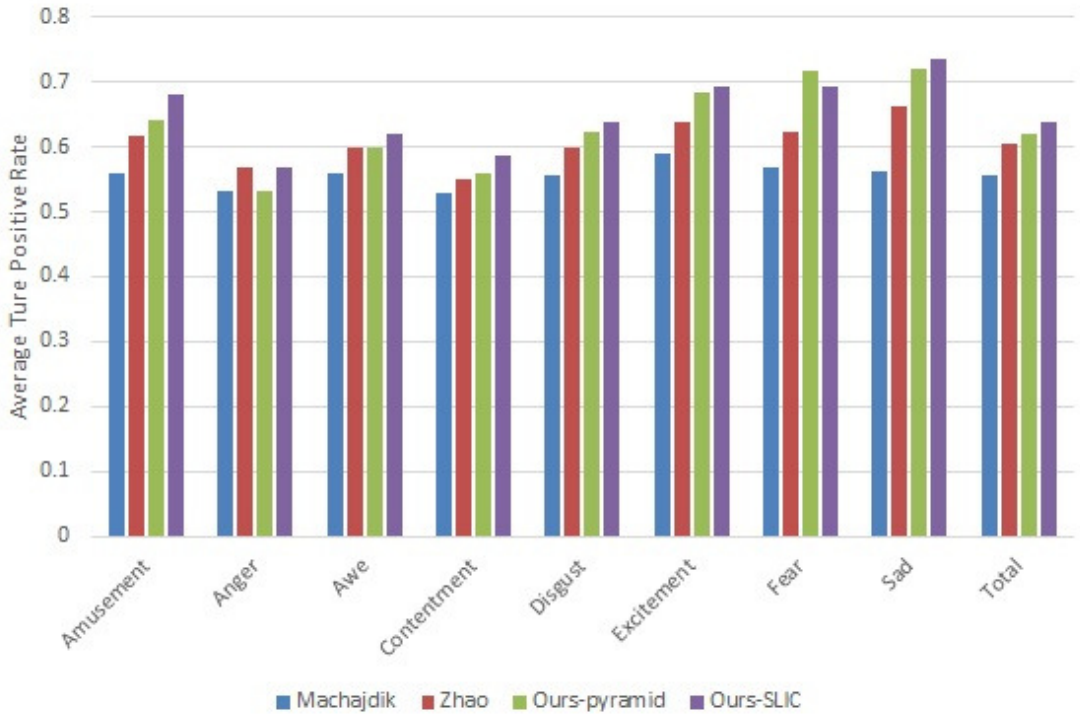

Figure 4.9 : Classification performance on the Abstract for the proposed method with pyramid segmentation and SLIC compared to Machajdik *et al.* (Machajdik and Hanbury, 2010*a*) and Zhao *et al.* (Zhao et al., 2014*a*)

4.3 Saliency Detection

Research in psychology and neuroscience reveals the influence of emotion on the human's attention. However, it is unable to utilize the image-level affective information extracted from existing methods for saliency detection. In this section, the affective map is incorporated into computational visual saliency to investigate the effectiveness of affective map.

To represent the different families of saliency detection methods, four saliency detection methods are chosen as baseline methods.

- Itti's method (Itti et al., 1998). This is the seminal work for saliency detection. It is also the earliest contrast-based method.

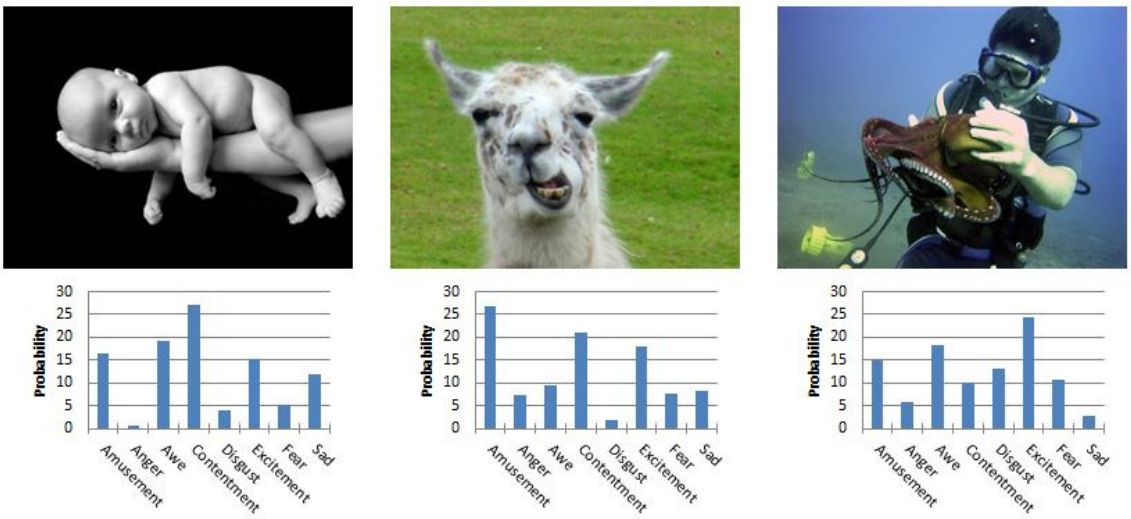

Figure 4.10 : Emotion distributions of eight emotion categories in three image examples

- The Attention based on Information Maximization (AIM) method (Bruce and Tsotsos, 2005). As an information-based saliency detection method, AIM uses the context information of an image to estimate the salient region within that image.

- The Saliency Using Natural images method (SUN) (Zhang, Tong, Marks, Shan and Cottrell, 2008). SUN method is also an information-based saliency detection method. The difference between the AIM and SUN is that information used in SUN indicates the prior knowledge learned from natural image dataset.

- The Graph-Based Vision Saliency (GBVS) (Harel, Koch and Perona, 2006). GBVS method is a contrast-based method, which calculates saliency using contrast by random walk on graphs.

Then, the framework of Judd (Judd, Ehinger, Durand and Torralba, 2009) is borrow and a linear SVM is employed to fuse the saliency map extracted from baseline methods with the proposed affective maps extracted from the images. The new

saliency map incorporating emotion information is used to evaluate the effectiveness of affective map.

4.3.1 Experimental Setup

The dataset used for testing is the National University of Singapore Eye Fixation (NUSEF) (Ramanathan, Katti, Sebe, Kankanhalli and Chua, 2010), a public eye tracking dataset, including 758 images in total. On average, about 25 users' eye gaze data are recorded for each image.

Five sets of results are compared in the experiments:

- affective map based on pyramid segmentation only
- affective map based on SLIC only
- baseline method
- baseline method incorporating with affective map based on pyramid segmentation
- baseline method incorporating with affective map based on SLIC

Receiver Operating Characteristic (ROC) curve is used to evaluate the performance through calculating the Area Under the Curve (AUC). A paired t-test is employed to evaluate the difference between two curves (baseline method and baseline method incorporating with affective map). The smaller the p-value of t-test is, the more significant difference between the two curves is.

4.3.2 Results and Discussions

In previous work, the impact of emotional factors in saliency detection has already been demonstrated(Liu, Xu, Wang, Rao and Burnett, 2016). In this paper,

Table 4.4 : AUC of ROC curves and p-values for testing affective map.

Baseline method	Baseline	with affective map based on pyramid	with affective map based on SLIC
ITTI	0.5625	0.6826 (<0.01)	0.7056 (<0.01)
AIM	0.6760	0.7447 (<0.01)	0.7569 (<0.01)
SUN	0.6447	0.7018 (<0.01)	0.7202 (<0.01)
GBVS	0.6388	0.7210 (<0.01)	0.7285 (<0.01)

the performance of the affective map generation method is improved by using SLIC to extract multi-scale blocks from images.

The AUC values and p-values are compared in Table 4.4. As shown in Table 4.4, both affective maps based on pyramid segmentation and based on SLIC can significantly improve the performance of saliency detection. Furthermore, SLIC makes the affective map more precise than pyramid segmentation, therefore achieve a better performance of saliency detection. The improvements of using affective map based on SLIC are 0.1431 of AUC for ITTI, 0.0809 of AUC for AIM, 0.0755 AUC for SUN and 0.0897 AUC for GVBS, average 15% improvement for these methods.

The ROC curves are shown in Fig. 4.11. It can be found out that, incorporating

with affective maps, the saliency detection results are largely improved. The p-values, which are less than 0.01, indicate that incorporating with affective map does change the ROC curve of the four baseline methods significantly.

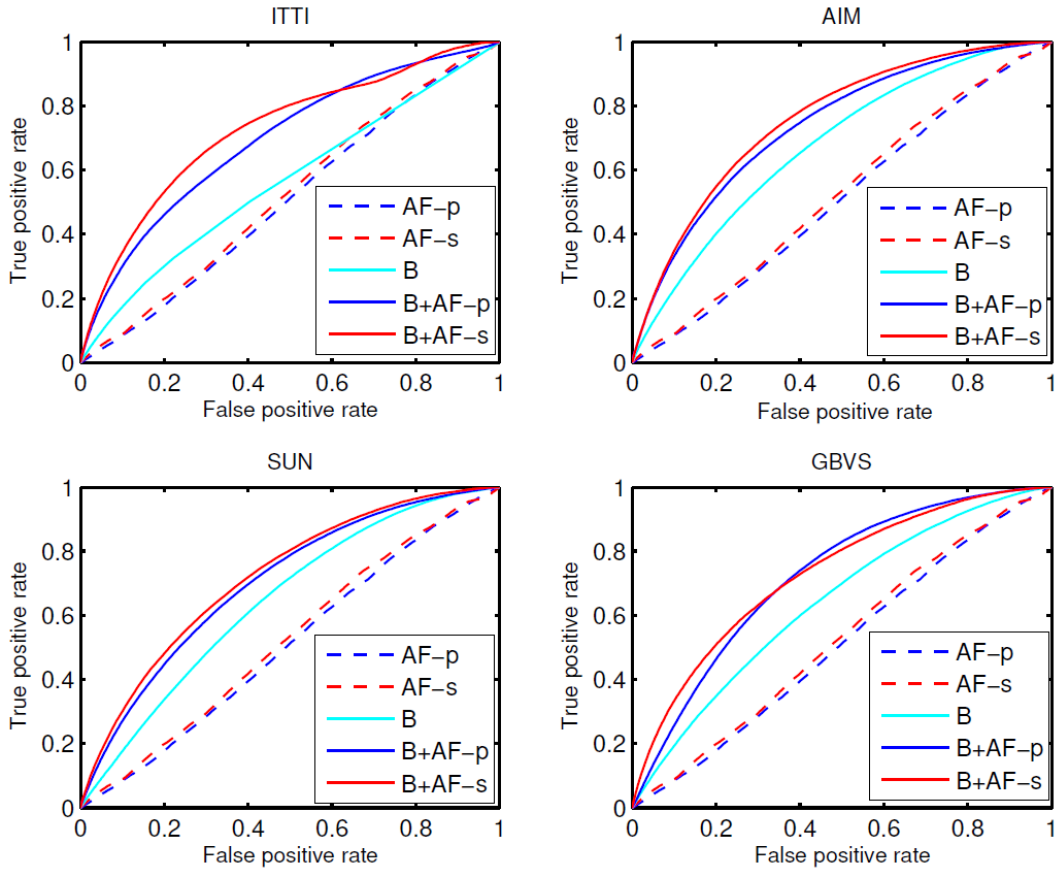

Figure 4.11 : The comparison of ROC curves for testing affective map (AF-p: affective map based on pyramid segmentation; AF-s: affective map based on SLIC; B:baseline method; B+AF-p: baseline method incorporating with affective map based on pyramid segmentation; B+AF-s: baseline method incorporating with affective map based on SLIC)

4.4 Discussions

In this work, the concept of affective map is proposed to analyze affective images at a region-level. The affective map can not only improve the precision of affective

image classification but also extend the application field of affective image analysis. Considering the difficulty to collect a well labeled dataset with emotional region, the proposed affective map also provide a way to demonstrate the effectiveness of emotional region. Affective map also makes those applications which require detailed emotion information possible. Experiments have been carried out on three popular datasets for affective image classification and NUSEF dataset for saliency detection. Promising experimental results indicate that: 1) affective map improves the affective image classification results compared to the state-of-the-art methods; 2) affective map incorporating with saliency detection method significantly improves the performance of saliency detection.

Chapter 5

Learning Multi-level Deep Representations for Image Emotion Classification

As shown in Figure 5.1, image emotion is related to complex visual features from high-level to low-level for both global and local views.

Recently, with the rapid popularity of Convolutional Neural Network (CNN), outstanding breakthroughs have been achieved in many visual recognition tasks, such as image classification (Krizhevsky et al., 2012), image segmentation (Long et al., 2015), object detection (Ren et al., 2015) and scene recognition (Zhou et al., 2014). Instead of designing visual features manually, CNN provides an end-to-end feature learning framework, which can automatically learn deep representations of images from the global view. Several researchers have also applied CNN to image emotion classification. However, as shown in Figure 5.2, the currently used CNN methods, such as AlexNet (Krizhevsky et al., 2012), for visual recognition cannot well deal with mid-level image aesthetics and low-level visual features from local view. In (Alameda-Pineda, Ricci, Yan and Sebe, 2016), the authors indicated that AlexNet is not effective enough to extract emotion information from abstract paintings, whose emotions are mainly conveyed by mid-level image aesthetics and low-level visual features.

To address the issue that traditional CNN can not utilize low-level visual features , in this work, a new deep network (MldrNet) that learns multi-level deep representations from both global and local views for image emotion classification is proposed. Figure 5.3 shows an overview of the proposed MldrNet network. The

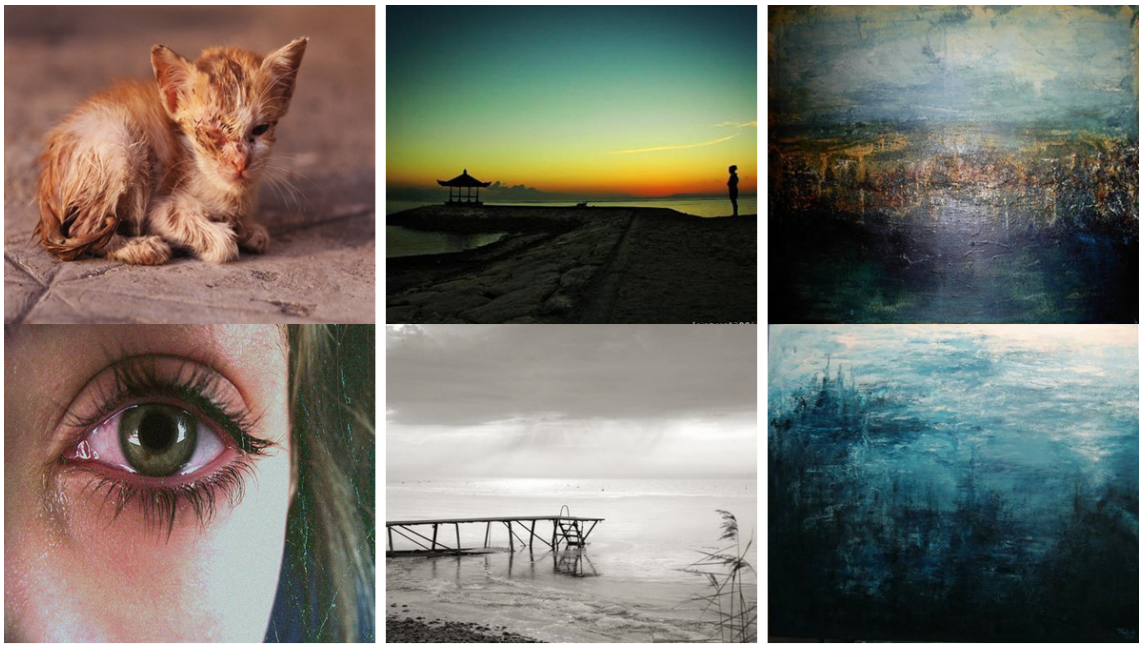

Figure 5.1 : Sample images from different datasets that evoke the same emotion *sadness*. It can be found out that image emotion is related to many factors. Left: web images whose emotions are mainly related to image semantics. Middle: art photos whose emotions are mainly related to image aesthetics, such as compositions and emphasis. Right: abstract paintings whose emotions are mainly related to low-level visual features, such as texture and color.

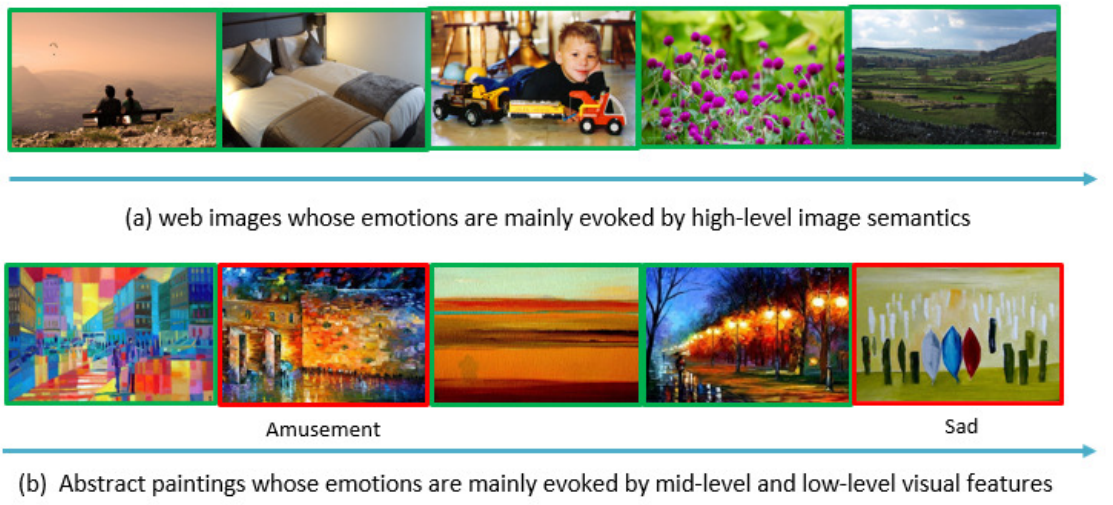

Figure 5.2 : Top 5 classification results for emotion category *contentment* using AlexNet (Krizhevsky et al., 2012) on web images and abstract paintings. *Green (Red)* box means correct (wrong) results, the correct label for wrong retrieve are provided. It is clear that AlexNet produces better matches for web images than abstract paintings. This means AlexNet deals high-level image semantics better than mid-level and low-level visual features.

traditional CNN method is designed for center-position object classification, which cannot effectively extract mid-level image aesthetics and low-level visual features from the local view. To perform end-to-end learning methods for different levels of deep representation from an entire image, we propose a CNN model with side branches to extract different levels of deep representations. Through a fusion layer, different levels of deep representations are integrated for classification. We notice that different fusion methods will largely affect the classification results when using noisy labeled data as training data. To demonstrate the effectiveness of the proposed MldrNet and explore the impact of different fusion methods, we conduct extensive experiments on several publicly available datasets for different kinds of images, e.g.

