UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Learning Robust Features for Recognition of Emotions in Images and Videos

by

Haimin Zhang

A THESIS SUBMITTED
IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2019

Certificate of Original Authorship

I, Haimin Zhang declare that this thesis, is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 20 November 2019

Learning Robust Features for Recognition of Emotions in Images and Videos

by

Haimin Zhang

Abstract

Today, recognition of emotions in images and videos has attracted increasing research attention. In terms of video emotion recognition, most existing approaches are based on spatial features extracted from video frames. The performance of these approaches is mainly restricted due to the broad affective gap between spatial image features and high-level emotions. To bridge the affective gap, we propose to recognize emotions with kernelized features. A polynomial kernel function is constructed based on rewritten the equation of the discrete Fourier transform as the linear kernel. Moreover, we propose to apply the sparse representation method to kernelized features to reduce the impact of noise contained in video frames. This method can further help contribute to performance improvement.

In the second work, we develop a weighted sum pooling method for video emotion representation. We present an end-to-end deep network for simultaneously image emotion classification and emotion intensity map prediction. The proposed network is build based on the feature pyramid network. The class activation mapping technique is utilized to generate pseudo intensity maps to train the network. The proposed network is first trained on a large-scale image emotion dataset and then used to extracted features and intensity maps for video frames. We empirically show that this approach is effective to improve recognition performance.

Recent work has shown that using local region information helps to improve image emotion recognition performance. In the third work, we develop an end-to-end deep neural network for image emotion recognition by utilizing emotion intensity. The proposed network is composed of an intensity prediction stream and a classification stream. The

class activation mapping technique is used to generated pseudo intensity maps to guide the intensity prediction network for emotion intensity learning. The predicted intensity maps are integrated to the classification stream for final recognition. The two streams are trained cooperatively with each other to improve the overall performance.

In the fourth work, we present a dual pattern learning network architecture with adversarial adaptation (DPLAANet). Unlike conventional networks, the proposed architecture has two input branches. The dual input structure allows the network to have a considerably large number of image pairs for training. This can help address the overfitting issue due to limited training data. Moreover, we introduce to use the adversarial training approach to reduce the domain difference between training data and test data. The experimental results show that the DPLAANets are effective for several benchmark datasets.

Thesis Supervisor: A/Prof. Min Xu
School of Electrical and Data Engineering

Acknowledgements

The four-year Ph.D study at UTS has been a wonderful experience. I would like to acknowledge several people who not only made this thesis well finished, but also brought a lot of support, joy, and happiness into this amazing academic journey.

First and foremost, I would like to sincerely thank my supervisor, A/Prof. Min Xu, for her continually supervision and encouragement during my Ph.D. study. We have collaborated on a number of awesome projects since I started the Ph.D program four years ago. Her sharp intuition and passion for knowledge have influenced me to dig deeper into the problems we are having and to discover something novel. This would be much helpful for my future career. I feel I am very fortunate to have been supervised by and working with her in the past four years.

I would like to express my thanks and appreciation to my co-supervisor Dr. Xiaoying Kong for help guidance and help. Many thanks to A/Prof. Qiang Wu and Dr. Wenjing Jia for their valuable advices and suggestions for my candidature assessment one and two, which are much helpful for improving the quality of this thesis. I would like to give thanks to A/Prof. Richard Xu, who has provided useful insight for my research and career. Special thanks go to Prof. Yu-gang Jiang at Fudan University who provided the datasets for conducting experiments.

I would like to acknowledge my labmates in the Aural and Visual Intelligence Lab: Dr. Tianrong Rao, Madhumita Takalkar, Lingxiang Wu, Zhongqin Wang, Zhiyuan Shi, Lei Sang, Yukun Yang, Wanneng Wu, Ruiheng Zhang, and Xiaoxu Li. I would like to thank my friends for their assistance: Dr. Cheng Luo and Dr. Ming Liu, Dr. Hao Li, Dr. Lin Ye, Dr. Zhichao Sheng, Dr. Ye Shi, etc.

I would like to acknowledge with gratitude the love of my family. Their support has always been unconditional. This thesis could not have been well finished without their

support.

Finally, I would like to thank the anonymous reviewers for reviewing this thesis.

List of Publications

The contents of this thesis are based on the following papers that have been published or accepted, or preprints that have been under submission or submitted to peer-reviewed journals.

Publications:

1. Haimin Zhang and Min Xu, "Recognition of Emotions in User-Generated Videos With Kernelized Features," *IEEE Transactions on Mulitmedia*, vol. 20, no. 10, pp. 2824-2835, 2018.

2. Haimin Zhang and Min Xu, "Modeling temporal information using discrete fourier transform for recognizing emotions in user-generated videos," *IEEE International Conference on Image Processing (ICIP)*, 2016.

3. Madhumita A. Takalkar, Haimin Zhang, and Min Xu, "Improving Micro-expression Recognition Accuracy Using Twofold Feature Extraction," *International Conference on MultiMedia Modeling (MMM)*, 2019.

4. Tianrong Rao, Xiaoxu Li, Haimin Zhang, and MinXu, "Multi-level region-based Convolutional Neural Network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429-439, March, 2019.

5. Shenghong Hu, Min Xu, Haimin Zhang, Chunxia Xiao, and Chao Gui, "Affective Content-aware Adaptation Scheme on QoE Optimization of Adaptive Streaming over HTTP," accepted to *ACM Transactions on Multimedia Computing, Communications, and Applications* .

Others:

1. Haimin Zhang and Min Xu, "Weakly Supervised Emotion Intensity Prediction for Recognition of Emotions in Images," under review by *IEEE Transactions on Mulitmedia*.

2. Haimin Zhang and Min Xu, "Improving the Performance of Deep Networks by Dual Pattern Learning with Adversarial Adaptation," under first revision by *IEEE Transactions on Circuits and Systems for Video Technology*.

3. Haimin Zhang and Min Xu, "Frame-level Emotion Intensity Prediction for Improving Video Emotion Recognition Performance," under submission to *IEEE Transactions on Affective Computing*.

Contents

List of Figures

List of Tables

Abbreviation

- CNN: convolutional neural network
- RNN: recurrent neural networks
- SVM: support vector machine
- GAN: generative adversarial network
- DFT: discrete Fourier transform
- FFT: Fast Fourier transform
- FV: Fisher vector
- VLAD: vector of locality aggregated vectors
- CAM: class activation mapping
- RMSE: root mean square error
- RMSEL: Root mean square error in log space
- SGD: stochastic gradient descent
- ITE: image transfer encoding
- LLC: locality-constrained linear coding
- DPL: dual pattern learning
- ERM: empirical risk minimization
- HMM: hidden Markov model
- MFCC: Mel-frequency cepstral coefficients

- STE: short-time energy
- FPN: feature pyramid network
- CAN: collaborative and adversarial networks
- ADDA: adversarial discriminative domain adaptation
- SymNets: domain-symmetric networks
- LSTM: long short-term memory