Poster Abstract: Multimodal Emotion Recognition by extracting common and modality-specific information

Wei Zhang
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
wzhang17@mails.tsinghua.edu.cn

Weixi Gu
University of California, Berkeley
guweixigavin@gmail.com

Fei Ma
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
mf17@mails.tsinghua.edu.cn

Shiguang Ni
Graduate School at Shenzhen
Tsinghua University
ni.shiguang@sz.tsinghua.edu.cn

Lin Zhang
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
linzhang@tsinghua.edu.cn

Shao-Lun Huang
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
shaolun.huang@sz.tsinghua.edu.cn

ABSTRACT

Emotion recognition technologies have been widely used in numerous areas including advertising, healthcare and online education. Previous works usually recognize the emotion from either the acoustic or the visual signal, yielding unsatisfied performances and limited applications. To improve the inference capability, we present a multimodal emotion recognition model, EMOdal. Apart from learning the audio and visual data respectively, EMOdal efficiently learns the common and modality-specific information underlying the two kinds of signals, and therefore improves the inference ability. The model has been evaluated on our large-scale emotional data set. The comprehensive evaluations demonstrate that our model outperforms traditional approaches.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → *Human computer interaction (HCI)*;

KEYWORDS

Multimodal Machine Learning, Multimodal Emotion Recognition

ACM Reference Format:
Wei Zhang, Weixi Gu, Fei Ma, Shiguang Ni, Lin Zhang, and Shao-Lun Huang. 2018. Poster Abstract: Multimodal Emotion Recognition by extracting common and modality-specific information . In *The 16th ACM Conference on Embedded Networked Sensor Systems (SenSys '18), November 4–7, 2018, Shenzhen, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3274783.3275200

1 INTRODUCTION

Emotion recognition, as an essential interaction between human and machine, has been widely employed in a surge of fields [1, 2]. For example, Bargal in [3] encodes image features on pretrained networks and classifies with a linear SVM. Chernykh in [4] recognize emotions from speech using a recurrent neural network. Nevertheless, these works learn the emotion from a single modality: either the visual or the acoustic signals. Lacking of the multi-modal analysis results in the unsatisfied performance. Although Kahou in [5] has built models to merge these signals together, the plug-and-play models rarely capture the underlying correlations shared by the visual and audio signals, which limits their generalization.

To cope with these issues, we proposed EMOdal, a multi-modality emotion recognition model according to both visual and acoustic signals inspired by the work in [6]. EMOdal consists a convolutional neural network for visual signals and a deep neural network for acoustic signals. Networks are trained together with a loss composed of three types of losses. The common loss adopts a function defined in [7] so that the maximally correlated information between the visual and acoustic modalities is extracted. The classify loss is utilized to capture the relationship between emotion states and the combination of visual and acoustic signals. The specific loss makes use of modality specific information. In this way, useful visual and acoustic signals could be efficiently extracted within lower-dimension representations.

A large-scale real-world dataset is constructed to evaluate our model. It is collected from 33 TV shows and 20 movies with four emotion states: angry, happy, sad and neutral. The evaluation on 11606 clips over four emotions shows that EMOdal could achieve 66.02% accuracy in general superior to unimodal methods. The outstanding performance and the flexible design of the framework allows integration of various neural networks structures for audios and images, providing great potential for better performance.

2 MODEL ARCHITECTURE

The architecture of EMOdal is depicted in Figure 1. In general, EMOdal first leverages a deep neural network to learn the acoustic signals and adopts a convolutional neural network to deal with the visual signals. The inference results are then fed into the common and modality-specific loss function for the model training.

Input Features Image raw data is utilized as input, of which the size is 128×128. Audios are downsampled to 16 kHz and 68-dimensional features (*e.g.* spectral entropy, MFCCs, energy) are extracted by pyAudioAnalysis[8].

Image Network We construct a convolutional neural network with three 3×3 convolutional layers, each of which is followed

SenSys '18, November 4–7, 2018, Shenzhen, China

ACM ISBN 978-1-4503-5952-8/18/11…$15.00
https://doi.org/10.1145/3274783.3275200

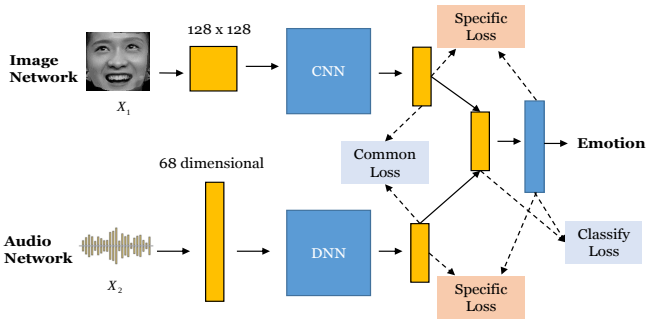

Figure 1: Model Architecture

by a max pooling layer. Features are flattened later. After a fully connected layer, the dropout strategy is deployed to prevent the network from overfitting. The last fully connected layer contains 10 units, so a 10-dimensional image feature is pulled out at last.