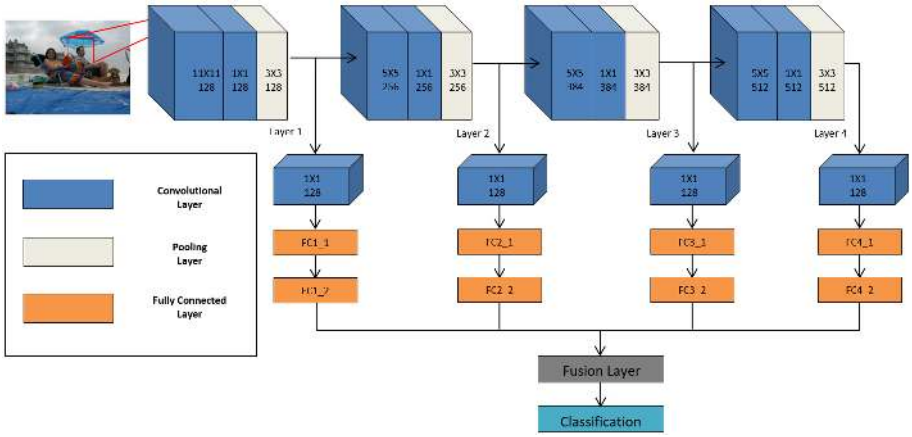

Figure 5.3 : Overview of the proposed multi-level deep representation network (MldrNet). Different levels of deep representations related to high-level, mid-level and low-level visual features are extracted from different convolutional layer and fuse using fusion layer. The fusion representations are finally used for classification

web images and abstract paintings.

The main contribution in this work is that a new CNN based method is proposed that combines different levels of deep representations, such as image semantics, image aesthetics and low-level visual features, from both global and local views for image emotion classification. By combining different levels of deep representations, the proposed method can effectively extract emotional information from images. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods using deep features or hand-crafted features on both Internet images and abstract paintings.

5.1 Multi-level deep representations for image emotion classification

In this section, the proposed method that learns multi-level deep representations (MldrNet) is introduced for image emotion classification. Consider image emotion is

related to different levels of features, i.e., high-level image semantics, mid-level image aesthetics and low-level visual features, the proposed method unifies different levels of deep representation within one CNN structure. In particular, we propose a fusion layer to support multi-level deep representations aggregation based on the characteristics of image emotion. Following the aforementioned discoveries, we divide the images into 8 emotion categories (positive emotion *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotion *Anger*, *Disgust*, *Fear*, *Sadness*) for visual emotion classification.

5.1.1 Convolutional Neural Network

Before introducing the proposed MldrNet, we first review the CNN model that has been widely used for computer vision tasks (Krizhevsky et al., 2012). Given one training sample $\{(x, y)\}$, where x is the image and y is the associated label, CNN extracts layer-wise representations of input images using convolutional layers and fully-connected layers. Followed by a softmax layer, the output of the last fully-connected layer can be transformed into a probability distribution $\mathbf{p} \in \mathbb{R}^m$ for image emotions of n categories. In this work, $n = 8$ indicates eight emotion categories. The probability that the image belongs to a certain emotion category is defined below:

$$p_i = \frac{exp(h_i)}{\sum_i exp(h_i)}, i = 1, ..., n, \tag{5.1}$$

where h_i is the output from the last fully-connected layer. The loss of the predicting probability can also be measured by using cross entropy

$$L = -\sum_i y_i \log(p_i), \tag{5.2}$$

where $y = \{y_i | y_i \in \{0, 1\}, i = 1, ..., n, \sum_{i=1}^{n} p_i = 1\}$ indicates the true emotion label of the image.

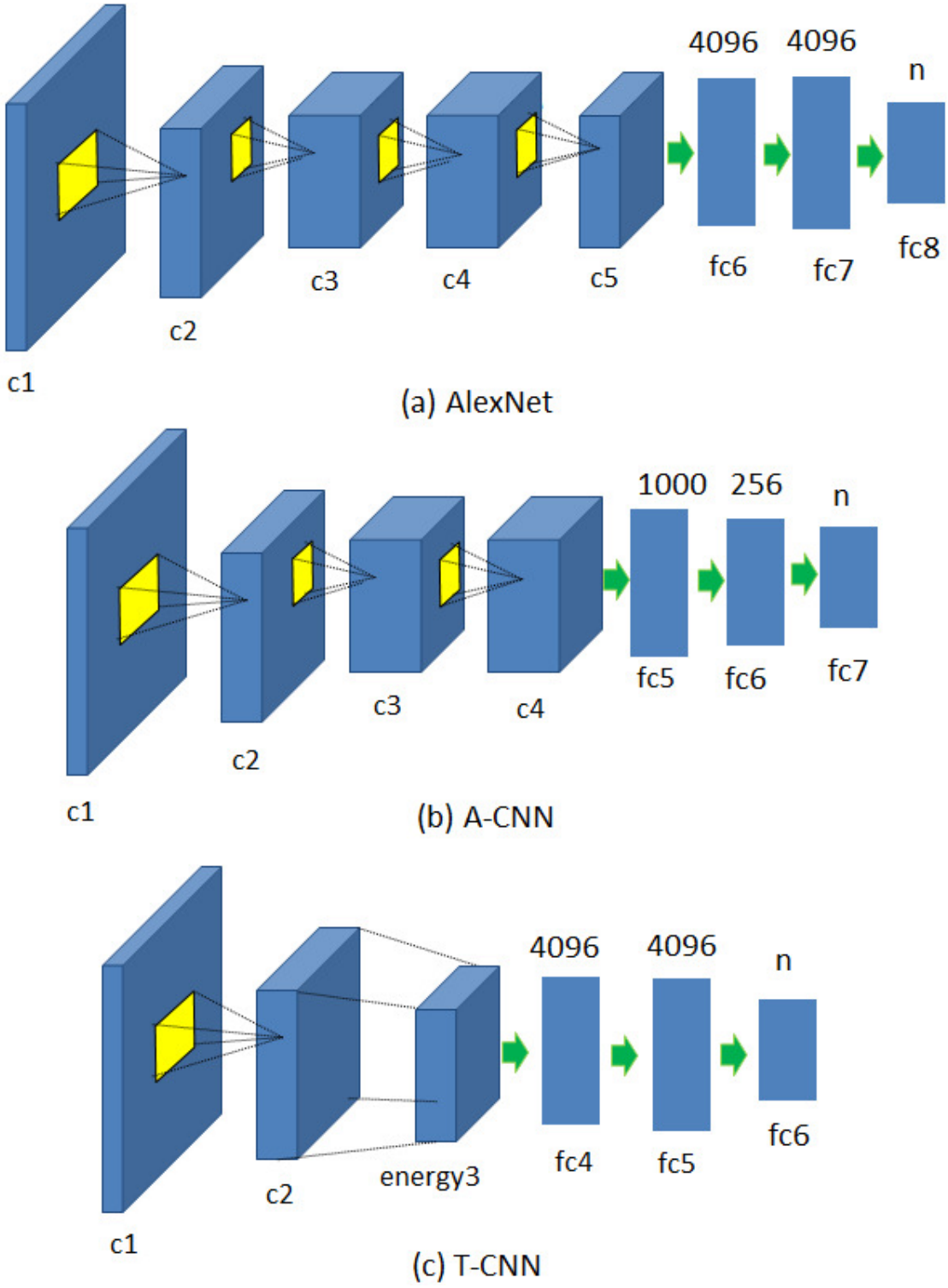

Figure 5.4 : The structures of different CNN models that deal with different levels of computer vision tasks.

AlexNet is used to classify image on the large-scale dataset. It contains five convolutional layers followed by max-pooling layers, and three fully-connected layers, which contains 4,096, 4,096 and 8 neurons, respectively. The structure of AlexNet is shown in Fig 5.4(a). AlexNet is mainly trained for semantic-level image classification and tends to extract high-level deep representation about image semantics. It cannot effectively extract emotion information from abstract painting whose emotion is mainly conveyed by mid-level image aesthetics and low-level visual features (Alameda-Pineda et al., 2016). As discussed in Section 1, AlexNet is likely not informative enough for image emotion classification.

5.1.2 Analysis of different CNN models

Emotion-related image features can be roughly divided into low-level visual features, such as color, line and texture, mid-level image aesthetics, including composition and visual balance, and high-level image semantics. As CNN models contain a hierarchy of filters, the level of representations learned from CNN models are higher if one goes "deeper" in the hierarchy (Zeiler and Fergus, 2014). This means that if a CNN structure contains more convolutional layers, the level of feature extracted from the CNN structure is higher. Various CNN structures used in different computer vision tasks have also demonstrated this conclusion. To extract deep representations about mid-level image aesthetics and low-level visual features, different kinds of CNN models inspired by AlexNet, which contain less number of convolutional layers, are developed (Lu, Lin, Jin, Yang and Wang, 2014; Andrearczyk and Whelan, 2016).

Image aesthetics has a close relationship with image emotion. To effectively deal with the mid-level image aesthetics, Aesthetics CNN(A-CNN) model has been developed (Lu et al., 2014) As shown in Figure 5.4(b), A-CNN consists of four convolutional layers and three fully-connected layers, which contains 1,000, 256 and

8 neurons, respectively. The first and second convolutional layers are followed by max-pooling layers. Even contains less convolutional layers compared to AlexNet, A-CNN has a better performance on image aesthetics analysis.

Texture has been proven as one of the important low-level visual features related to image emotion classification (Machajdik and Hanbury, 2010*b*). To extract deep representations about the texture of images, an efficient CNN model, T-CNN, is designed for texture classification (Andrearczyk and Whelan, 2016). As shown in Figure 5.4(c), T-CNN removes the last three conventional layers of AlexNet, and adds an energy layer (average-pooling layer with the kernel size as 27) behind the second convolutional layers. Following the energy layer, there are still three fully-connected layers, which contains 4,096, 4,096 and 8 neurons, respectively.

From the aforementioned CNN models, it can be found that the structures of different CNN models are similar, the main difference is the number of convolutional layers. This means some parameters for different CNN models that can extract different levels of deep representations can be shared. Based on the observation, different CNN models is unified into one CNN structure, which can not only improve the classification accuracy of image emotion but also gain a better parameter efficiency.

5.1.3 Deep Network Learning Multi-level Deep representations

To effectively unify different levels of deep representations in one CNN model, a multi-level deep representations network (MldrNet) is proposed, which contains the main network and four side branches. Different levels of deep representation from both the global view and local view can be extracted from different convolutional layers in the proposed MldrNet. As shown in Figure 5.4, the proposed MldrNet consists of 4 convolutional layers, whose filter size are 11×11, 5×5, 5×5 and 5×5, respectively. For each convolutional layer, it is followed by two fully connected layers.

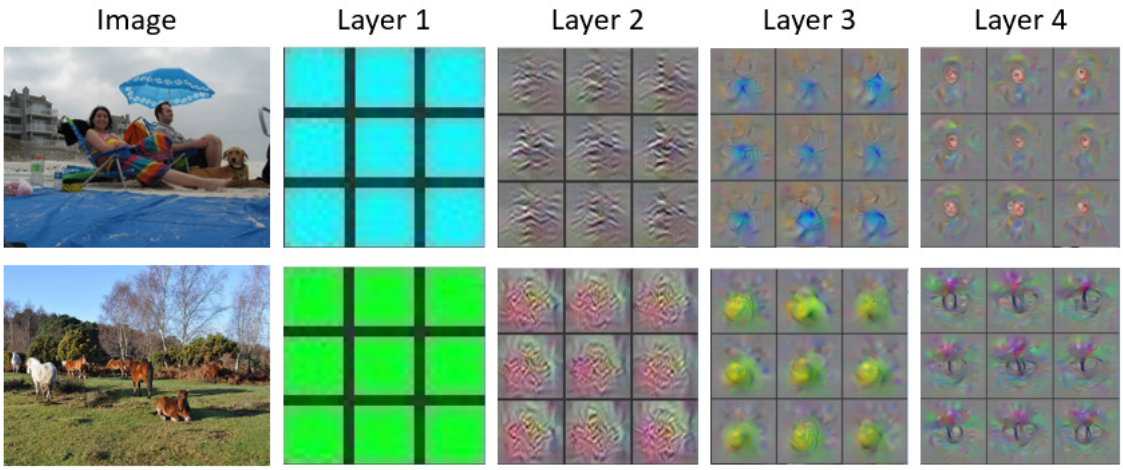

Figure 5.5 : Visualization of the weights of filter, which produce an activation map with the highest activation, in each convolutional layer.

One problem for the proposed MldrNet is that the dimension of the output of each convolutional layer is different. Inspired by the structure of GoogleNet (Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke and Rabinovich, 2015), a 1×1 convolutional layer is inserted with 128 filters between the pooling layer and fully connected layer for each layer of the proposed MldrNet. The 1×1 convolutional layer unifies the output dimension of each layer and rectifies linear activation.

Compared to high-level image semantic information extracted from the highest layer of the proposed MldrNet, deep representations extracted from the lower layers provide the additional information, such as low-level color and texture and mid-level composition and visual balance, which is related to image emotion. Existing research on image emotion analysis has demonstrated that, with the additional information related to low-level and mid-level image features, the performance of image emotion classification will be significantly improved (Zhao et al., 2014*b*).

When designing the proposed MldrNet, two problems need to be considered. First, the proper number of layers in the proposed deep network should be inves-

tigated. As mentioned before, making the network deeper may not improve the image emotion classification results. If the number of layers is too high, the number of parameters will largely increase because each layer needs to have its own weights, while the contribution of these layers to emotion classification may be very little. However, if the number of layers is too low, the extracted deep representations may not be able to well represent image emotion. To show the differences of various deep representations extracted from each layer in the proposed network, the weights of the filter is visualized, which produces an activation map with the highest activation, in each convolutional layer in Figure 5.5. It is obvious that the deep representations extracted from layer 1 and layer 2 are related to low-level features, while in layer 3, the deep representations focus on abstract concepts that reflect image aesthetics. In the highest layer, the deep representations mainly represent concrete objects in images, i.e. human face and horse. Experiments are also conduct to investigate the impact of the different number of layers in MldrNet for image emotion classification in Section 5.2.2

Second, the role of deep representations extracted from the different layer of MldrNet for evoking emotions may vary for different kinds of images. To effectively combine the different levels of deep representations, the fuse function needs to be carefully chosen. The most commonly used fusion functions are introduced in the proposed MldrNet, including concatenation, $min(\cdot)$, $max(\cdot)$ and $mean(\cdot)$. The detailed discussion of the fusion layer will be shown in Section 5.1.4.

5.1.4 Fusion Layer

Fusion layer is the core component of the proposed multi-level deep representations network, which is comprised of a collection of fusion functions. As some image information will be disregarded when passing a convolutional layer, some existing models, i.e. ResNet (Szegedy et al., 2015) and DenseNet (He et al., 2016), com-

bines information from different convolutional layers to improve the performance. However, they just simply concatenate multi-level features through skip-connection, which means information extracted from different convolutional layers has equal weights. While in image emotion analysis, different level of features may have a different impact on evoking emotions. To choose a suitable fusion function for image emotion classification result, different fusion functions are used in the fusion layer to fuse multi-level deep representations.

The deep representation extracted from the ith layer is defined as h_i and the fusion function is $f(x)$. Then the representation of the entire image can be aggregated by using the representation of each layer

$$\hat{h} = f(\hat{h_1}, \hat{h_2}, ..., \hat{h_i}). \tag{5.3}$$

The distribution of emotion categories of the image and the loss L can be computed as

$$p_i = \frac{exp(\hat{h})}{\sum_i exp(\hat{h})} and L = -\sum_i y_i \log(p_i), \tag{5.4}$$

In the experiments, the fusion function can be $f(x) = min, max, mean$. It can be easily found out that the function $mean(\cdot)$ assigns the same weight to deep representations extracted from each convolutional layer, while the function $max(\cdot)$ and $min(\cdot)$ would encourage the model to increase the weight of one out of all layers of deep representations. The choice of the fusion function is of critical importance in the proposed method. The comparison results of utilizing different fusion functions are shown in Section IV.

5.2 Experiments

In this section, the performance of the proposed MldrNet is evaluated on different datasets. The recently published large-scale dataset for emotion recognition (You et al., 2016) and three popular used small datasets: IAPS-Subset (Mikels et al., 2005*b*), ArtPhoto and Abstract (Machajdik and Hanbury, 2010*b*) are used to evaluate the classification results over 8 emotion categories. The MART dataset (Yanulevskaya, Uijlings, Bruni, Sartori, Zamboni, Bacci, Melcher and Sebe, 2012) is used to evaluate the classification result on abstract paintings over 2 emotion categories (positive and negative).

5.2.1 Experimental Settings

Implementation Details

The proposed model is implement by using the pyTorch framework on two Nvidia GTX1080. The detailed parameters of the proposed model are presented in Fig 5.3 and the input images are cropped as 375×375 from center and corners. The batch size is set to 64. The proposed model is optimized using stochastic gradient descent(SGD). The initial learning rate is empirically set as 0.001, the momentum is 0.9, and weight decay is 0.0005. The parameters in these optimizers are initialized by using the default setting.

Datasets

Large Scale Dataset For Emotion classification: This dataset is newly published in (You et al., 2016) to evaluate the classification result over 8 different emotion categories (positive emotions *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotions *Anger*, *Disgust*, *Fear*, *Sad*). To collect this dataset, 90,000 noisy labeled images are first downloaded from Instagram and Flickr by using the names of emotion categories as the keywords for searching. Then, the downloaded

images were submitted to Amazon Mechanical Turk (AMT) for further labeling. Finally, 23,308 well-labeled images were collected for emotion recognition.

Small Scale Datasets For Emotion Classification: Three small datasets that are widely used in previous works for image emotion classification are introduced below.

(1)**IAPS-Subset**: The *International Affective Picture System* (IAPS) is a standard stimulus image set which has been widely used in affective image classification. IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, landscapes and others (Lang et al., 2008*a*). Among all IAPS images, Mikels *et al.* (Mikels et al., 2005*b*) selected 395 images and mapped arousal and valence values of these images to the above mentioned eight discrete emotion categories.

(2)**ArtPhoto**: In the ArtPhoto dataset, 806 photos are selected from some art sharing sites by using the names of emotion categories as the search terms (Machajdik and Hanbury, 2010*b*). The artists, who take the photos and upload them to the websites, determine emotion categories of the photos. The artists try to evoke a certain emotion for the viewers of the photo through the conscious manipulation of the emotional objects, lighting, colors, etc. In this dataset, each image is assigned to one of the eight aforementioned emotion categories.

(3)**Abstract**: This dataset consists of 228 abstract paintings. Unlike the images in the IAPS-Subset and ArtPhoto dataset, the images in the Abstract dataset represent the emotions through overall color and texture, instead of some emotional objects (Machajdik and Hanbury, 2010*b*). In this dataset, each painting was voted by 14 different people to decide its emotion category. The emotion category with the most votes was selected as the emotion category of that image.

