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: ____________________________

Date: 2019.02.20

Production Note: Signature removed prior to publication.
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**
  
  

- **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
  

- **Chapter 6**
  

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

## Contents

Certificate

Abstract

Acknowledgments

List of Publications

List of Figures

List of Tables

Abbreviation

1 Introduction

2 Literature Review

2.1 Affective Modeling and Classification Methods

2.1.1 Affective Modelling

2.1.2 Classification Methods

2.2 Emotion Features Extraction

2.2.1 Low-level Visual Features

2.2.2 Mid-level Visual Features

2.2.3 High-level Visual Features

2.3 Convolutional Neural Networks

2.3.1 Region-based CNN

2.4 Affective Image Datasets
3  Overview of Research Goals and Contributions

4  Region-based Affective Image Analysis

4.1  Affective Map Generation

4.1.1  Multi-scale Block Extraction

4.1.2  BoVW Description

4.1.3  pLSA Representation

4.1.4  MIL Estimation

4.1.5  Affective Map Generation

4.2  Affective Image Classification

4.2.1  Experimental Setup

4.2.2  Datasets

4.2.3  Parameter Tuning

4.2.4  Results and Discussions

4.3  Saliency Detection

4.3.1  Experimental Setup

4.3.2  Results and Discussions

4.4  Discussions

5  Learning Multi-level Deep Representations for Image Emotion Classification

5.1  Multi-level deep representations for image emotion classification

5.1.1  Convolutional Neural Network

5.1.2  Analysis of different CNN models

5.1.3  Deep Network Learning Multi-level Deep representations

5.1.4  Fusion Layer
5.2 Experiments .............................................. 65
  5.2.1 Experimental Settings ............................... 65
  5.2.2 Emotion Classification on Large Scale and Noisy Labeled Dataset ................. 68
  5.2.3 Emotion Classification on small Scale Datasets ......................... 74
  5.2.4 Emotion Classification on Abstract Paintings ...................... 75

5.3 RNN for Visual Emotion Recognition ................. 77
  5.3.1 RNN for Visual Emotion Recognition .................. 79
  5.3.2 Experiment ........................................... 81

5.4 Discussions ........................................... 85

6 Multi-level Region-based Convolutional Neural Network for Image Emotion Classification 88

  6.1 Preliminaries ........................................... 89
    6.1.1 Feature Pyramid Network (FPN) ...................... 90
    6.1.2 Faster R-CNN ...................................... 91

  6.2 Emotion Analysis using Multi-level R-CNN ................. 92
    6.2.1 Emotional Region Extraction ...................... 92
    6.2.2 Emotion Distribution Estimation .................. 94
    6.2.3 Classifier and Loss Function ...................... 96

  6.3 Experiments and Results ................................ 98
    6.3.1 Dataset ........................................... 98
    6.3.2 Implementation Details ......................... 99
    6.3.3 Baseline ....................................... 100
    6.3.4 Experimental Validation ...................... 101
6.3.5 Comparison with State-of-the-art Methods . . . . . . . . . . . . [108]

6.4 Discussions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . [111]

7 Conclusion and Future Work [113]

7.1 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . [113]

7.2 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . [114]

Bibliography [117]
List of Figures

2.1 Sample images show the impact of content for image emotion. The two images are formally similar, but have totally opposite emotional impact. ................................................................. [19]

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 $32 \times 32$. Right: Two-by-two maxpooling with non-overlapping pooling regions. ................................. [21]

4.1 Affective maps for different emotion categories. In this research, eight basic emotion categories which are defined in (Mikels, Fredrickson, Larkin, Lindberg, Maglio and Reuter-Lorenz, 2005a) is applied ................................................................. [30]

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. ................................................................. [32]

4.3 Samples of affective maps based on SLIC (a) and affective maps based on pyramid segmentation (b) for 8 emotion categories. ....... [38]

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. ................................................................. [42]
4.5 Affective image classification results for different number of words combined with different number of topics .................................. 44

4.6 Affective image classification results for different levels using two image segmentation methods ............................................. 45

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, 2010a) and Zhao et al. (Zhao, Gao, Jiang, Yao, Chua and Sun, 2014a) .................................................. 46

4.8 Classification performance on the IAPS for the proposed method with pyramid segmentation and SLIC compared to Machajdik et al. (Machajdik and Hanbury, 2010a) and Zhao et al. (Zhao et al., 2014a) ..................... 47

4.9 Classification performance on the Abstract for the proposed method with pyramid segmentation and SLIC compared to Machajdik et al. (Machajdik and Hanbury, 2010a) and Zhao et al. (Zhao et al., 2014a) ..................... 48

4.10 Emotion distributions of eight emotion categories in three image examples ................................................................. 49

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) .................................................. 52
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.

5.2 Top 5 classification results for emotion category *contentment* using AlexNet (Krizhevsky, Sutskever and Hinton, 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.

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.

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

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

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

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

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

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

5.11 The proposed unified CNN-RNN framework for visual emotion recognition. Different levels of features from multiple branches in the CNN models 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.)

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

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

5.14 Performance evaluation on the ArtPhoto dataset.

5.15 Performance evaluation on the IAPS-Subset.
6.1 The overview of the proposed framework. The framework consists of 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.

6.2 Structure of Feature Pyramid Network (FPN).

6.3 Mikels’ emotion wheel and example of emotion distance.

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

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

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.

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

4.1 The percentage of images in one emotion category that can evoke another emotion in IAPS dataset (positive emotions) ............. 41
4.2 The percentage of images in one emotion category that can evoke another emotion in IAPS dataset (negative emotions) .......... 41
4.3 The number of the images per emotional categories in three datasets 43
4.4 AUC of ROC curves and p-values for testing affective map ........... 51

5.1 Emotion classification accuracy for MldrNet Models of different number of convolutional layer ........................................ 69
5.2 Emotion classification accuracy for MldrNet Models of different fusion function training on both well dataset and noisy dataset .... 70
5.3 Emotion classification accuracy for different methods on the large scale dataset for image emotion classification .......................... 71
5.4 Emotion classification accuracy of different methods on the MART dataset .............................................................. 78
5.5 Emotion classification accuracy of different methods on the large scale emotion dataset .............................................. 82
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.

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($L_p$).

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

7.1 Performance of included works in this thesis.
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