Audio Network A deep neural network with three fully connected layers containing 128, 64 and 10 units is adopted as the audio network. The network produces a 10-dimensional audio feature.

Loss Function and Model Training A combined loss function is adopted to extract the common and modality-specific information from acoustic and visual signals. It consists of three types of losses: a common loss, a classify loss and two specific losses. It has been verified in [6] that if visual signals X_1 and acoustic signals X_2 are weakly dependent, the optimal solution that maximizes the log-likelihood of the joint distribution of X_1 and X_2 is to maximize the score defined in [7]. A *common loss* is defined by placing a negative sign ahead of the score. It is desired that the common loss could extract the correlation between visual and acoustic signals efficiently. To capture the relationship between modalities and targets, we send the combined feature to a softmax layer and utilize a categorical crossentropy loss as a *classify loss*. In addition, two more losses are added in as *specific losses* to incorporate the complementary information from image and audio. They are categorical crossentropy losses based on targets and predictions made with image and audio features respectively. We combine all the losses together and select Adam to train our model with Keras.

3 EVALUATION

Data Collection and Processing Our data set is collected from films and TV programs composed of 4 basic emotion categories: angry, happy, sad and neutral. We extract the the middle frame for each video clip and crop frames to the largest face detections. Then we resize images to 128×128, do face frontalization and convert them to grayscale. Since faces temporally disappear in some frames, we only use frames with clear faces. We apply data augmentation techniques and extend our data set to 11606 clips, 80% of which are utilized as training data.

Performance We implement experiments on 1)audios only; 2)images only; 3)audios and images. Table 1 compares our accuracy result to unimodal methods. The overall accuracy of 66.02% shows that our model outperforms the unimodal methods. Table 2 shows the confusion matrix of EMOdal over the test sets. As seen, the predicted emotion states match the ground truth labels in most cases. Most of the error predicted points belongs to 'neutral' and 'sadness'.

Table 1: Evaluation Results

Model	Audio-only	Image-only	Audio+Image
Acc(%)	52.20	63.73	66.02

The main reason might be the ambiguity of suppressed expressions and real emotions for Chinese. For example, when people feel sad, they tend to suppress their emotions rather than expressing in an explicit way.

Table 2: Confusion Matrix

Ground Truth	**Predictions(%)**			
	angry	happy	neutral	sad
anger	69.74	8.55	9.21	12.5
happiness	9.09	76.97	9.70	4.24
neutral	15.13	17.11	55.26	12.5
sadness	11.64	9.59	17.81	60.96

4 CONCLUSION

In this work, we propose a model in multimodal emotion recognition, which could extract effective information between audios and images. The proposed model outperforms others on our large-scale real-world database. In the future, we might treat gender or age group as additional modalities since expressions and voices vary in people of different age and gender. Inspired by [9–11], we plan to adopt a multi-task learning framework on emotion detection to improve the inference detection ability, and adopt the transfer learning to generalize our model [12] on other language learning.

REFERENCES

[1] Zhi Liu, Wenjing Zhang, Jianwen Sun, Hercy N. H. Cheng, Xian Peng, and Sanya Liu. Emotion and associated topic detection for course comments in a mooc platform. In *International Conference on Educational Innovation Through Technology*, 2017.

[2] Yisi Liu, Olga Sourina, and Mohammad Rizqi Hafiyyandi. Eeg-based emotion-adaptive advertising. In *Affective Computing and Intelligent Interaction*, pages 843–848, 2013.

[3] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. Emotion recognition in the wild from videos using images. In *ACM International Conference on Multimodal Interaction*, pages 433–436, 2016.

[4] Vladimir Chernykh and Pavel Prikhodko. Emotion recognition from speech with recurrent neural networks. 2017.

[5] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Acm International Conference on Multimodal Interaction*, pages 467–474, 2015.

[6] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang. An efficient approach to informative feature extraction from multimodal data. Submitted to 2019 the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).

[7] Shao-Lun Huang, Xiangxiang Xu, Lizhong Zheng, and Gregory W. Wornell. Feature projection in deep neural networks. Submitted to 2018 Advances in Neural Information Processing Systems (NIPS 2018).

[8] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *Plos One*, 10(12):e0144610, 2015.

[9] Weixi Gu, Zimu Zhou, Yuxun Zhou, Miao He, Han Zou, and Lin Zhang. Predicting blood glucose dynamics with multi-time-series deep learning. 2017.

[10] Weixi Gu. Phd forum abstract: Non-intrusive blood glucose monitor by multi-task deep learning. 2017.

[11] Weixi Gu, Yuxun Zhou, Zimu Zhou, Xi Liu, Han Zou, Pei Zhang, Costas J Spanos, and Lin Zhang. Sugarmate: Non-intrusive blood glucose monitoring with smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):54, 2017.

[12] Weixi Gu, Zimu Zhou, Yuxun Zhou, Han Zou, Yunxin Liu, Costas J Spanos, and Lin Zhang. Bikemate: Bike riding behavior monitoring with smartphones. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MobiQuitous 2017*. ACM, 2017.