MART: The MART dataset is a collection of 500 abstract paintings from the

Museum of Modern and Contemporary Art of Trento and Rovereto. These artworks were realized since the beginning of the 20 century until 2008 by professional artists, who have theoretical studies on art elements, such as colors, lines, shapes, and textures, and reflect the results of studies on their paintings. Using the relative score method in (Sartori, Culibrk, Yan and Sebe, 2015), the abstract paintings are labeled as positive or negative according to the emotion type evoked by them.

Compared Methods

To demonstrate the effectiveness of the proposed MldrNet, the proposed method is compared with state-of-the-art image emotion classification methods and the most popular CNN models:

Machajdik(Machajdik and Hanbury, 2010*b*): using the low-level visual features and mid-level features inspired by psychology and art theory which are specific to artworks.

Zhao(Zhao et al., 2014*b*): using principles-of-art-based emotion features, which are the unified combination of representation features derived from different principles, including *balance*, *emphasis*, *harmony*, *variety*, *gradation*, and *movement* and its influence on image emotions.

Rao(Rao, Xu, Liu, Wang and Burnett, 2016): using different visual features extracted from multi-scale image patches.

AlexNet+SVM(You et al., 2016): using AlexNet to extract emotion related deep features and classify them through SVM.

AlexNet(Krizhevsky et al., 2012): pre-trained based on ImageNet and fine-tuned using the large scale dataset for emotion classification.

VGGNet-19(Simonyan and Zisserman, 2014): pre-trained based on ImageNet and fine-tuned using the large scale dataset for emotion classification.

ResNet-101(He et al., 2016): pre-trained based on ImageNet and fine-tuned using the large scale dataset for emotion classification.

To fully quantify the role of different fusion functions of the proposed model and detect the suitable architecture of the proposed model, different variants of the proposed model are compared:

MldrNet-concat: simply concatenate deep representations extracted from each layer in the fusion layer.

MldrNet-max: using *max* as fusion function in the fusion layer.

MldrNet-min: using *min* as fusion function in the fusion layer.

MldrNet-mean using *mean* as fusion function in the fusion layer.

5.2.2 Emotion Classification on Large Scale and Noisy Labeled Dataset

The well labeled 23,164 images are randomly split into the training set (80%, 18,532 images), the testing set (15%, 3,474 images) and the validation set (5%, 1,158 images). Meanwhile, to demonstrate the effectiveness of the proposed approach on the noisy labeled dataset, a noisy labeled dataset is created for training by combining the images, which have been submitted to AMT for labeling but labeled from different emotion categories, with the training set of well-labeled images. The noisy labeled dataset contains 83,664 images for training. The well-labeled dataset is called as *well* dataset and noisy labeled dataset is c as *noisy* dataset. The *well* dataset and *noisy* dataset are used for training models. The testing dataset is used to test the proposed models.

Choice of Number of Layers in MldrNet

The proposed MldrNet model can utilize multi-level deep representations to classify image emotion by increasing or decreasing the number of convolutional layers.

Choosing a proper number of convolutional layers in MldrNet needs to be explored in order to achieve the best emotion classification performance. The experiments using MldrNet models with different number of convolutional layer are conducted.

Model	Accuracy
MldrNet-2 layer	52.12%
MldrNet-3 layer	58.34%
MldrNet-4 layer	67.24%
MldrNet-5 layer	67.55%
MldrNet-6 layer	67.68%

Table 5.1 : Emotion classification accuracy for MldrNet Models of different number of convolutional layer.

As shown in Table 5.1, changing the number of convolutional layers in the proposed MldrNet model will affect the classification accuracy. The models with fewer layers perform worse than the models with more than 4 layers. The main reason may be that the models with fewer layers lack the information related to high-level image features. What's more, the models with more than 4 layers cannot significantly improve the emotion classification accuracy, which indicates the contribution of these layers may be very little. Meanwhile, with more convolutional layers, the number of parameters needed to be computed is increased, therefore, the time for training these models are largely increased. Due to the above considerations, MldrNet with 4 layers is the best model the following experiments.

Choice of Fusion Function

Another important component of the proposed MldrNet model is the fusion layer. As discussed previously, the fusion function will affect the emotion classification

accuracy. It can also find that fusion function plays an important role in dealing with different training datasets.

Model	Accuracy	
	well dataset	*noisy* dataset
MldrNet-concat	66.92%	55.34%
MldrNet-max	64.79%	53.68%
MldrNet-min	62.44%	49.32%
MldrNet-mean	**67.24%**	**59.85%**

Table 5.2 : Emotion classification accuracy for MldrNet Models of different fusion function training on both *well* dataset and *noisy* dataset.

In Table 5.2, the result of the proposed MldrNet is presented with a variety of fusion functions using both *well* dataset and *noisy* dataset for training. Compare to the MldrNet model of using $max(\cdot)$ and $min(\cdot)$ as fusion function, the performances of MldrNet model of using $mean(\cdot)$ and $concat(\cdot)$ as fusion functions are better. Especially for fusion function $mean(\cdot)$, the proposed MldrNet achieves the best performance when using different training dataset. Unlike $max(\cdot)$ and $min(\cdot)$, when using the $mean(\cdot)$ and $concat(\cdot)$ as fusion function, the model can keep more emotional information extracted from each convolutional layers. Using $mean(\cdot)$ as fusion function in the proposed model can better fuse the emotional information for image emotion classification.

Comparison of Different Methods

To investigate the effectiveness of the proposed MldrNet model, the proposed model is compared with different image emotion classification methods, including the state-of-the-art method using hand-crafted features and popular deep learning

models. All methods use *well* dataset as training dataset. The results are shown in Table 5.3.

Methods	Accuracy
Zhao	46.52%
Rao	51.67%
AlexNet-SVM	57.89%
AlexNet	58.61%
VGGNet-19	59.32%
ResNet-101	60.82%
MldrNet	67.24%

Table 5.3 : Emotion classification accuracy for different methods on the large scale dataset for image emotion classification.

From Tabel 4.3, it has the following observations. First of all, methods using deep representations outperforms the methods using hand-crafted features. These hand-crafted features are designed based on several small-scale datasets composed of images from specific domains, which cannot comprehensively describe image emotion compared to deep representations. Then, for methods using deep representations, we can find that, compared to AlexNet, even though containing more convolutional layers and providing higher deep representations in VGGNet-19 and ResNet-101, the performances are just slightly improved. Only containing 4 convolutional layers, the proposed MldrNet model considering both mid-level and low-level deep representations significantly improves the emotion classification accuracy, compared with other 'deeper' CNN models. Finally, when using the noisy dataset for training, the proposed MldrNet model can still achieve competitive emotion classification accuracy. This means the proposed method can utilize the images which are directly

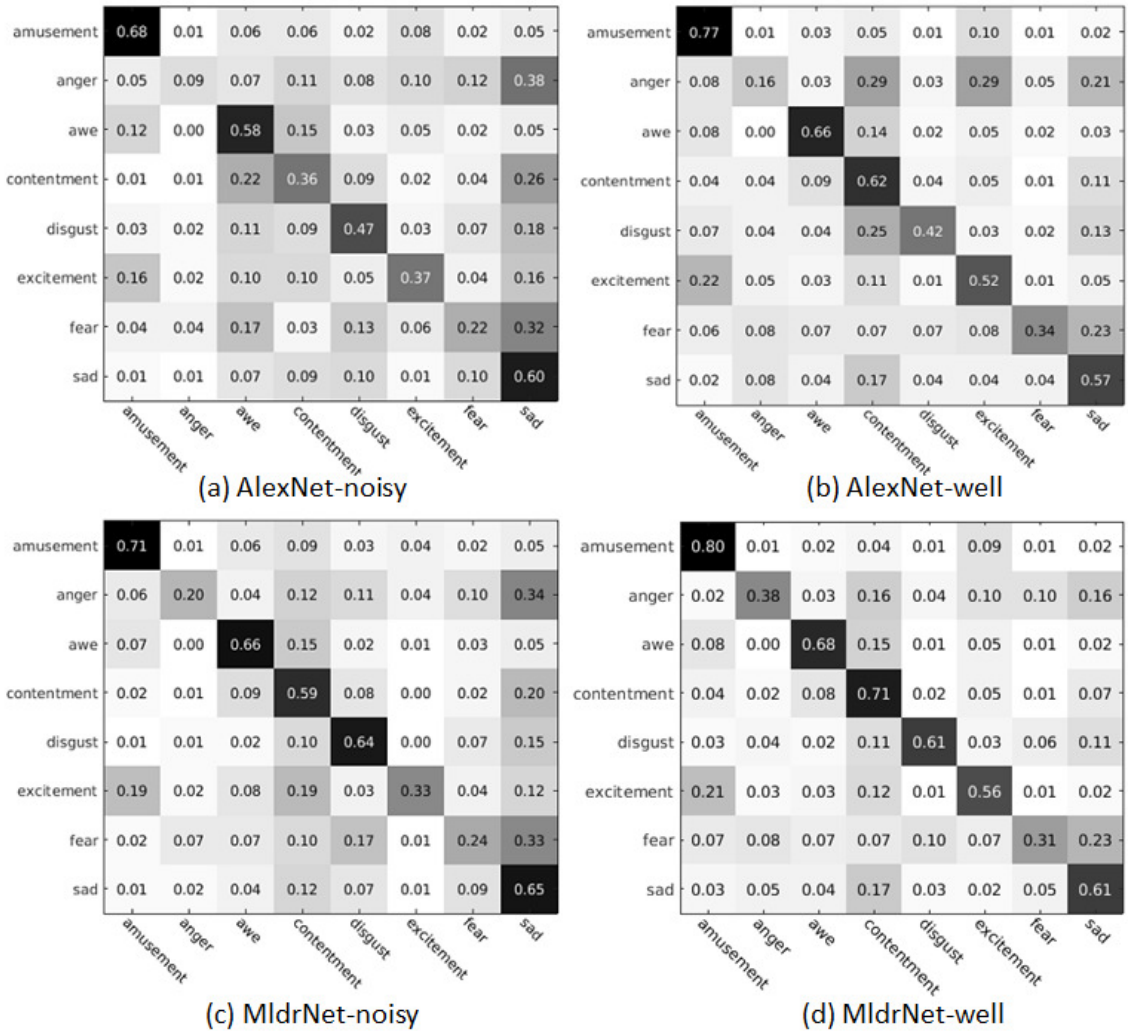

Figure 5.6 : Confusion matrices for AlexNet and the proposed MldrNet when using the *well* dataset and the *noisy* dataset as training dataset.

collected for the Internet, which makes the proposed method can be applied for many applications, such as, recommending system, social network and personalized advertising.

To further compared the proposed methods with AlexNet, the confusion matrix of the two methods on the testing dataset is reported. Considering the significant performance improvements by using deep representations compared to hand-crafted features, we only show the confusion matrices of the proposed MldrNet and AlexNet

using the *well* dataset as the training dataset (MldrNet-well and AlexNet-well) and the *noisy* dataset as the training dataset (MldrNet-noisy and AlexNet-noisy). As shown in Figure 5.6, the performances of AlexNet using both *well* and *noisy* as the training dataset in most emotional categories are lower than the proposed MldrNet. AlexNet tend to confuse some emotions, such as *fear* and *sad*. This indicates that image emotions cannot be clearly analyzed only relying on high-level image semantics. What's more, compared to Alexnet, the proposed MldrNet shows a more robust emotion classification result when using different training dataset. This means the proposed MldrNet can effectively extract emotional information even using a training dataset that contains false labels.

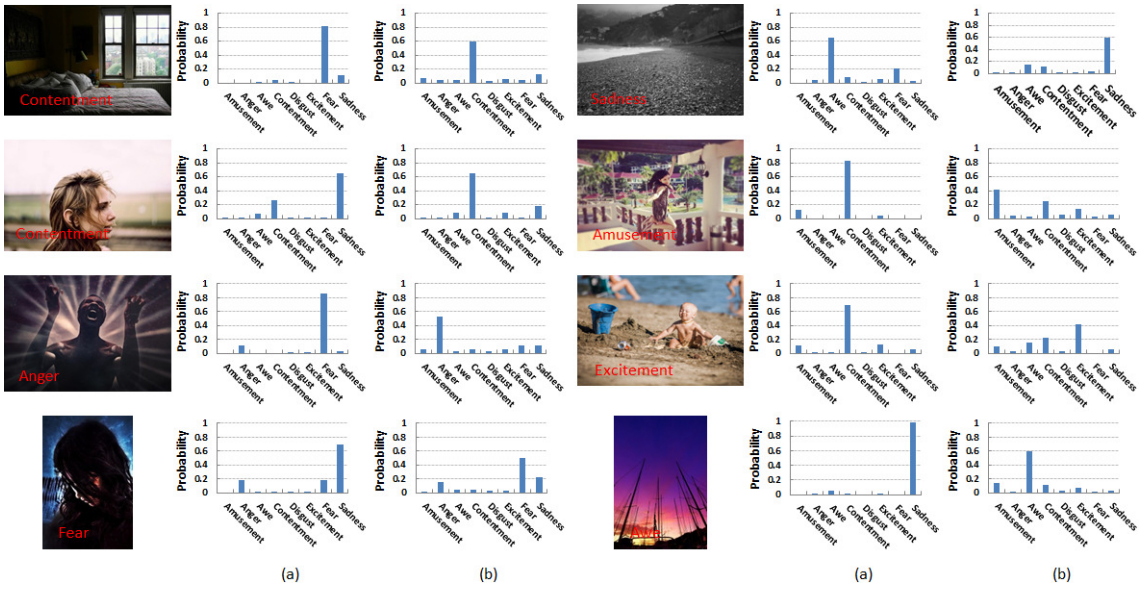

Figure 5.7 : Sample images correctly classified by the proposed MldrNet but misclassified by AlexNet. The column (a) shows the emotion distribution predicted by AlexNet and the column (b) shows the emotion distribution predicted by the proposed MldrNet. The red label on each image indicates the ground-truth emotion category.

A couple of sample images is also visualized that are correctly classified by the

proposed MldrNet but incorrectly classified by AlexNet to qualitatively analyze the influence of mid-level and low-level deep representations for image emotion classification. As shown in Figure 5.7, the emotions of the images misclassified by AlexNet are mainly conveyed by mid-level and low-level visual features, such as color, texture and image aesthetics. Combining the emotion information related to mid-level and low-level deep representations can significantly improve the emotion classification accuracy.

5.2.3 Emotion Classification on small Scale Datasets

Several image emotion analysis methods using hand-crafted features have been introduced. To better evaluate the effectiveness of MldrNet, the proposed method is compared with state-of-the-art methods based on hand-crafted features and Alexnet for each emotion categories.

Follow the same experimental settings described in (Machajdik and Hanbury, 2010*b*). Due to the imbalanced and limited number of images per emotion category, the "one against all" strategy is employed to train the classifier. The image samples from each category are randomly split into five batches and 5-fold cross validation strategy is used to evaluate the different methods. The images is used to train the last fully connected layer in the proposed MldrNet and AlexNet. Also, the *true positive rate per class* suggested in (Machajdik and Hanbury, 2010*b*) is calculated to compare the results. Note that in IAPS-Subset and Abstract dataset, only eight and three images are contained in the emotion category *anger*, so the 5-fold cross validation are unable to perform for this category. Therefore, the *true positive rate per class* of emotion category *anger* in these two datasets is not reported.

The emotion classification accuracies for each emotion categories are reported in Figure 5.8,5.9 and 5.10, respectively. For most of the emotion categories, deep learning methods significantly outperform the state-of-the-art hand-crafted meth-

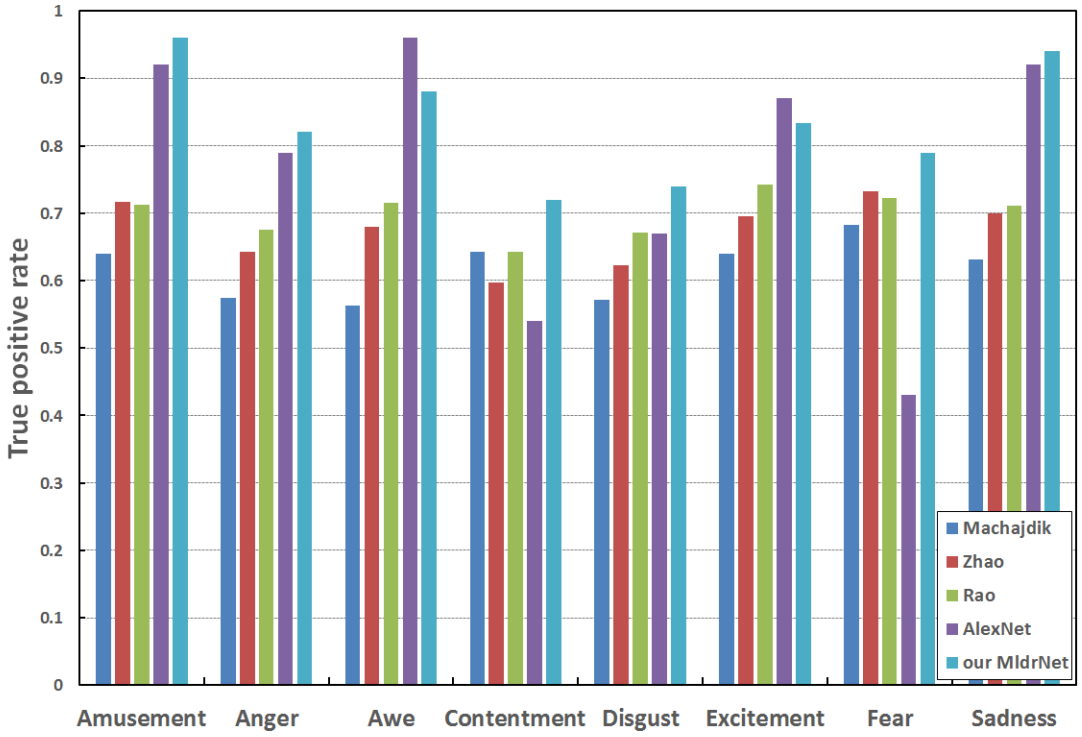

Figure 5.8 : Performance evaluation for each emotion categories on the ArtPhoto dataset.

ods. However, the performances of AlexNet in Abstract and ArtPhoto dataset are relatively low, this may because emotions of images in these two datasets are mainly conveyed by mid-level and low-level visual features. In contrast, MldrNet achieves the best performance in almost all emotion categories for the three dataset, which shows a robust result.

