

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 generate 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 Multitmedia*, 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 Multimedia*.
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

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	
List of Figures	
List of Tables	
Abbreviation	
1 Introduction	1
1.1 Background and Problem Statement	1
1.2 Thesis Objectives and Contributions	3
1.3 Thesis Outline	4
2 Related Work	6
2.1 Emotion Modelling	6
2.2 Deep Neural Networks	7
2.3 Domain Adaptation	9
2.4 Kernel Methods	10
2.5 Emotion Recognition in Videos	12
2.6 Emotion Recognition in Images	14
3 Recognition of Emotions in User-generated Videos with Kernel-	

ized Features	17
3.1 Introduction	17
3.2 The Proposed Approach	20
3.2.1 Frame-level Feature Extraction	20
3.2.2 Apply Kernel Method to CNN Features	21
3.2.3 Sparse Representation for Denoising	24
3.2.4 Video-level Representation and Classification	27
3.3 Experimental Results	27
3.3.1 Experimental Setup	27
3.3.2 Results on VideoEmotion-8	28
3.3.3 Results on Ekman-6	36
3.4 Conclusion	42
3.A Appendix	43

4 Frame-level Emotion Intensity Prediction for Improving Video

Emotion Recognition Performance	45
4.1 Introduction	45
4.2 Methodology	47
4.2.1 The network architecture	48
4.2.2 CAM guided Pseudo intensity map generation	49
4.2.3 The loss functions	50
4.2.4 Video representation and classification	52
4.2.5 Training details	52
4.3 Experiments	53
4.3.1 Experimental Setup	53

4.3.2	Results on VideoEmotion-8	53
4.3.3	Results on Ekman-6	56
4.4	Conclusions	60
5	Weakly Supervised Emotion Intensity Prediction for Recognition of Emotions in Images	62
5.1	Introduction	62
5.2	Methodology	65
5.2.1	Pseudo intensity map generation	65
5.2.2	The network architecture	67
5.2.3	The loss functions	68
5.3	Experiments	72
5.3.1	Experimental setup	72
5.3.2	Results on Emotion-6	73
5.3.3	Results on FI-8	76
5.3.4	Results on WEBEmo	77
5.3.5	Image sentiment analysis	80
5.4	Conclusion	81
5.A	Appendix	82
6	Dual Pattern Learning with Adversarial Adaptation	83
6.1	Introduction	83
6.2	Methodology	87
6.2.1	Dual pattern learning	87
6.2.2	Adversarial domain adaptation	90
6.3	Experiments	91

6.3.1	CIFAR-10 and CIFAR-100	91
6.3.2	Image emotion recognition	98
6.3.3	Google commands dataset	100
6.3.4	MNIST classification	101
6.3.5	Experiments on Small Datasets	102
6.4	Conclusion	103
7	Conclusion and Future Work	104
7.1	Conclusions	104
7.2	Future Work	105
	Bibliography	107

List of Figures

3.1	An illustration of space transformation using a kernel function. The kernel function might not be explicitly written out.	18
3.2	An overview of the proposed approach for recognition of emotions in user-generated videos.	19
3.3	An illustration of the effect of interpolation for two signals.	25
3.4	Examples of recognition accuracy for each emotion category on VideoEmotion-8. A check mark (✓) represents a correct recognition result, while an X mark (×) represents an incorrect recognition result. . .	30
3.5	Recognition accuracy for each emotion category on VideoEmotion-8 using the proposed approach.	31
3.6	Illustration of the effect of denoising using LLC for a signal in the frequency domain.	32
3.7	Recognition accuracy for each emotion category on Ekman-6 using the proposed approach.	37
3.8	Confusion matrix on Ekman-6 using kernelized features with denoising. .	37
3.9	Examples of recognition accuracy for each emotion category on Ekman-6. A check mark (✓) represents a correct recognition result, while an X mark (×) represents an incorrect recognition result.	38
4.1	Unlike average pooling, each frame-level feature is associated with an emotion intensity value in our method, the weighted summation is calculated as video features.	47

4.2	An overview of the proposed approach for video representation and recognition. Each selected video frame is passed to a pretrained deep neural network. The activations before the last fully connected layer are extracted as a frame-level feature, and the average value of the predicted intensity map is calculated as the weight for the frame-level feature. The weighted sum pooling is applied to generate video-level features. Finally, an SVM is trained for prediction.	48
4.3	Confusion matrix on VideoEmotion-8 using the proposed approach.	55
4.4	Recognition accuracy for each emotion category on VideoEmotion-8.	57
4.5	Confusion matrix on Ekman-6 using the proposed approach.	59
4.6	Recognition accuracy for each emotion category on Ekman-6.	60
4.7	Examples of predicted emotion intensity maps for video frames on Ekman-6.	61
5.1	Sample images and corresponding emotion intensity maps synthesized with the original image. As shown in the second row, emotion intensity maps highlight discriminative regions that invoke an emotion.	63
5.2	An overview of the proposed end-to-end network architecture for image emotion recognition. This network consists of an emotion intensity prediction stream and a classification stream. The predicted intensity maps are integrated to the classification stream for final emotion recognition. The proposed network is trained with a multi-task loss function. The two streams are trained cooperatively with each other to improve overall performance.	66
5.3	The emotion intensity prediction subnetwork. The subnetwork is built on top of the FPN. The CAM technique is used to generate pseudo intensity maps to guide the subnetwork for emotion intensity learning.	70