5.2.4 Emotion Classification on Abstract Paintings

To further evaluate the benefits of MldrNet. The proposed MldrNet on the MART dataset is also tested, which consists of abstract paintings. Followed the experimental approach in (Alameda-Pineda et al., 2016), 10-fold cross-validation is employed to compare the proposed MldrNet model with other six baseline methods

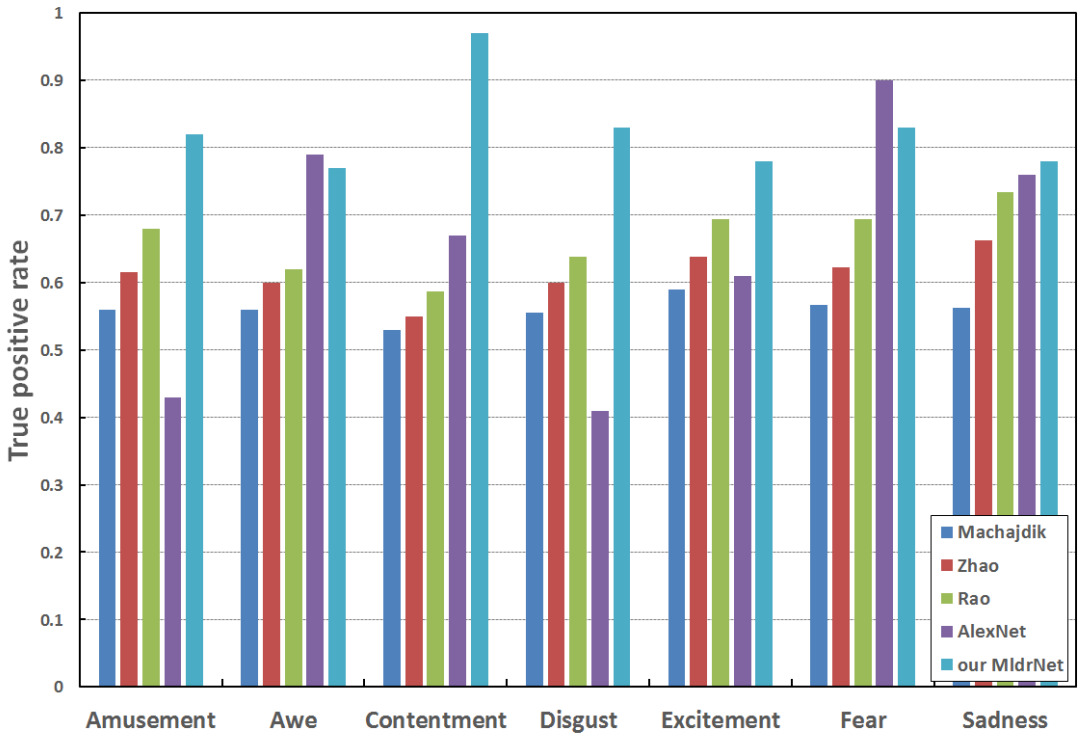

Figure 5.9 : Performance evaluation for each emotion categories on the Abstract dataset.

on the MART dataset. The baseline methods are: kernel transductive SVM (TSVM (Joachims, 1999)), linear matrix completion (LMC (Chen, Patel and Chellappa, 2015)), Lasso and Group Lasso both proposed in (Sartori et al., 2015), non-linear matrix completion (NLMC (Alameda-Pineda et al., 2016)) and AlexNet (Krizhevsky et al., 2012). The results shown in Table 5.4 demonstrate that the proposed MldrNet can effectively extract emotion information from abstract paintings when compared with all other methods. Compared to traditional CNN models, MldrNet Model is especially good at dealing with the image emotion related to low-level and mid-level visual features, i.e., color, texture and image aesthetics.

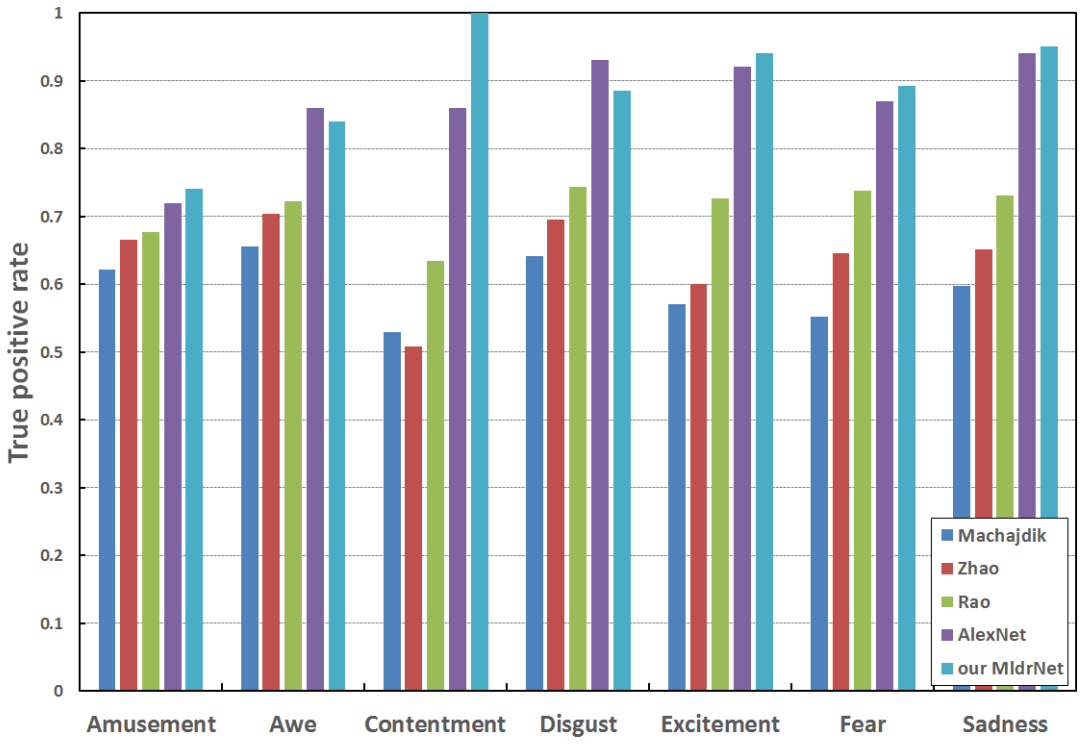

Figure 5.10 : Performance evaluation for each emotion categories on the IAPS-Subset.

5.3 RNN for Visual Emotion Recognition

Considering there exist strong correlations among different levels of features. For example, for middle-level features, such as textures, it is composed of low-level features, such as lines, and meanwhile, it leads to high-level features, such as the parts of objects. Such dependency among different levels of features benefits visual emotion recognition because different types of emotional stimuli in this task need to be considered. To this end, a new bidirectional model is proposed for feature fusion by exploiting the dependency among different levels of features. Experimental results justify and demonstrate the effectiveness of the proposed newly proposed Recurrent Neural Network (RNN) method for feature fusion.

Model	Accuracy
TSVM	69.2%
LMC	71.8%
Lasso	68.2%
Group Lasso	70.5%
NLMC	72.8%
AlexNet	69.8%
MldrNet	76.4%

Table 5.4 : Emotion classification accuracy of different methods on the MART dataset.

Specifically, a unified CNN-RNN framework for visual emotion recognition, which effectively learns different levels of features and integrates them by exploring the dependencies. As shown in Fig. 5.11, the total framework consists of two parts, i.e., feature extraction and feature fusion. The image features are extracted from multiple branches of CNN, which can represent different levels of features from the local view to global view. Considering the dependencies among different levels of features, a bidirectional RNN model consisting of the gated recurrent unit is proposed to capture this relationship and integrate different levels of features together. Finally, the fused features extracted from the proposed RNN model is concatenated to predict the emotion.

Recurrent neural networks can effectively model the long-term dependency on sequential data. It has been widely applied in many tasks including machine translation (Sutskever, Vinyals and Le, 2014), sequential modeling (Chung, Kastner, Dinh, Goel, Courville and Bengio, 2015), and so on. In this work, the results show that the

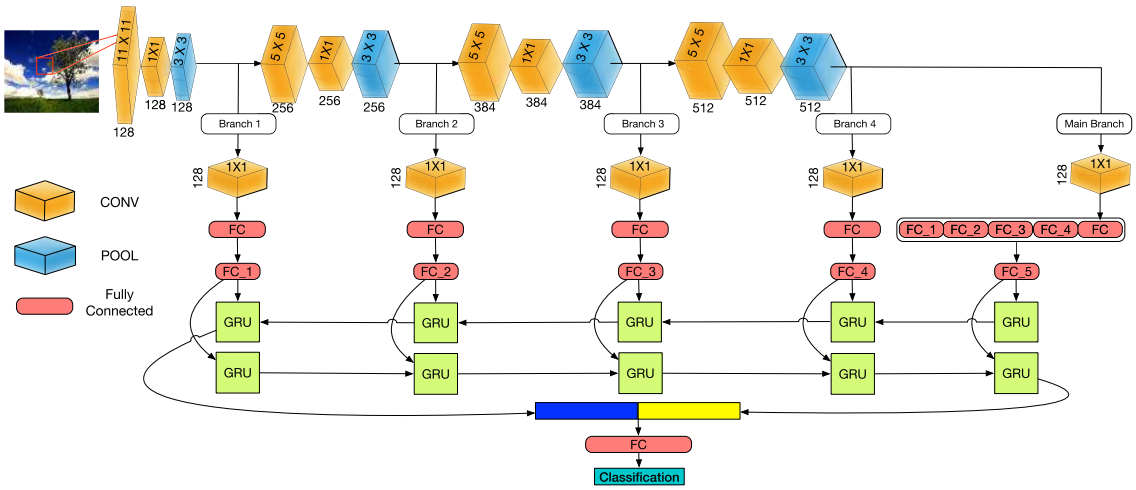

Figure 5.11 : The proposed unified CNN-RNN framework for visual emotion recognition. Different levels of features from multiple branches in the CNN modelis first extracted, which include low-level features (e.g. color, edge), middle-level features (e.g. texture) and high-level features (e.g. part, object). Then different levels of features flow into the proposed newly proposed Bidirectional Gated Recurrent Unit (GRU) model to integrate these features and exploit their dependencies. Two features generated from the proposed Bi-GRU model are concatenated as the final features to predict the emotion from images. (Best viewed in color.)

RNN can also exploit the relation between low-level features and high-level features.

5.3.1 RNN for Visual Emotion Recognition

In order to exploit the dependency among different levels of features, the features from the lower level to the higher level and from the higher level to the lower level are treated as two sets of sequential data and propose a new RNN based approach with bidirectional connections to better model such dependencies. Specifically, a new bidirectional GRU model is proposed to integrate the features from different levels as the proposed comprehensive experiments on the benchmark datasets show that the GRU method can achieve better performance than the LSTM method for

visual emotion recognition.

$\mathbf{V_t}$ is used to represent the visual features extracted from the branch at the time step t. Then the total pipeline in a GRU (illustrated in Fig. 5.12) at the time step t can be presented as follows:

$$r_t = \sigma(\mathrm{W}_{vr}\mathbf{V_t} + \mathrm{W}_{hr}h_{t-1} + b_r) \tag{5.5}$$

$$z_t = \sigma(\mathrm{W}_{vz}\mathbf{V_t} + \mathrm{W}_{hz}h_{t-1} + b_z) \tag{5.6}$$

$$\tilde{h}_t = \tanh(\mathrm{W}_{v\tilde{h}}\mathbf{V_t} + \mathrm{W}_{h\tilde{h}}(r_t \odot h_{t-1} + b_{\tilde{h}})) \tag{5.7}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{5.8}$$

where r_t, z_t, $\tilde{h}_t$, h_t are the reset gate, update gate, hidden candidate, and hidden state respectively. $\mathrm{W}_{[\cdot][\cdot]}$ are the weight matrices and $b_{[\cdot]}$ are the bias terms. In addition, σ stands for the sigmoid function in the proposed Bi-GRU and $\odot$ represents the element-wise multiplication. These gate mechanisms allow the GRU to capture information from local to global view and produce the output based on different levels of features.

Since the dependencies among different levels of features can be estimated from both local to global view and global to local view, the bidirectional GRUs is utilized that consist of a forward GRU and a backward GRU (illustrated in Fig. 5.12) to model the relationships from two different views, which follows the practical intuition. The final hidden states from the proposed bidirectional GRUs model are concatenated to be fed into the softmax classifier.

$$\mathbf{H} = [\overrightarrow{h_T}, \overleftarrow{h_T}] \tag{5.9}$$

The same loss, i.e. negative log-likelihood(NLL) function, is also used to train the proposed Bi-GRU model.

$$\mathcal{L} = L_{cls}(\mathbf{H}) + \lambda\|\theta\|_2 \tag{5.10}$$

where λ is the weight factor and θ represents the weight parameters in Bi-GRU.

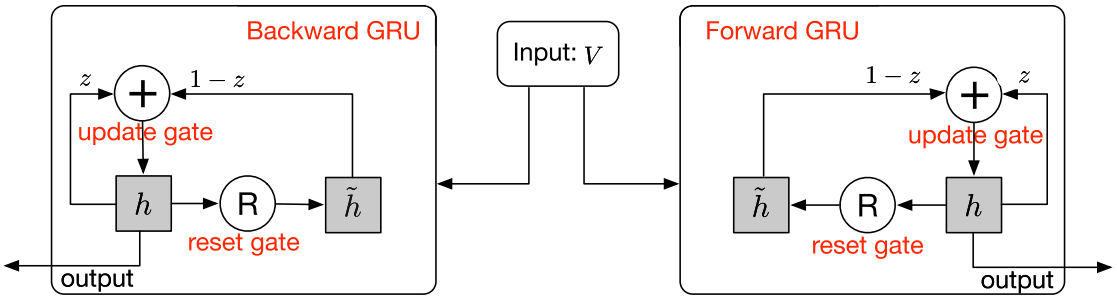

Figure 5.12 : The information flow of bidirectional gate recurrent unit. The bidirectional GRU consists of a forward GRU (right) and a backward GRU (left).

5.3.2 Experiment

The CNN part is same as mentioned before. For the RNN part, the Bi-GRU has 512 hidden units and the dropout on top of the output of Bi-GRU is applyed to avoid overfitting. $\lambda = 0.5$ in E.q (5.10) is set to balance the loss function and the regularization term, and set margin $\mu = 1$. The Bi-GRU is optimized by using Rmsprop (Tieleman and Hinton, 2012) with the learning rate as 0.0001.

To verify the contributions of the RNN part, different variants of the proposed model is designed as follows:

- **CNN+5 Branches+Ensemble:** It uses five branches (i.e. four branches and one main branch) to train five classifiers respectively, then predicts emotions according to the average category scores from five classifiers.
- **CNN+5 Branches+LSTM:** It uses five branches to extract the features at different levels and utilizes a unidirectional LSTM to integrate these features.
- **CNN+5 Branches+GRU:** It uses five branches to extract the features at different levels and utilizes a unidirectional GRU model to integrate these features.

Methods	Accuracy
Zhao	46.52%
Rao	51.67%
AlexNet+SVM	57.89%
ResNet-101	60.82%
MldrNet	67.24%
CNN+5 Branches+Ensemble	66.78%
CNN+5 Branches+LSTM	70.52%
CNN+5 Branches+Bi-LSTM	72.24%
CNN+5 Branches+GRU	71.33%
CNN+5 Branches+Bi-GRU	**73.03%**

Table 5.5 : Emotion classification accuracy of different methods on the large scale emotion dataset.

- **CNN+5 Branches+Bi-GRU:** It uses five branches to extract the features at different levels and utilizes the proposed bidirectional GRU model to integrate these features.

Experiments on Large Scale Emotion Dataset

The large scale emotion dataset is recently published in (You et al., 2016), which contains 8 different emotion categories including positive emotions: *Amusement*, *Awe*, *Contentment* and *Excitement* and negative emotions: *Anger*, *Disgust*, *Fear* and *Sad* as mentioned before.

The proposed proposed RNN based feature fusion method and its variants are compared with these baseline methods on the large-scale emotion dataset. The re-

sults are shown in Table 5.5. It has the following observations. First, the methods using deep representation outperform the methods using the hand-crafted features. Then these models based on different levels of features (MldrNet and the proposed method) outperform the methods using single-level features (ResNet and AlexNet). Moreover, the proposed feature fusion methods using the RNN (LSTM/GRU) significantly outperform all baseline methods.

More specifically, RNN based feature fusion methods with bidirectional GRU achieve better performance than methods using LSTM or unidirectional GRU, which demonstrates the bidirectional GRU plays a crucial role in improving the emotion classification performance.

To further demonstrate the performance of RNN, the confusion matrix is reported. As shown in Fig. 5.13, RNN based feature fusion method achieves a more balanced performance, especially for some negative emotions, such as *anger* and *fear*.

Experiments on Two Small Scale Datasets

Two small-scale datasets including ArtPhoto and IAPS as mentioned before are also widely used in previous works. For a fair comparison, the previous work (Machajdik and Hanbury, 2010*b*) is followed to evaluate different methods on the small scale emotion datasets, in which the same "one against all" strategy is used to train the emotion classifier. In addition, the proposed model is pre-trained using the large-scale emotion dataset and fine-tune the last fully connected layer by using the small-scale emotion datasets. The dataset is separated into the training and testing sets with K-fold cross-validation (K=5). The *true positive rate per class* is reported to evaluate different methods. Note that there are only eight images in the category *anger* in the IAPS-Subset dataset, it is unable to train a classifier for this category, and only the results over seven categories for this dataset are reported.

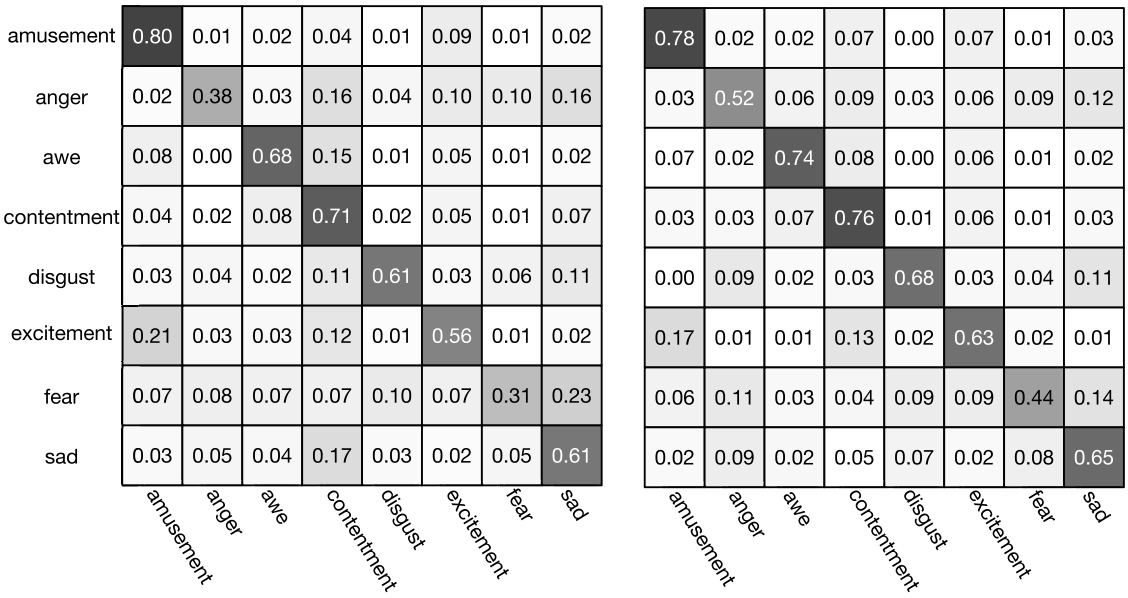

Figure 5.13 : The confusion matrix of MldrNet (left) and RNN based feature fusion method (right).

The RNN-based feature fusion method is compared with five aforementioned baseline methods on the two small-scale datasets. The results of the Artphoto dataset is presented in Fig. 5.14. Not only the true positive rate per class but also the average true positive rate is reported overall emotion categories in the last column. According to the results, the RNN-based feature fusion method is better on average than other methods, especially for some difficult emotions categories including *Fear* and *Sad*.

Fig. 5.15 shows the performances on the IAPS-Subset dataset. From the results, it can observed that the deep learning methods outperform methods using the hand-crafted features. The methods based on multiple levels of features generally outperform the methods which only utilize the high-level features. Furthermore, although the RNN-based feature fusion method does not achieve the best results in all emotion categories due to the limited number of training images, on average

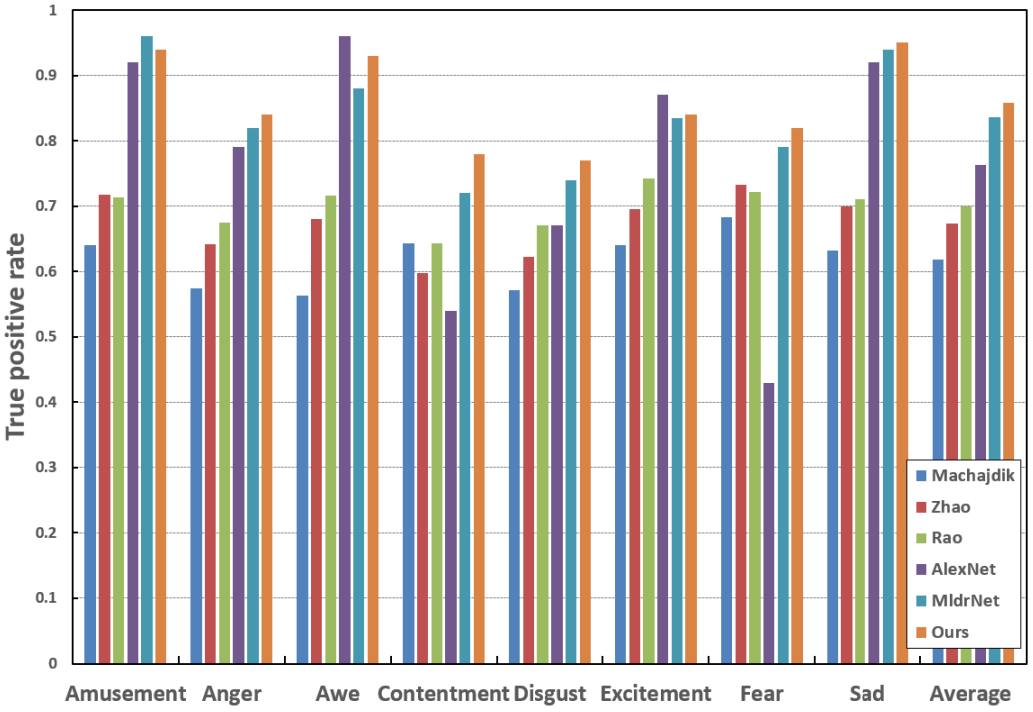

Figure 5.14 : Performance evaluation on the ArtPhoto dataset.

the method still achieves the best performance because the dependency for feature fusion is exploited.

5.4 Discussions

In this work, a new network that learns multi-level deep representations is proposed for image emotion classification. It has been demonstrated that image emotion is not only affected by high-level image semantics but also related mid-level and low-level visual features, such as color, texture and image aesthetics. the proposed MldrNet successfully combine the deep representations extracted from the different layer of the deep convolutional network for image emotion classification. In the experiments, MldrNet achieves consistent improvement in image emotion classification accuracy with fewer convolutional layers compared to popular CNN models

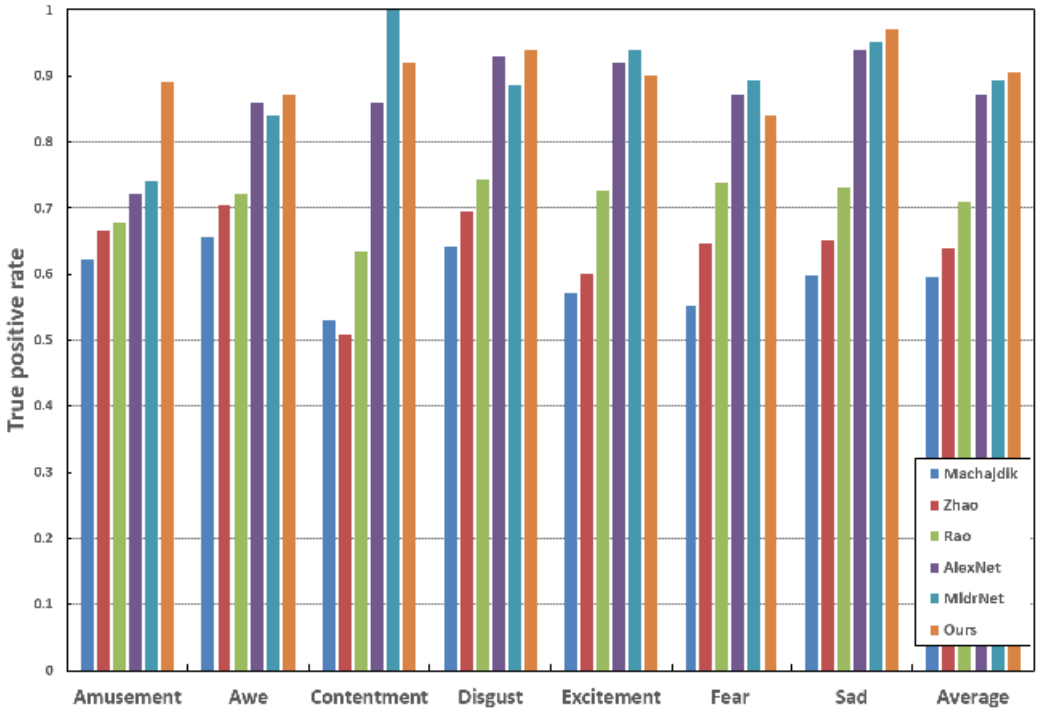

Figure 5.15 : Performance evaluation on the IAPS-Subset.

for different kind of image emotion datasets. Moreover, MldNet shows more robust results when using different training dataset, especially the *noisy* dataset directly collected from the Internet. This will decrease the demand for reliable training data, which will help us to utilize huge amount of images. An unified CNN-RNN model is also proposed for visual emotion recognition, which leverages different levels of features from multiple branches in CNN and effectively integrates these features by exploiting the dependencies among them with the bidirectional GRU approach.

Compared to linear deep convolutional neural network models, MldrNet model combining with deep representations extracted from different convolutional layers are better at dealing with abstract-level computer vision tasks, i.e. image emotion classification, image aesthetics analysis and photo quality assessment.

Chapter 6

Multi-level Region-based Convolutional Neural Network for Image Emotion Classification

In this work, a multi-level region-based convolutional neural network is proposed that can automatically extract multi-level deep representations of local image regions. Usually, an image may contain two types of regions: object region and emotional region. Object region is the region that only contains a specific object and emotional region is the region that only contains a specific type pf emotion. Multi-level deep features can better represent different kinds of affective images and utilizing features extracted from emotional regions can effectively avoid the noisy information containing in non-emotion regions. Moreover, a new loss function is proposed in this paper to estimate an emotion distribution derived from emotion class probability, which can effectively counteract the factor of subjectivity existing in image emotion labels. The overview of the proposed framework is shown in Figure 6.1. Emotional regions of different size are extracted based on different scales of feature maps which combine multi-level deep features. Subsequently, the local deep representations extracted from these emotional regions are combined with the global deep representations extracted from the whole image for emotion classification. Compared with existing methods mainly based on single-level visual features from a global view, the multi-level emotion information from both global and local view utilized in the proposed model can provide a robust performance on various kinds of images.

The contributions of this work are summarized as follows:

- A feature pyramid network(FPN) is employed to extract multi-scale deep feature maps that related to image emotion. The multi-scale deep feature maps extracted from different convolutional layers can combine high-level semantic features with low-level deep features, and thus significantly improve the performance of emotion region detection.

- A region-based CNN model is built that can effectively extract local emotional information from the emotional regions of the image. Ignoring the noisy information generating from non-emotional regions can significantly improve the emotion classification performance.

- Image emotion labeling is a highly subjective task and the uncertain emotion labels will degrade the classification accuracy. Thus, the loss function is mmodified to consider the emotion class probability, rather than a hard class label, into image emotion classification to overcome the subjectivity in emotion analysis.

Extensive experiments are conducted to evaluate the proposed Multi-level R-CNN model on multiple datasets including Flickr&Instagram(FI) (You et al., 2016), IAPSsubset (Machajdik and Hanbury, 2010*b*), ArtPhoto (Machajdik and Hanbury, 2010*b*), *etc.* The experimental results demonstrate the effectiveness of the proposed method for effectively detecting emotional regions with multi-level deep features and dealing with the problem of subjectivity existing in image emotion.

6.1 Preliminaries

As shown in Figure 6.1(a), candidates of emotional regions with multi-level deep features are extracted using faster R-CNN based on FPN.

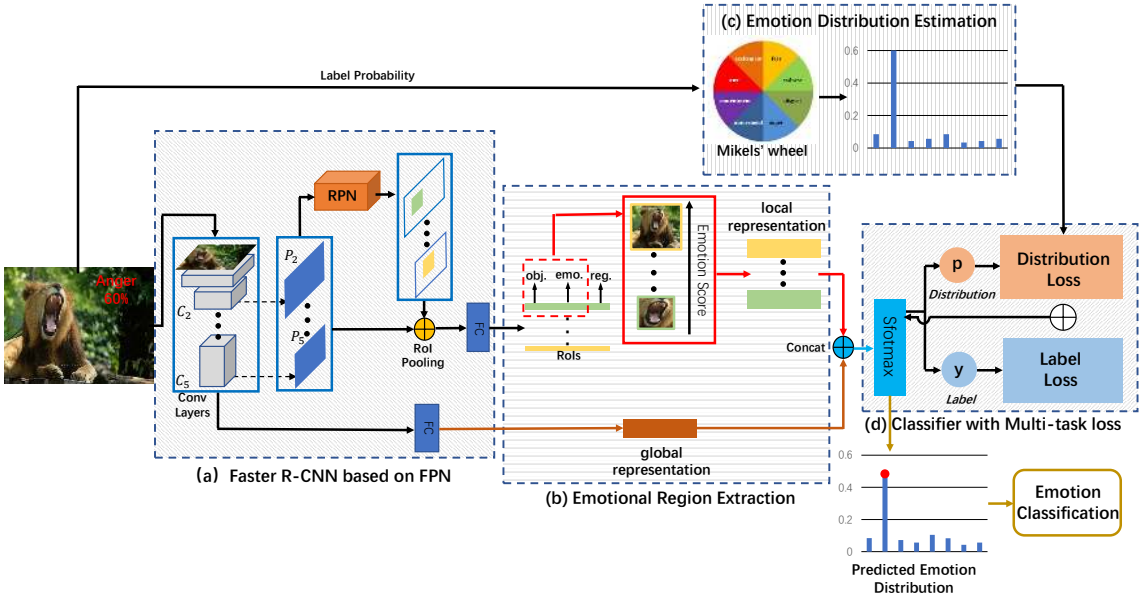

Figure 6.1 : The overview of the proposed framework. The framework consists 4 components:(a) faster R-CNN based on FPN, (b) emotional region extraction based, (c) emotion distribution estimation and (d) classifier with multi-task loss

6.1.1 Feature Pyramid Network (FPN)

To extract multi-level deep representations for image emotion analysis, a Feature Pyramid Network (FPN) (Lin, Dollár, Girshick, He, Hariharan and Belongie, 2017) is employed to extract multi-scale feature maps. Compared to the existing pyramidal feature hierarchy structure in (Rao, Xu and Xu, 2016; Zhu et al., 2017), in which the lower level feature maps are high-resolution but with low-level deep features that harm their representational capacity for object recognition and emotion classification. The detailed structure of FPN is shown in Figure 6.2. As shown in the figure, FPN consists of two parts, a bottom-up pathway and a top-down pathway, between them is the lateral connections.

The bottom-up pathway is the feed-forward computation of normal backbone convolutional network (e.g., (Krizhevsky et al., 2012; Simonyan and Zisserman,

2014; He et al., 2016)). In this work, ResNet101 (He et al., 2016) is used as the backbone network. From the bottom-up pathway, feature hierarchy which contains feature maps of different size can be computed. The output of the last layer of each bottleneck in the ResNet101 is selected as the reference set of feature maps to create the pyramid. The output of these bottlenecks are defined as $\{C_2, C_3, C_4, C_5\}$ for conv2, conv3, conv4 and conv5 outputs, and note that conv1 is excluded for the pyramid due to the large memory consuming for the massive feature map.

The top-down pathway is used to combine different levels of feature maps extracted from the bottom-up pathway. Feature map from the highest pyramid level, which is semantically stronger but spatially coarser, is upsampled to fit the size of lower-level feature maps in feature pyramid, which are higher resolution but only contains low-level deep features. The upsampled map is then merged with the corresponding bottom-up maps (a 1×1 convolutional layer is added behind each bottom-up map to reduce channel dimensions) by element-wise addition. The process is iterated until the last (finest resolution) merged map is generated. The set of final feature maps is defined as $\{\{P_2, P_3, P_4, P_5\}$, which is corresponding to $\{C_2, C_3, C_4, C_5\}$ in the same spatial size respectively.

6.1.2 Faster R-CNN

Detecting concrete visual objects in images has been widely studied in computer vision (Redmon, Divvala, Girshick and Farhadi, 2016; Liu, Anguelov, Erhan, Szegedy, Reed, Fu and Berg, 2016; Ren et al., 2015). In this work, Faster R-CNN model (Ren et al., 2015) is used to extract emotional region from the image. Faster R-CNN is a two-stage detector mainly consisting of three major parts: shared bottom convolutional layers which is FPN in the proposed model, a region proposal network (RPN) and a classifier built for region-of-interest (ROI). The detailed structure is shown in the left part of Figure 6.1.

First, an input image is represented as multi-scale feature maps which combine different levels of deep features by FPN. Then, RPN generates candidate object proposals based on the feature maps. Since the single-scale feature map using in Faster R-CNN is replaced with multi-scale feature maps, single-scale anchors with size $\{32^2, 64^2, 128^2, 256^2\}$ pixels are applied for multi-level feature maps $\{P_2, P_3, P_4, P_5\}$ with different receptive fields respectively. Finally, ROI-pooling is used to extract features representing ROI and ROI-wise classifier predicts the category label based on the features. The training loss is composed of two terms:

$$L_{det} = L_{obj} + L_{reg} \tag{6.1}$$

here L_{obj} is the classification loss over two class (if the candidate region contains an object or not). $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$ is the regression loss on the box coordinates for better localization, in which t_i is a 4-d vector representing the coordinates of the predicted bounding box, t_i^* is the coordinates of the ground-truth box and R is the robust loss function(L1 smooth) defined in (Girshick, 2015). More detailed information about the architecture and training procedure of Faster R-CNN can be found in (Ren et al., 2015).

6.2 Emotion Analysis using Multi-level R-CNN

6.2.1 Emotional Region Extraction

Detecting concrete visual objects in images has been widely studied in computer vision (Redmon et al., 2016; Liu, Anguelov, Erhan, Szegedy, Reed, Fu and Berg, 2016; Ren et al., 2015). However, compared to object detection, detecting emotional content is extremely challenging. The main difficulty is that both the concrete objects and the surrounding background contribute to image emotions (Chen, Yu, Chen, Cui, Chen and Chang, 2014; Sun, Yang, Wang and Shen, 2016). Due

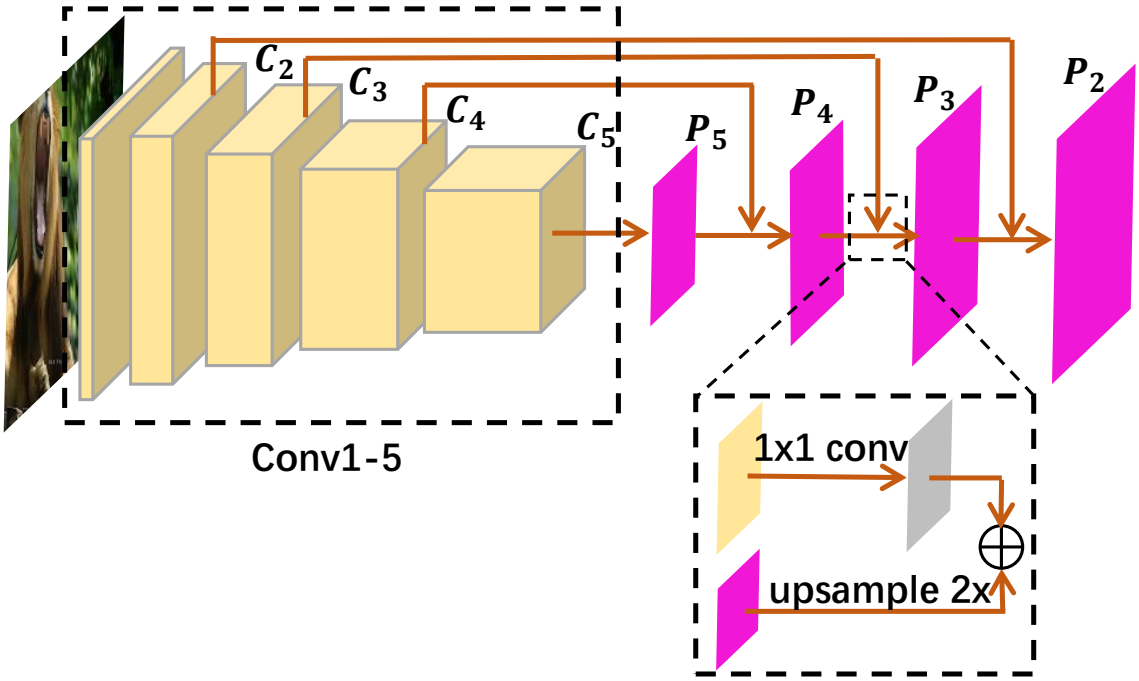

Figure 6.2 : Structure of Feature Pyramid Network (FPN).

to the strong co-occurrence relationships between objects and local emotional regions (Wang et al., 2016), object detection methods could still be utilized to select potential emotional regions. However, to select proper emotional regions from the candidate object proposals generating through RPN, the Faster R-CNN is modified to contain emotional information as shown in Figure 6.1(b).

Following the definition of objectness score S_{obj} in (Ren et al., 2015), which measures the membership to set of object classes *vs.* background, emotion score S_{emo} is defined to evaluate the probability of a region evoking emotions. To compute the emotion score, a binary class label (of the emotional region or not) is assigned to each anchor in RPN. The positive label is assigned to an anchor with the highest Intersection-over-Union (IoU) overlap with a ground-truth emotional region or an anchor that has an IoU overlap higher than 0.7 with any ground-truth emotional region. The negative label is assigned to anchor with IoU overlap lower than 0.3

with any ground-truth emotional regions. Using the samples collected from RPN, a softmax classifier can be trained to predict to probability p_{emo} of the region evoking emotions.