5.4	Examples of emotion intensity maps generated by CAM and predicted with the proposed network.	75
5.5	The confusion matrix on Emotion-6 using the proposed network based on ResNet-101.	76
5.6	The confusion matrix on FI-8 using the proposed network based on ResNet-101.	77
5.7	The confusion matrix on WEBEmo-25 using the proposed network based on ResNet-101.	78
5.8	The confusion matrix on WEBEmo-6 using the proposed network based on ResNet-101.	79
6.1	An illustration that shows humans learn knowledge by analyzing dual images. They may have more interest in learning one image than the other image. In this figure, the human is more interested in, or pays more attention to, learning dog (boldness of lines represents interest value). . .	84
6.2	An illustration of the proposed DPLAANet framework. This framework consists of a DPLNet and an adversarial adaptation module. The DPLNet has two input branches which share the same parameters. Feature maps generated by the two input branches are fused together to backbone network. We perform random weighted fusion. A value λ is sampled from the standard uniform distribution as weight for one branch, and $1 - \lambda$ for the other branch. The weight associated with each branch can be considered as an interest value for learning the corresponding image. The adversarial training approach is used to reduce the domain difference between fused image features and real image features.	86
6.3	Two test approaches at test time: (a) Pass a test image to both input branches and set λ to 0.5; (b) Give an image as input to one branch and set corresponding λ to 1 while ignoring the other input branch.	89

6.4	Test errors on CIFAR-10 and CIFAR-100 for ResNets, DPLNets, and DPLAANets.	97
-----	---	----

List of Tables

3.1	Overall recognition accuracy of the proposed approach on VideoEmotion-8.	29
3.2	Impact of the number of interpolated points on the overall recognition accuracy.	32
3.3	Performance of the proposed approach using features extracted from different CNN architectures on VideoEmotion-8.	33
3.4	Results of the proposed approach using features extracted from the ResNet fine-tuned on video frames on VideoEmotion-8.	34
3.5	Performance of the proposed approach using different pooling methods on VideoEmotion-8.	34
3.6	Comparison with previous work on VideoEmotion-8.	35
3.7	Overall recognition accuracy of the proposed approach on Ekman-6.	39
3.8	Performance of the proposed approach using features extracted from different CNN architectures on Ekman-6.	40
3.9	Results of the proposed approach using features from the ResNet fine-tuned on video frames on Ekman-6.	40
3.10	Performance of the proposed approach using different pooling methods on Ekman-6.	41
3.11	Comparison with previous work on Ekman-6.	42
3.12	Impact of the number of Gaussian components for FV on the overall recognition accuracy on VideoEmotion-8.	43

3.13	Impact of the number of clusters for VLAD on the overall recognition accuracy on VideoEmotion-8.	43
3.14	Impact of the number of Gaussian components for FV on the overall recognition accuracy on Ekman-6.	44
3.15	Impact of the number of clusters for VLAD on the overall recognition accuracy on Ekman-6.	44
4.1	Performance of the proposed approach and comparison with features extracted from other deep networks on VideoEmotion-8.	54
4.2	Comparison with previous work on VideoEmotion-8.	55
4.3	Performance of the proposed approach and comparison with features extracted from other deep networks on Ekman-6.	58
4.4	Comparison with previous work on Ekman-6.	58
5.1	Recognition accuracy (%) of the proposed network on Emotion-6.	73
5.2	Impact of loss function on performance on Emotion-6. The ResNet-50 was used as the backbone network in the experiments.	74
5.3	Recognition accuracy of the proposed network on FI-8 and comparison with previous work.	80
5.4	Recognition accuracy of the proposed network on WEBEmo-6 and WEBEmo-25.	80
5.5	Image sentiment recognition results using the proposed network on Ekman-2, FI-2, and WEBEmo-2, and comparison with previous work. . .	81
6.1	Test errors (%) on CIFAR-10 and CIFAR-100.	91
6.2	Test errors (%) on CIFAR-10 and CIFAR-100.	92

6.3	Test errors (%) on CIFAR-10 and CIFAR-100. k indicates the growth rate of network.	93
6.4	Test errors (%) on CIFAR-10 and CIFAR-100.	93
6.5	Ablation study. Performance comparison among DPLAANets, DPLNets, and vanilla ResNets on CIFAR-10 and CIFAR-100.	94
6.6	Impact of number of input branches on performance. We did not use adversarial adaptation for branch number equal to 1.	95
6.7	Comparison with previous work on CIFAR-10 and CIFAR-100.	96
6.8	Recognition accuracies (%) on FI-8.	99
6.9	Error rates (%) on the Google commands dataset.	99
6.10	Error rates (%) on MNIST.	99
6.11	Error rates (%) on subsets of CIFAR-10.	100
6.12	Error rates (%) on subsets of CIFAR-100.	100
6.13	Error rates (%) on subsets of MNIST. TPLAANet represents triple pattern learning with adversarial adaptation networks, in which three input branches are used.	100

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