In Faster R-CNN, the emotional region classifier is introduced into the ROI-wise classifier and fix the object classifier. The new training loss function is:

$$L^*_{det} = L_{emo} + L_{reg} \tag{6.2}$$

where L_{emo} is the classification loss of the RoI being an emotional region or not. Therefore, the RoI-wise classifier can compute both the probability of the RoI evoking emotions p_{emo} and the probability of the ROI containing an object p_{obj}. The Faster R-CNN can train on the loss L^*_{det} related to emotional regions.

As mentioned before, considering the emotional region is related to the probability of the region containing an object p_{obj} and the probability of the region evoking emotions p_{emo}, the emotion score of the region can be computed considering both probabilities:

$$S_{emo} = \sqrt{p^2_{emo} + p^2_{obj}} \tag{6.3}$$

The proposed emotion score S_{emo} can reflect how likely a region evoking emotions. The 10 regions with the highest emotion score are selected as the emotional regions of the image and used for image emotion classification.

6.2.2 Emotion Distribution Estimation

The majority voting strategy is widely employed to obtain the ground truth emotional label for most of affective image datasets (Machajdik and Hanbury, 2010*b*; You et al., 2016). Many images in these datasets have emotional labels with probabilities instead of hard emotional labels. To handle the impact of labeling image emotions

with emotion class probabilities, either estimate an emotion distribution based on label probabilities or directly import label probabilities into a loss function for training is considered. For emotion distribution estimation, since the subjectivity existing in humans' emotional response to images, the emotional response to an image is more likely a distribution of several emotions rather than a single emotion.

Inspired by the study of emotion theory (Plutchik, 2001), the degree of similarity between two emotions, which determines the relationship of the two emotions, from similar to complete opposite, can be represented through Mikels' Wheel (Zhao, Yao, Gao, Ji and Ding, 2017). Figure 6.3 shows Mikels' wheel and the method to compute emotion distance revealing the similarity between two emotions. The low distance d_{ij} between emotion i and emotion j indicates that the two emotions are similar to each other. Using Mikels' Wheel as weak prior knowledge, the probability of different emotion classes can be assigned to an image based on the dominant emotion class of that image. Therefore, if the image has a dominant emotion j with probability p_j^*, the emotion distribution for the i-th emotion of the image can be generated through triangular distribution as shown in Figure 6.1(c):

$$f(i) = \begin{cases} p_j^* & i = j \\ \\ \frac{\frac{1}{dis_{ij}}(1-p_j)}{\sum_{i \neq j} \frac{1}{dis_{ij}}} & i \neq j \end{cases} \tag{6.4}$$

in which, the emotion classes being closer to the dominant emotion class are assigned with higher probabilities. The sum of all emotion class probabilities $\sum f(i)$ is normalized to 1.

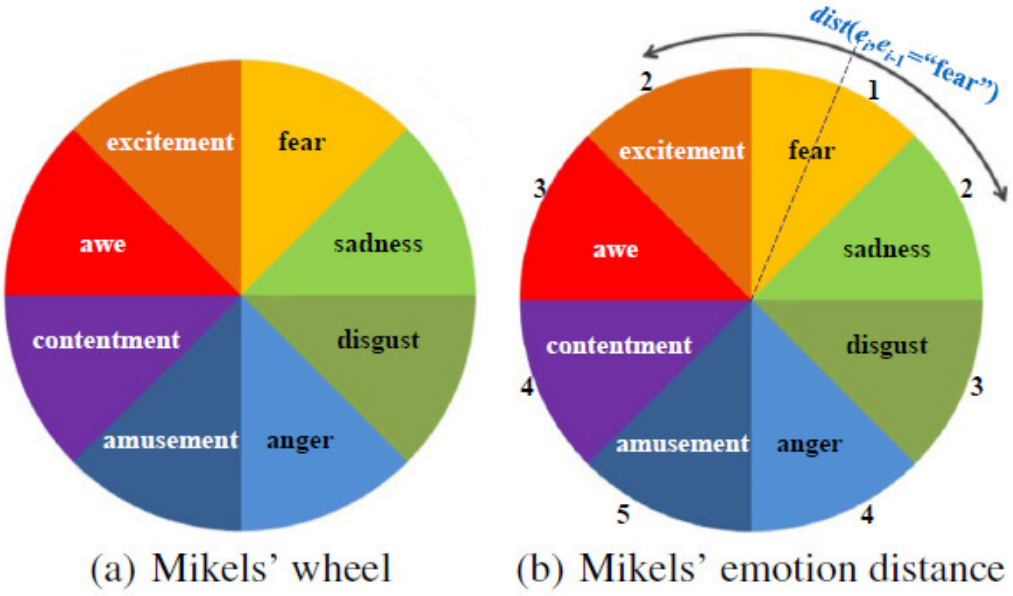

Figure 6.3 : Mikels' emotion wheel and example of emotion distance.

6.2.3 Classifier and Loss Function

Through Faster R-CNN, a set of local deep representations of emotional regions is collected $\{X_{local}\}_{j=1}^{K}$, where K is the number of emotional regions extracted from one image. Considering that an image may not contain many local emotion regions, only top-ranked major emotion regions is utilized for classification, by setting $K = 10$. The global deep representation of the whole image X_{global} extracted from the ResNet101 is concatenated with the local deep representations $\{X_{local}\}_{j=1}^{K}$:

$$X = [X_{global}, \{X_{local}\}_{j=1}^{K}] \tag{6.5}$$

.

Followed by a softmax layer, X is transformed into a probability distribution of different emotions, where the emotion category with the highest probability is considered as the predicted label of the image. Considering the two methods (as

discussed in Section 6.2.2) deal with the subjectivity existing in emotions, two loss functions can be applied in the proposed approach.

Multi-task Loss: Taking both emotional label and estimated emotion distribution into account, the multi-task loss function consists of two terms:

$$L_{multi} = (1 - \lambda)L_{cls} + \lambda L_{ed} \tag{6.6}$$

where L_{cls} is the traditional classification loss, which can be computed as:

$$L_{cls} = -\sum_{i} y_i \log(p_i) \tag{6.7}$$

where $y = \{y_i | y_i \in \{0, 1\}, i = 1, ..., n, \sum_{i=1}^{n} y_i = 1\}$ indicates the ground-truth label of the image, and p_i is the probability of an image belonging to the ith emotion category.

$L_{ed}$is the loss from emotion distribution $f(i)$. The KL loss defined in (Gao, Xing, Xie, Wu and Geng, 2017) is employed. λ controls the trade-off between the two weights. The KL loss is the measurement of the similarity between the emotion distribution $f(i)$ and the predicted emotion distribution p_i:

$$L_{ed} = -\sum_{i} f(i) \log(p_i) \tag{6.8}$$

The loss function can be optimized by stochastic gradient descent(SGD). $\{a_i | i = 1, 2, ..., N\}$ is defined to be the activation values of class i in the last fully connected layer. The gradient can be computed by:

$$\begin{aligned} \frac{\partial L}{\partial a_i} =& (1 - \lambda) \sum_{i} \frac{\partial L_{cls}}{\partial p_i} \frac{\partial p_i}{\partial a_i} + \lambda \sum_{i} \frac{\partial L_{ed}}{\partial p_i} \frac{\partial p_i}{\partial a_i} \\ &= p_i + (1 - \lambda) y_i + \lambda f(i) \end{aligned} \tag{6.9}$$

Loss with Probability: Another instinctive thought to deal with label with probability is to directly introduce label probability into loss function. Similar to (Gal, 2016), the classification loss with probability L_p can be defined as:

$$L_p = -\sum_i y_i \log(p_i^\theta), p_i^\theta = \frac{\exp(p_j^{*2} \cdot p_i)}{\sum_i \exp(p_j^{*2} \cdot p_i)} \tag{6.10}$$

The class prediction is weighted by the label probability p_j^*. By introducing the explicit simplifying assumption $p_j^* \sum_i \exp(p_j^{*2} \cdot p_i) \approx (\sum_i \exp(p_i))^{p_j^*}$ which becomes equal when $p_j^* \to 1$, the classification loss with probability can be rewritten as:

$$\begin{aligned} L_p &= -\sum_i \log(\exp(p_j^{*2} \cdot p_i)) + \log(\sum_i \exp(p_j^{*2} \cdot p_i)) \\ &\approx -p_j^{*2} \sum_i y_i \log(p_i^\theta) + \log(\frac{1}{p_j^{*2}}) \end{aligned} \tag{6.11}$$

The label with lower probability P_j^* will reduce the contribution of the classification loss. With the above equation, label probability is introduced into the loss for training.

6.3 Experiments and Results

In this section, the proposed model is evaluated against state-of-the-art emotion classification methods through comprehensive experiments to demonstrate the effectiveness of the proposed framework for different emotion classification tasks.

6.3.1 Dataset

Experiments are carried out on normally used image emotion datasets:

Flickr and Instagram (FI)(8 categories)(You et al., 2016): This dataset is collected from social websites using the names of emotion categories as search keywords. Workers from Amazon Mechanical Turk (AMT) are then hired to further label the images. Finally, 23,308 well-labeled images are collected for emotion recognition.

EmotionRoI(2 categories)(Peng et al., 2016): The dataset contains 1,980 affective images from Flickr with labeled emotional regions. This dataset can be used for training the R-CNN.

IAPSsubset(8 categories)(Machajdik and Hanbury, 2010*b*): The *International Affective Picture System*(IAPS) is a standard stimulus image set, which has been widely used in affective image classification. IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, landscapes and others. Among all IAPS images, Mikels et al. (Mikels et al., 2005*b*) selected 395 images and mapped arousal and valence values of these images to the above mentioned eight discrete emotion categories.

ArtPhoto(8 categories)(Machajdik and Hanbury, 2010*b*): In the ArtPhoto dataset, 806 photos are selected from some art sharing sites by using the names of emotion categories as the search terms. The artists, who take the photos and upload them to the websites, determine emotion categories of the photos. The artists try to evoke a certain emotion for the viewers of the photo through the conscious manipulation of the emotional objects, lighting, colors, etc. In this dataset, each image is assigned to one of the eight aforementioned emotion categories.

Abstract(8 categories)(Machajdik and Hanbury, 2010*b*): This dataset consists of 228 abstract paintings. Unlike the images in the IAPS-Subset and ArtPhoto dataset, the images in the Abstract dataset represent the emotions through overall color and texture, instead of some emotional objects. In this dataset, each painting was voted by 14 different people to decide its emotion category. The emotion category with the most votes was selected as the emotion category of that image.

6.3.2 Implementation Details

The backbone network of the proposed model is the FPN (Lin et al., 2017). In this work, a two-step training strategy is used. At the first step, the same strategy in

(Lin et al., 2017) is used to fine-tune the multi-level R-CNN on COCO pre-trained weights using the EmotionROI dataset. Note, the aspect ratio of an anchor is set to {1:1}. At the second step, the learning rate of the last two fully-connected layers is initialized as 0.001 and fine-tuned by SGD. The batch size is 128 and a total of 100 epochs are run to update the parameters. All the experiments are carried out on four NVIDIA GTX 1080 GPUs with 32GB of GPU memory.

6.3.3 Baseline

The proposed framework is compared with the state-of-the-art methods for image emotion classification, which use various features, including hand-crafted features and deep features.

Hand-crafted features

- **GCH/LCH/GCH+BoW/LCH+BoW**(Siersdorfer et al., 2010): 64-bin color histogram features for global view(GCH) and local view(LCH), and with SIFT-based bag-of-words features.

- **Zhao**(Zhao et al., 2014*b*): low-level and mid-level features based on principle of art.

- **Rao(a)**(Rao, Xu, Liu, Wang and Burnett, 2016): SIFT-based bag-of-visual features for both global and local view based on the image blocks extracted from images.

- **SentiBank**(Borth, Ji, Chen, Breuel and Chang, 2013): 1200-dim adjective noun pairs(ANPs) features as mid-level representation with linear SVM classifier.

Deep features

- **AlexNet**(Krizhevsky et al., 2012): AlexNet fine-tuned on ImageNet pre-trained weights.
- **VGG-16**(Simonyan and Zisserman, 2014): VGGNet fine-tuned on ImageNet pre-trained weights.
- **ResNet101**(He et al., 2016): ResNet Fine-tuned on ImageNet pre-trained weights.
- **DeepSentiBank**(Chen, Borth, Darrell and Chang, 2014): 2,089-dim ANPs features based on CNN.
- **PCNN**(You, Luo, Jin and Yang, 2015): a novel progressive CNN architecture based on VGGNet (Simonyan and Zisserman, 2014).
- **Rao(b)**(Rao, Xu and Xu, 2016): a CNN architecture based on AlexNet with side branch to utilize multi-level deep features.
- **Zhu**(Zhu et al., 2017): a unified CNN-RNN architecture for visual emotion recognition.

6.3.4 Experimental Validation

For methods using deep features, they are first fine-tuned on the large scale dataset(**FI**). The **FI** dataset is split randomly into 80% training, 5% validation and 15% testing sets. For the 4 dataset(**FI, IAPSsubset, ArtPhoto, Abstract**), with 8 emotional categories(positive emotion *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotion *Anger*, *Disgust*, *Fear*, *Sadness*), they can be converted to 2 emotional categories with labeling 4 positive emotions as positive and 4 negative emotions as negative. To compare the results for all datasets, the classification results is presented for both 8 emotional categories and 2 emotional categories.

The effectiveness of local emotional region:

To demonstrate the effectiveness of considering the proposed local emotional regions. Experiments are designed to performe on the **FI** dataset to compare: 1) ResNet101(He et al., 2016) only using the global feature extracted from the last convolutional layer; 2) the proposed framework only with features extracted from object regions extracted using Faster R-CNN with FPN(Lin et al., 2017). 3) the proposed framework only with features extracted from emotional regions; 4) the proposed framework with object regions extracted using Faster R-CNN with FPN; 5) the proposed framework with features extracted from both the whole image and emotion regions. Table **??** shows the performance of the five different methods on the test set of **FI**. As shown in table 6.1, compared to ResNet101, the proposed method with object regions improves the performance by 7.65% for 8 classes **FI** and 7.08% for 2 classes **FI** and the proposed method with emotional regions improves the performance by 14.64% and 12.84%. This result reveals that emotional information from local regions can largely improve the emotion classification accuracy than a single-column CNN-based global feature extracted from the whole image. However, the emotion recognition performances reduced significantly without using features extracted from the global view. This demonstrates the effectiveness of the global features extracted from the whole image.

Although both the proposed framework with object regions and that with emotion regions improve the emotion classification performance, it is clear that using emotional regions in the proposed method outperforms than using object regions by 6.99% and 5.76% for 8 classes **FI** and 2 classes **FI** respectively. The results of the proposed methods only using local features extracted from object regions and emotional regions also indicate that emotional regions contain more emotional information than object regions

Table 6.1 : Classification accuracy for both 8 classes and 2 classes on the test set of **FI**. The proposed method with different configurations, *i.e.*, combing with object region and emotional region is compared with single column ResNet101 without local information and using object region and emotional region as local information only.

Method	**FI**(8 classes)	**FI**(2 classes)
ResNet101	60.82%	74.67%
object regions only	54.82%	88.44%
emotional regions only	59.78%	72.57%
The proposed method+object regions	68.47%	81.75%
The proposed method+emotional regions	**75.46**%	**87.51**%

Figure 6.4 shows examples of object regions and emotion regions. It can be found that emotional regions are larger than object regions by containing objects and the surrounding background which may evoke emotions.

In Figure 6.5, the confusion matrix of ResNet101 and the proposed method with different configuration is reported. It is clear that applying local information in image emotion classification can improve the performance and provide a more balanced classification result for each emotion category. Especially applying emotional regions as local information in the proposed method achieves the best classification result on most of the emotion categories. This also demonstrates the effectiveness of the emotional region.

The effectiveness of multi-level features:

Previous methods have already had already indicated that multi-level features can significantly improve the image emotion classification performances (Rao, Xu

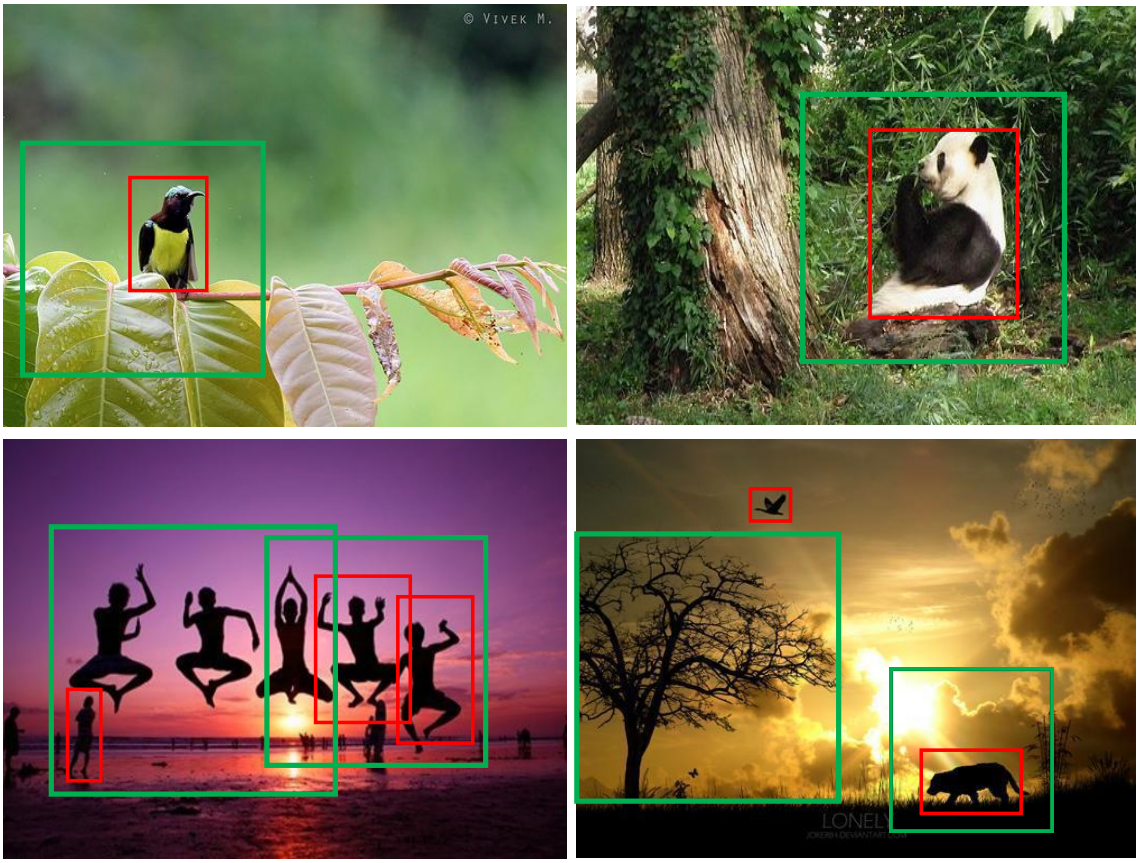

Figure 6.4 : Examples of object regions with highest objectness scores(red bounding box) and emotional regions with highest emotion probability(green bounding box).

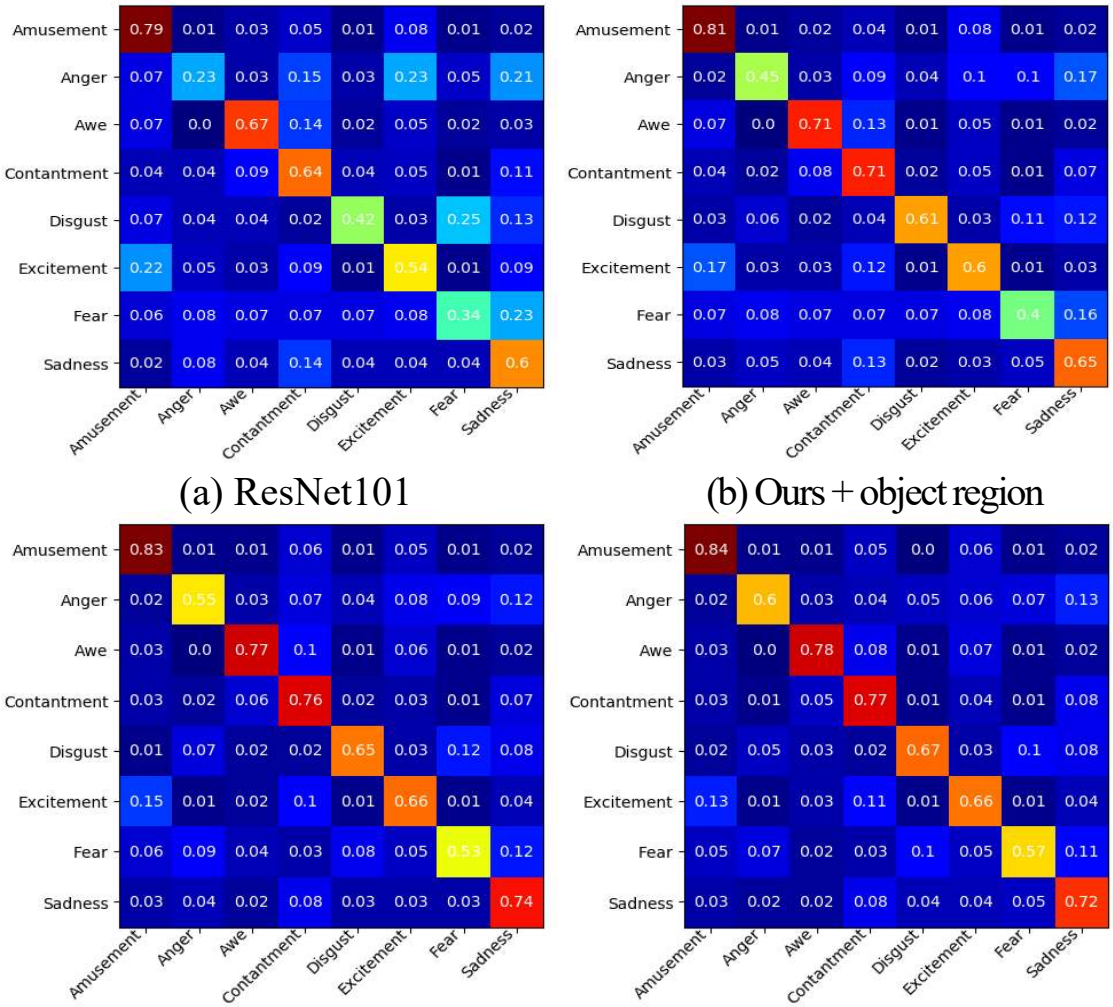

Figure 6.5 : Confusion matrix for the proposed method with different configurations and ResNet101.

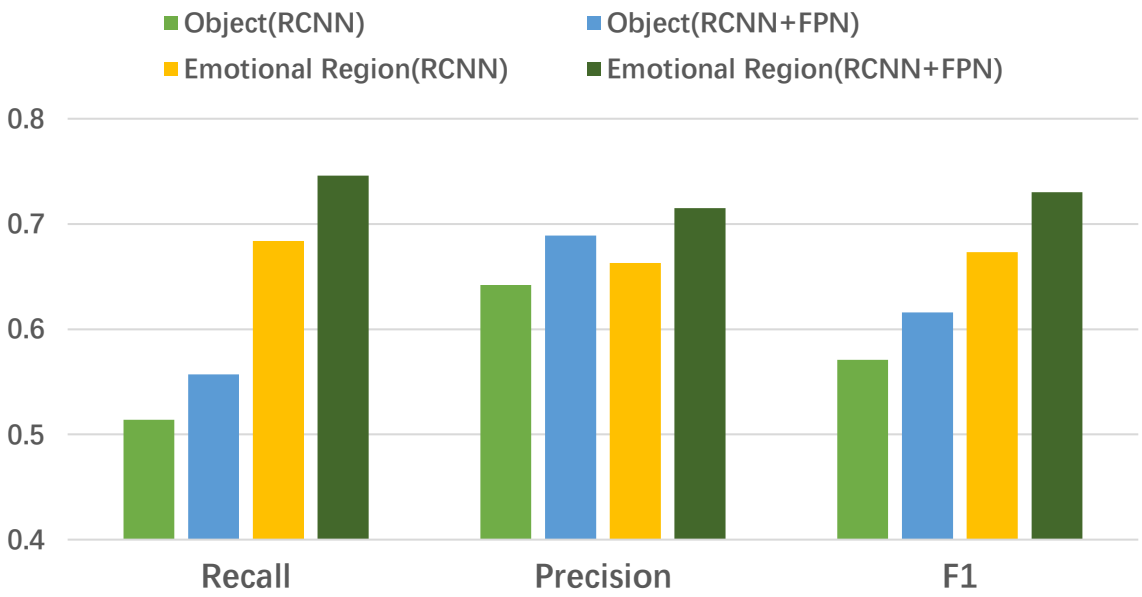

Figure 6.6 : Comparison of Emotional region detection performance on the test set of EmotionROI dataset using object detection methods and emotional region detection methods with single level features and multi-level features.

and Xu, 2016; Zhu et al., 2017). However, the effectiveness of multi-level features in emotional region detection still need to be proved. Figure 6.6 performs the detection results on the test set of EmotionROI dataset. It can be noticed that multi-level features improve both performances of object region detection and emotional region detection. The reason is that the multi-level framework provides features maps with different scales of respective fields, which can effectively detect objects with different size in an image. What's more, the multi-level features can improve the accuracy of predicting emotional score (Rao, Xu and Xu, 2016), which can further promote the emotional region detection performance.

Choice of the loss functions:

As discussed earlier, the subjectivity of the emotion is one of the main challenges for visual emotion recognition. Compared to traditional softmax loss L_{cls} widely

Table 6.2 : Classification accuracy for both 8 classes and 2 classes on the test set of **FI** using popular CNN models and the proposed method with traditional softmax loss(L_{cls}), multi-task loss(L_{multi}) and loss with probability(L_p).

Method	**FI**(8 classes)	**FI**(2 classes)
AlexNet+L_{cls}	58.61%	70.44%
ResNet101+L_{cls}	60.82%	74.67%
The proposed method+L_{cls}	73.05%	85.94%
AlexNet+L_p	57.44%	68.72%
ResNet101+L_p	59.28%	74.15%
The proposed method+L_p	73.58%	86.07%
AlexNet+L_{multi}	60.32%	72.83%
ResNet101+L_{multi}	62.77%	77.15%
The proposed method+L_{multi}	**75.46%**	**87.51%**

used in different CNN models, the two loss functions introduced before both taking label probability into account. Experiments are conducted on the **FI** dataset for popular CNN model and the proposed method using the loss functions mentioned above. The results are shown in Table 6.2. Though both L_{multi} and L_p introduce label probability into loss function, the performances of them are quite different. For L_p, the classification performance is worse than using L_{cls} while the performance of using L_{multi} is 2% better than that of using L_{cls}. The main reason is that, compared to L_p, L_{multi} introduce the inter-class relationship, rather than simply abandon the low-probability labels, which contribute to the overall classification performance. Therefore, L_{multi} is more suitable for emotion classification and the multi-task loss function is applied in the following experiments.

Choice of the parameter λ:

The parameter λ controls the two portion of the proposed loss function. $\lambda = 0$ means the proposed loss function is equal to cross entropy loss and $\lambda = 1$ means the proposed loss function is equal to KL loss. Considering the estimate emotion distribution $\hat{p}_i$ generated only using label probability and weak prior knowledge of emotion distance defined in Mikels' wheel(Figure 6.3), the parameter λ is not recommended to set too high. As **FI** dataset contains different kinds of affective images, the λ is only adjusted on the **FI** dataset to ensure the generalization of the proposed method. Figure 6.7 shows the effectiveness of parameter λ in the proposed loss function. When increases from 0 to 0.4, the classification performance is improved dramatically. However, further increasing over 0.5 leads to significant decreasing of the accuracy, since the large weight of L_{ed} introduces excess ambiguity. Therefore, $\lambda = 0.4$ is chosen in all experiments for a comprehensive considering of the hard emotional label and emotion distribution.

6.3.5 Comparison with State-of-the-art Methods

The results of the proposed method and state-of-the-art methods on the aforementioned 5 datasets(**FI**, **EmotionROI**, **IAPSsubset**, **ArtPhoto** and **Abstract**) is represented. For a fair comparison with the **EmotionROI** dataset, which only has two emotional classes, the classification performance of the other 4 datasets for both 8 classes and 2 classes are shown. The label conversion method is introduced in Section 6.3.4. For the small-scale datasets(**IAPSsubset**,**ArtPhoto**,**Abstract** and **EmotionROI**), the parameters of deep learning methods on the **FI** dataset can be transfered. The same experimental settings described in (Machajdik and Hanbury, 2010*b*) is followed. Due to the imbalanced and limited number of images per emotion category, the "one against all" strategy is employed to train the classifier. The image samples from each category are randomly split into five batches

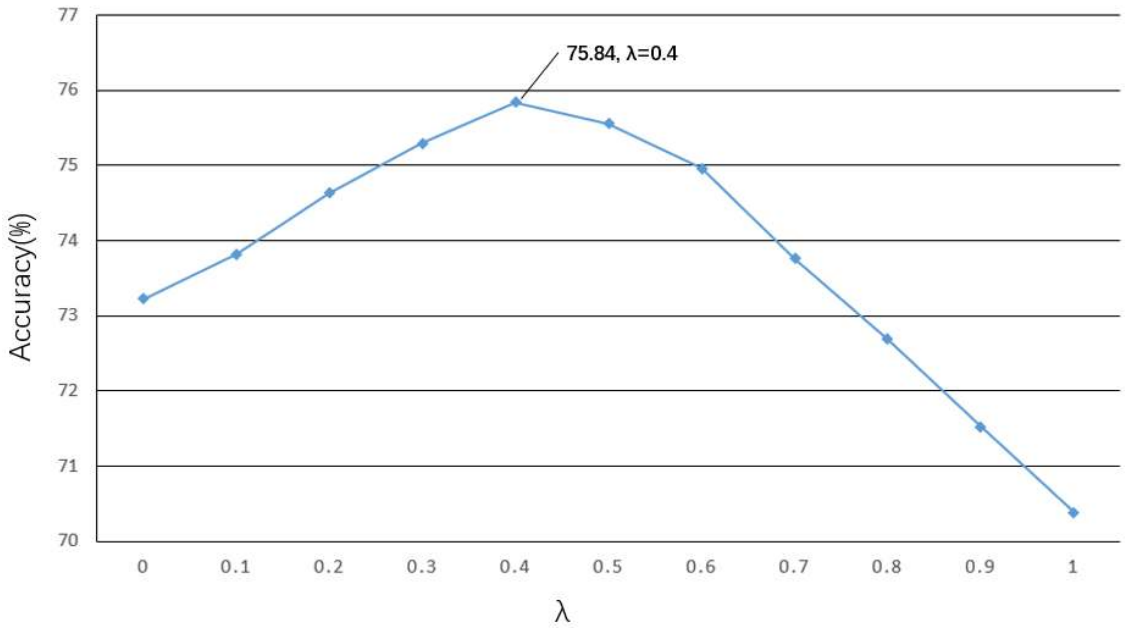

Figure 6.7 : Impact of different λ on the validation set of the **FI** dataset. $\lambda = 0.4$ achieves the best performance and is used in all experiments.

and 5-fold cross validation strategy is used to evaluate the different methods.

Table 6.3 shows the comparisons of the proposed methods to several state-of-the-art methods, including methods using hand-crafted features and deep features. It is clear that methods using deep features outperform methods using handcrafted features on large-scale dataset **FI**s. However, hand-crafted features show their effectiveness for some specific kinds of images on small-scale datasets.

For hand-crafted features, low-level features like color are very suitable to classify abstract paintings, which mainly consist of color and texture. While for other kinds of images, the simple color feature seems not enough for emotion classification. Multi-level features are combined in Zhao's method (Zhao et al., 2014*b*) and achieve an acceptable result for the small-scale dataset. The reason is that image emotion is related to various kinds of visual features from different levels, comprehensive

consideration of different visual features can benefit the classification result. In Rao(a) (Rao, Xu, Liu, Wang and Burnett, 2016), the local emotional region is extracted using image segmentation method and represented with SIFT feature and bag-of-words. SIFT feature is a texture representation, which can be used to detect concrete objects, *e.g.* face, building, animal *etc.*. The performance of the method demonstrates the effectiveness of both concrete objects and local regions for image emotion analysis.

For deep features, the performances of three popular CNN frameworks, which are AlexNet(Krizhevsky et al., 2012), VGGNet(Simonyan and Zisserman, 2014) and ResNet(He et al., 2016) are first compared. It can be found that as the CNN goes deeper, the emotion classification accuracy just slightly improves. The results show that high-level image semantics cannot be used for image emotion classification independently. Other deep methods utilize only one kind of features, like DeepSentibank(Chen, Borth, Darrell and Chang, 2014) and PCNN(You et al., 2015) also show limited performance. Both Rao(b)(Rao, Xu and Xu, 2016) and Zhu(Zhu et al., 2017) utilize the multi-level deep features extracted from the different level of CNN and achieve relatively high performance. Except for the multi-level features, the regional information contained in lower levels of convolutional layers also contributes to the improvement.

Employed both multi-level deep features and local emotional regions, the proposed framework outperforms both hand-crafted feature based methods and deep approaches in all datasets. Also, the proposed method shows a robust performance on different kinds of images, such as abstracting paintings consisting of color and texture and images from **IAPSsubset** whose emotions are evoked by certain objects. This means the proposed method effectively combine different levels of visual features from both global and local view.

6.4 Discussions

In this paper, the problem of image emotion recognition is investigated. Inspired by the observation that multi-level features and local regions with high emotional response contribute much to image emotion, a framework is proposed to automatically detect emotional regions on multi-level deep feature maps. The local emotional information extracted from emotional regions is combined with global information extracted from the whole image for image emotion classification. The label probability of the affective images are also utilized to leverage the ambiguity and subjectivity of the emotional labels. The experimental results show that the proposed method outperforms the state-of-the-art methods on different affective image datasets.

Table 6.3 : Classification results for different state-of-the-art methods on 5 different datasets. For **FI**, **IAPSsubset**, **Artphoto** and **Abstract**, classification results for both 2 classes and 8 classes is presented .

Method	**FI**		**IAPSsubset**		**ArtPhoto**		**Abstract**		**EmotionROI**
	8 classes	2 classes	8 classes	2 classes	8 classes	2 classes	8 classes	2 classes	2 classes
GCH	34.76%	47.95%	55.15%	69.96%	52.14%	66.53%	54.74%	67.33%	66.85%
LCH	32.42%	45.37%	43.15%	52.84%	50.41%	64.33%	55.45%	70.93%	63.79%
GCH+BoW	36.63%	50.05%	57.18%	71.63%	57.41%	71.30%	54.26%	68.92%	67.48%
GCH+BoW	34.58%	48.26%	47.61%	56.07%	52.05%	66.72%	58.39%	72.48%	65.67%
Zhao	46.52%	58.42%	63.61%	65.77%	66.37%	68.42%	60.60%	66.23%	73.45%
Rao(a)	51.67%	62.79%	70.32%	78.34%	69.74%	71.53%	62.17%	67.82%	74.51%
SentiBank	44.49%	56.47%	73.58%	80.57%	53.96%	67.33%	50.68%	64.30%	65.73%
AlexNet	58.61%	68.63%	72.24%	84.58%	67.03%	69.27%	61.96%	65.49%	71.60%
VGG-16	59.75%	73.95%	74.78%	87.20%	68.16%	70.48%	62.41%	65.88%	72.49%
ResNet101	61.82%	75.76%	75.09%	88.15%	69.36%	71.08%	63.56%	66.64%	73.92%
DeepSentiBank	53.16%	64.39%	75.88%	86.31%	68.54%	70.26%	66.46%	69.07%	70.38%
PCNN	56.16%	73.59%	76.87%	88.65%	68.93%	71.47%	67.17%	70.26%	74.06%
Rao(b)	67.24%	79.54%	78.08%	90.53%	69.75%	74.83%	67.81%	71.96%	78.99%
Zhu	73.03%	84.26%	82.39%	91.38%	71.63%	75.50%	68.45%	73.88%	80.52%
The proposed method	**75.46**%	**87.51**%	**84.71**%	**93.66**%	**74.58**%	**78.36**%	**70.77**%	**77.28**%	**82.94**%

Chapter 7

Conclusion and Future Work

7.1 Conclusions

Each of the included works in this thesis provides an in-depth analysis of corresponding approaches for image emotion classification. The performance of the inculded works is shown in table 7.1

Table 7.1 : Performance of included works in this thesis

Methods	FI	IAPSsubset	ArtPhoto	Abstract
Region-based affective image analysis (Rao et al., 2017)	51.67%	70.32%	69.74%	62.17%
Multi-level deep representations for image emotion classification (Rao, Xu and Xu, 2016)	67.24%	78.08%	69.75%	67.81%
Multi-level Region-based Convolutional Neural Network for Image Emotion Classification (Rao, Li, Zhang and Xu, 2019)	75.46%	84.71%	74.58%	70.77%

These works have shown that deep learning feature can achieve great success in the task of image emotion classification in more realistic scenarios. Therefore, following the new development and applying CNN for image emotion classification is a worthwhile direction for research. However, most CNN architectures

are designed specifically for semantic-level computer vision tasks, such as object recognition(Krizhevsky et al., 2012), object detection(Girshick et al., 2014), scene detection (Zhou et al., 2014) and image segmentation (Long et al., 2015), without considering the complexity and subjectivity existing in abstract-level computer vision tasks, including photo quality assessment, image aesthetics analysis and image emotion classification. What's more, the limited number of affective images with the reliable emotional label also confines the performance and generalization of CNN architectures in image emotion classification.

As aforementioned, the long-term goal of our research is to design a CNN structure for image emotion classification considering the complexity and subjectivity in emotion recognition. The work in this thesis contributes towards that goal, showing that combining different level of deep representations extracted from both global and local view. Two of the articles in this thesis shows how to build CNN architecture specifically for image emotion classification. The article in Chapter 5, compared to popular CNN architectures, proposed to combine the different level of deep representations extracted from different convolutional layers in CNN for image emotion classification. The correlations between different level of deep representations are also studied. For the next article in Chapter 6, we improve our model by combining region-based CNN, which can automatically discover local emotional regions from the image, and considering the noisy labeled training images in large-scale dataset. The proposed model significantly outperforms the previous methods based on single column CNN architecture.

7.2 Future Work

A logical follow-up project would be introducing attention model into CNN framework for image emotion classification. Through RCNN can discover the emotional regions in images more efficiently, it relies on a large amount of images with

labeled emotional regions for training. Considering the existing dataset only containing over 1,000 images with the labeled emotional region, the performance of the RCNN has been limited. Compared to RCNN, the performance of attention based model affected less by the training data. Such a model could learn a good emotion representation from multiple small- to medium-scale datasets with different parametrization.

Another important direction for future research is to collect a large scale affective image dataset with diverse labeling information, including emotion types, emotional regions, caption with emotion information, *etc.* Compared to the large number of datasets with richly annotated images for semantic-level image analysis, such as MS COCO (Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick, 2014) and Visual Genome (Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma et al., 2017), which largely promote the research on corresponding tasks, most of datasets for image emotion analysis just has limited annotated information like emotion type. Building a large scale with richly annotated affective information, such as emotion types, emotional regions and caption with emotion information would significantly help the research on image emotion analysis. It can also extend the applications of image emotion classification for other computer vision tasks, e.g. image captioning, object detection and image segmentation.

Consider the subjectivity existing image emotion analysis, multi-task learning and multi-label learning is also worth to be applied in this area. Compared using a single label to describe image emotion, emotion distribution or multiple emotion labels can better represent viewers' opinions towards one image (Zhao et al., 2017). They can represent different feelings of each viewer on the same image or the multiple emotion reactions for one viewer on the image. Therefore, applying multi-task learning and multi-label learning would be very useful for image emotion analysis.

Combining visual emotion can also produce more interesting applications. Internet users usually post images on social networks, personality analysis can be done based on these images for recommendation and advertisement. Image emotion can also benefit the area of computer based automatic design. Combining with visual emotion can make the products of computer based automatic design diversification and individuation.

Bibliography

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Susstrunk, S. (2012), 'Slic superpixels compared to state-of-the-art superpixel methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(11), 2274–2282.

Alameda-Pineda, X., Ricci, E., Yan, Y. and Sebe, N. (2016), Recognizing emotions from abstract paintings using non-linear matrix completion, *in* 'CVPR'.

Andrearczyk, V. and Whelan, P. F. (2016), 'Using filter banks in convolutional neural networks for texture classification', *Pattern Recognition Letters* **84**, 63–69.

Antonisse, H. J. (1982), 'Image segmentation in pyramids', *Computer Graphics and Image Processing* **19**(4), 367–383.

Aronoff, J. (2006), 'How we recognize angry and happy emotion in people, places, and things', *Cross-cultural research* **40**(1), 83–105.

Ballard, D. H. (1981), 'Generalizing the hough transform to detect arbitrary shapes', *Pattern recognition* **13**(2), 111–122.

Bar, M. and Neta, M. (2006), 'Humans prefer curved visual objects', *Psychological science* **17**(8), 645–648.

Benini, S., Canini, L. and Leonardi, R. (2011), 'A connotative space for supporting movie affective recommendation', *IEEE Transactions on Multimedia* **13**(6), 1356–1370.

Bianchi-Berthouze, N. (2003), 'K-dime: an affective image filtering system', *Multi-Media, IEEE* **10**(3), 103–106.

Borth, D., Ji, R., Chen, T., Breuel, T. and Chang, S.-F. (2013), Large-scale visual sentiment ontology and detectors using adjective noun pairs, *in* 'ACM MM'.

Bosch, A., Zisserman, A. and Muñoz, X. (2006), Scene classification via plsa, *in* 'Computer Vision–ECCV 2006', Springer, pp. 517–530.

Bruce, N. and Tsotsos, J. (2005), Saliency based on information maximization, *in* 'Advances in neural information processing systems', pp. 155–162.

Chen, C.-H., Patel, V. M. and Chellappa, R. (2015), Matrix completion for resolving label ambiguity, *in* 'CVPR'.

Chen, T., Borth, D., Darrell, T. and Chang, S.-F. (2014), 'Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks', *arXiv preprint arXiv:1410.8586* .

Chen, T., Yu, F. X., Chen, J., Cui, Y., Chen, Y.-Y. and Chang, S.-F. (2014), Object-based visual sentiment concept analysis and application, *in* 'ACM MM'.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C. and Bengio, Y. (2015), A recurrent latent variable model for sequential data, *in* 'NIPS', pp. 2980–2988.

Collingwood, R. G. (1938), *The principles of art*, Vol. 11, Oxford University Press.

Colombo, C., Del Bimbo, A. and Pala, P. (1999), 'Semantics in visual information retrieval', *IEEE MultiMedia* (3), 38–53.

Corridoni, J. M., Del Bimbo, A. and Pala, P. (1999), 'Image retrieval by color semantics', *Multimedia systems* **7**(3), 175–183.

Datta, R., Joshi, D., Li, J. and Wang, J. Z. (2006), Studying aesthetics in photographic images using a computational approach, *in* 'ECCV', pp. 288–301.

Daubechies, I. (1992), *Ten lectures on wavelets*, Vol. 61, Siam.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* 'CVPR'.

Fei-Fei, L. and Perona, P. (2005), A bayesian hierarchical model for learning natural scene categories, *in* 'IEEE Computer Society Conference on Computer Vision and Pattern Recognition', Vol. 2, pp. 524–531.

Gal, Y. (2016), 'Uncertainty in deep learning', *University of Cambridge* .

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J. and Geng, X. (2017), 'Deep label distribution learning with label ambiguity', *IEEE Transactions on Image Processing* **26**(6), 2825–2838.

Girshick, R. (2015), 'Fast r-cnn', *arXiv preprint arXiv:1504.08083* .

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014), Rich feature hierarchies for accurate object detection and semantic segmentation, *in* 'CVPR', pp. 580–587.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2016), 'Region-based convolutional networks for accurate object detection and segmentation', *IEEE transactions on pattern analysis and machine intelligence* **38**(1), 142–158.

Guo, M., Zhao, Y., Zhang, C. and Chen, Z. (2014), 'Fast object detection based on selective visual attention', *Neurocomputing* **144**, 184–197.

Hanjalic, A. (2006*a*), 'Extracting moods from pictures and sounds: Towards truly personalized tv', *IEEE Signal Processing Magazine* **23**(2), 90–100.

Hanjalic, A. (2006*b*), 'Extracting moods from pictures and sounds: Towards truly personalized tv', *IEEE Signal Processing Magazine* **23**(2), 90–100.

Hanjalic, A. and Xu, L.-Q. (2005), 'Affective video content representation and modeling', *IEEE Transactions on Multimedia* **7**(1), 143–154.

Harel, J., Koch, C. and Perona, P. (2006), Graph-based visual saliency, *in* 'Advances in neural information processing systems', pp. 545–552.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep residual learning for image recognition, *in* 'CVPR', pp. 770–778.

Hobbs, J. A., Salome, R. A. and Vieth, K. (1995), *The visual experience*, Davis Publications.

Hubel, D. H. and Wiesel, T. N. (1962), 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *The Journal of physiology* **160**(1), 106–154.

Irie, G., Satou, T., Kojima, A., Yamasaki, T. and Aizawa, K. (2010), 'Affective audio-visual words and latent topic driving model for realizing movie affective scene classification', *IEEE Transactions on Multimedia* **12**(6), 523–535.

Itten, J. and Van Haagen, E. (1962), *The Art of Color; the Subjective Experience and Objective Rationale of Colour*, Reinhold.

Itti, L., Koch, C. and Niebur, E. (1998), 'A model of saliency-based visual attention for rapid scene analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (11), 1254–1259.

Jia, J., Wu, S., Wang, X., Hu, P., Cai, L. and Tang, J. (2012), Can we understand van gogh's mood?: learning to infer affects from images in social networks, *in* 'ACM MM'.

Joachims, T. (1999), Transductive inference for text classification using support vector machines, *in* 'ICML'.

Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.-T., Wang, J. Z., Li, J. and Luo, J. (2011), 'Aesthetics and emotions in images', *IEEE Signal Processing Magazine* **28**(5), 94–115.

Judd, T., Ehinger, K., Durand, F. and Torralba, A. (2009), Learning to predict where humans look, *in* '12th international conference on Computer Vision', pp. 2106–2113.

Kang, H.-B. (2003), Affective content detection using hmms, *in* 'ACM MM'.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al. (2017), 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *International Journal of Computer Vision* **123**(1), 32–73.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'NIPS'.

Lang, P. J., Bradley, M. M. and Cuthbert, B. N. (2008*a*), 'International affective picture system (iaps): Affective ratings of pictures and instruction manual', *Technical report A-8* .

Lang, P. J., Bradley, M. M. and Cuthbert, B. N. (2008*b*), 'International affective picture system (iaps): Affective ratings of pictures and instruction manual', *Technical report A-8* .

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2324.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017), Feature pyramid networks for object detection, *in* 'CVPR', p. 4.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014), Microsoft coco: Common objects in context, *in* 'ECCV', pp. 740–755.

Liu, H., Xu, M., Wang, J., Rao, T. and Burnett, I. (2016), 'Improving visual saliency computing with emotion intensity', *IEEE transactions on neural networks and learning systems* **27**(6), 1201–1213.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C. (2016), Ssd: Single shot multibox detector, *in* 'ECCV', pp. 21–37.

Long, J., Shelhamer, E. and Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'CVPR'.

Lopes, A. T., de Aguiar, E., De Souza, A. F. and Oliveira-Santos, T. (2017), 'Facial expression recognition with convolutional neural networks: coping with few data and the training sample order', *Pattern Recognition* **61**, 610–628.

Loy, G. and Eklundh, J.-O. (2006), Detecting symmetry and symmetric constellations of features, *in* 'ECCV', pp. 508–521.

Loy, G. and Zelinsky, A. (2003), 'Fast radial symmetry for detecting points of interest', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25**(8), 959–973.

Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J. Z. (2014), Rapid: rating pictorial aesthetics using deep learning, *in* 'ACM MM'.

Lu, X., Suryanarayan, P., Adams Jr, R. B., Li, J., Newman, M. G. and Wang, J. Z. (2012), On shape and the computability of emotions, *in* 'Proceedings of the

20th ACM international conference on Multimedia', pp. 229–238.

Machajdik, J. and Hanbury, A. (2010*a*), Affective image classification using features inspired by psychology and art theory, *in* 'Proceedings of the international conference on Multimedia', pp. 83–92.

Machajdik, J. and Hanbury, A. (2010*b*), Affective image classification using features inspired by psychology and art theory, *in* 'ACM MM', pp. 83–92.

Mao, Q., Dong, M., Huang, Z. and Zhan, Y. (2014), 'Learning salient features for speech emotion recognition using convolutional neural networks', *IEEE Transactions on Multimedia* **16**(8), 2203–2213.

Maron, O. and Ratan, A. L. (1998), Multiple-instance learning for natural scene classification., *in* 'ICML', Vol. 98, pp. 341–349.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J. and Reuter-Lorenz, P. A. (2005*a*), 'Emotional category data on images from the international affective picture system', *Behavior research methods* **37**(4), 626–630.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J. and Reuter-Lorenz, P. A. (2005*b*), 'Emotional category data on images from the international affective picture system', *Behavior research methods* **37**(4), 626–630.

Pang, L., Zhu, S. and Ngo, C.-W. (2015), 'Deep multimodal learning for affective analysis and retrieval', *IEEE Transactions on Multimedia* **17**(11), 2008–2020.

Peng, K.-C., Chen, T., Sadovnik, A. and Gallagher, A. C. (2015), A mixed bag of emotions: Model, predict, and transfer emotion distributions, *in* 'CVPR'.

Peng, K.-C., Sadovnik, A., Gallagher, A. and Chen, T. (2016), Where do emotions come from? predicting the emotion stimuli map, *in* 'ICIP', pp. 614–618.

Plutchik, R. (2001), 'The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice', *American scientist* **89**(4), 344–350.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M. and Chua, T.-S. (2010), An eye fixation database for saliency detection in images, *in* 'European Conference on Computer Vision', pp. 30–43.

Rao, T., Li, X., Zhang, H. and Xu, M. (2019), 'Multi-level region-based convolutional neural network for image emotion classification', *Neurocomputing* **333**, 429–439.

Rao, T., Xu, M. and Liu, H. (2017), 'Generating affective maps for images', *Multimedia Tools and Applications* pp. 1–21.

Rao, T., Xu, M., Liu, H., Wang, J. and Burnett, I. (2016), Multi-scale blocks based image emotion classification using multiple instance learning, *in* 'ICIP'.

Rao, T., Xu, M. and Xu, D. (2016), 'Learning multi-level deep representations for image emotion classification', *arXiv preprint arXiv:1611.07145* .

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016), You only look once: Unified, real-time object detection, *in* 'CVPR', pp. 779–788.

Ren, S., He, K., Girshick, R. and Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, *in* 'NIPS'.

Sartori, A., Culibrk, D., Yan, Y. and Sebe, N. (2015), Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings, *in* 'ACM MM'.

Siersdorfer, S., Minack, E., Deng, F. and Hare, J. (2010), Analyzing and predicting sentiment of images on the social web, *in* 'ACM MM', pp. 715–718.

Simonyan, K. and Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition', *CoRR* **abs/1409.1556**.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000), 'Content-based image retrieval at the end of the early years', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(12), 1349–1380.

Soleymani, M., Larson, M., Pun, T. and Hanjalic, A. (2014), 'Corpus development for affective video indexing', *IEEE Transactions on Multimedia* **16**(4), 1075–1089.

Solli, M. and Lenz, R. (2009), Color based bags-of-emotions, *in* 'CAIP'.

Sun, M., Yang, J., Wang, K. and Shen, H. (2016), Discovering affective regions in deep convolutional neural networks for visual sentiment prediction, *in* 'ICME', pp. 1–6.

Sun, X., Yao, H. and Ji, R. (2012), What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency, *in* 'Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on', pp. 1552–1559.

Sun, X., Yao, H., Ji, R. and Liu, S. (2009), Photo assessment based on computational visual attention model, *in* 'ACM MM', pp. 541–544.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014), Sequence to sequence learning with neural networks, *in* 'NIPS', pp. 3104–3112.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015), Going deeper with convolutions, *in* 'CVPR'.

Tarvainen, J., Sjoberg, M., Westman, S., Laaksonen, J. and Oittinen, P. (2014), 'Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments', *IEEE Transactions on Multimedia* **16**(8), 2085–2098.

Tieleman, T. and Hinton, G. (2012), 'Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude', *COURSERA: Neural Networks for Machine Learning* **4**(2).

Valdez, P. and Mehrabian, A. (1994), 'Effects of color on emotions.', *Journal of Experimental Psychology: General* **123**(4), 394.

Van de Weijer, J., Schmid, C. and Verbeek, J. (2007), Learning color names from real-world images, *in* 'CVPR', pp. 1–8.

Wang, J., Fu, J., Xu, Y. and Mei, T. (2016), Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks., *in* 'IJCAI', pp. 3484–3490.

Wang, W. and He, Q. (2008), A survey on emotional semantic image retrieval., *in* 'ICIP'.

Warriner, A. B., Kuperman, V. and Brysbaert, M. (2013), 'Norms of valence, arousal, and dominance for 13,915 english lemmas', *Behavior research methods* **45**(4), 1191–1207.

Wei-ning, W., Ying-lin, Y. and Sheng-ming, J. (2006), Image retrieval by emotional semantics: A study of emotional space and feature extraction, *in* 'IEEE International Conference on Systems, Man and Cybernetics', Vol. 4, pp. 3534–3539.

Wu, Q., Zhou, C. and Wang, C. (2005), Content-based affective image classification and retrieval using support vector machines, *in* 'Affective Computing and Intelligent Interaction', pp. 239–247.

Xu, L., Yan, Q., Xia, Y. and Jia, J. (2012), 'Structure extraction from texture via relative total variation', *ACM Transactions on Graphics (TOG)* **31**(6), 139.

Xu, M., Luo, S. and Jin, J. S. (2008), Affective content detection by using timing features and fuzzy clustering, *in* 'Advances in Multimedia Information Processing-PCM 2008', Springer, pp. 685–692.

Yang, J., Sun, M. and Sun, X. (2017), Learning visual sentiment distributions via augmented conditional probability neural network., *in* 'AAAI', pp. 224–230.

Yang, Y., Jia, J., Zhang, S., Wu, B., Chen, Q., Li, J., Xing, C. and Tang, J. (2014), How do your friends on social media disclose your emotions?, *in* 'AAAI', Vol. 14, pp. 1–7.

Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., Melcher, D. and Sebe, N. (2012), In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings, *in* 'ACM MM'.

Yanulevskaya, V., Van Gemert, J., Roth, K., Herbold, A.-K., Sebe, N. and Geusebroek, J.-M. (2008), Emotional valence categorization using holistic image features, *in* 'ICIP'.

You, Q., Jin, H. and Luo, J. (2017), Visual sentiment analysis by attending on local image regions., *in* 'AAAI', pp. 231–237.

You, Q., Luo, J., Jin, H. and Yang, J. (2015), Robust image sentiment analysis using progressively trained and domain transferred deep networks., *in* 'AAAI', pp. 381–388.

You, Q., Luo, J., Jin, H. and Yang, J. (2016), Building a large scale dataset for image emotion recognition: The fine print and the benchmark, *in* 'AAAI'.

Zeiler, M. D. and Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* 'ECCV'.

Zhang, H., Gönen, M., Yang, Z. and Oja, E. (2015), 'Understanding emotional impact of images using bayesian multiple kernel learning', *Neurocomputing* .

Zhang, L., Tong, M. H., Marks, T. K., Shan, H. and Cottrell, G. W. (2008), 'Sun: A bayesian framework for saliency using natural statistics', *Journal of vision* **8**(7), 32–32.

Zhang, S., Huang, Q., Jiang, S., Gao, W. and Tian, Q. (2010), 'Affective visualization and retrieval for music video', *IEEE Transactions on Multimedia* **12**(6), 510–522.

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S. and Sun, X. (2014*a*), Exploring principles-of-art features for image emotion recognition, *in* 'Proceedings of the ACM International Conference on Multimedia', pp. 47–56.

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S. and Sun, X. (2014*b*), Exploring principles-of-art features for image emotion recognition, *in* 'ACM MM'.

Zhao, S., Yao, H., Gao, Y., Ji, R. and Ding, G. (2017), 'Continuous probability distribution prediction of image emotions via multi-task shared sparse regression', *IEEE Transactions on Multimedia* **19**(3), 632–645.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A. (2014), Learning deep features for scene recognition using places database, *in* 'NIPS'.

Zhu, X., Li, L., Zhang, W., Rao, T., Xu, M., Huang, Q. and Xu, D. (2017), Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition, *in* 'IJCAI', pp. 3595–3601.