Video-Based Sentiment Analysis with hvnLBP-TOP Feature and bi-LSTM

Haoran Li,[1,2,3] **Hua Xu**[1,2]

[1]State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology（BNRist), Beijing 100084, China
[3]Department of Automation, Tsinghua University, Beijing 100084, China
lihr15@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

Abstract

In this paper, we propose a new feature extraction method called hvnLBP-TOP for video-based sentiment analysis. Furthermore, we use principal component analysis (PCA) and bidirectional long short term memory (bi-LSTM) for dimensionality reduction and classification. We achieved an average recognition accuracy of 71.1% on the MOUD dataset and 63.9% on the CMU-MOSI dataset.

Introduction

With the prosperity of video sharing websites and social network applications, multimodal sentiment analysis is becoming increasingly popular among researchers. Beyond information from natural language, visual information contains important sentiment features in the speakers' gestures and facial expressions. Therefore, video-based sentimental analysis and facial expression recognition (FER) on video are of great importance in analyzing multimodal sentiment.

Many previous works are associated with facial expression recognition on images. In the early stage, facial landmarks (Hong, Neven, and von der Malsburg 1998) have been used to describe the geometric feature of faces. Later, texture features including LBP (Ahonen, Hadid, and Pietikainen 2006) have been used for expression recognition. LBP has many more advanced derivatives, and horizontal and vertical neighborhood comparison LBP (hvnLBP) (Mistry et al. 2017) is the most advanced among them. These features compare each pixel with its border pixels and generate histograms of patterns of such comparison results. In recent years, machine learning techniques such as shallow convolutional neural network (CNN) (Burkert et al. 2015) have also been used to extract sentiment features. Besides image-based features, video-based sentiment analysis utilizes video features including LBP-TOP (Zhao and Pietikainen 2007). Unlike image-based features, there exist significant opportunities for accuracy improvements in applying video-based features.

This paper proposes a novel feature for video-based sentiment analysis. We combine the hand-crafted feature hvnLBP and the approach of concatenating features on three orthogonal planes (TOP) to create hvnLBP-TOP feature. By using principal components analysis (PCA), we reduce the length of our feature to 512. Finally, we select bidirectional LSTM architecture for sentiment classifying and regression. We verify our feature on MOUD dataset (Pérez-Rosas, Mihalcea, and Morency 2013) and CMU-MOSI dataset (Zadeh et al. 2016), and the results reveal that our model achieves better accuracy and efficiency than other approaches.

Proposed Method

The architecture contains 2 main modules: Feature Extraction Module and Sequential Learning Module, plus an additional preprocessing module. The details of each module are explained as follows.

Preprocessing

First, the video is cut into frames and stored as jpeg pictures. Second, we detect the biggest human face using the face detection API. Third, all frames in the video are cropped by the same face location, so that we get the human face video. Lastly we resize them into 98*98 png picture sequences.

Feature Extraction

As a texture descriptor, hvnLBP involves static facial states, but not movements of facial muscles. As temporal change is similar to spatial texture, we compute hvnLBP on three orthogonal planes, i.e. XY, XT and YT planes, in order to represent both facial state and movements.

To optimize this feature, we adjusted the computation of hvnLBP. For all hvnLBP results, there exist only 65 possible digits ranging from 0 to 255. Given this, we create a map between such digits in range 0-64. We then concate-

nate all histograms for each 12*12*5 grid in every 5 connected frames (1/6 seconds). To further reduce the dimension of the feature, principal component analysis (PCA) is conducted to compress the feature dimension to 512.

Sequential Learning

Once the feature is generated, it is put in the following sequential learning module with layers including:

1) Bi-LSTM layer (Hochreiter and Schmidhuber 1997);
2) Batch Normalization layer;
3) Two dense layers and dropout layer;
4) Dense layer for classification (or regression).

The activation function in the last layer is dependent on the specific task: SoftMax for classification task, and sigmoid for regression task.

Experiments

In order to test the effect of the proposed feature, we design several experiments on two datasets: MOUD dataset (Pérez-Rosas, Mihalcea, and Morency 2013) and CMU-MOSI dataset (Zadeh et al. 2016). The MOUD dataset consists of 498 utterances of product review video in Spanish. Each video consists of several utterances of product review with sentiment labels of positive, negative or neutral, and we only keep 450 positive or negative labeled utterances for experimental use. The CMU-MOSI dataset contains 2199 segments from 93 opinion videos of movie reviews on YouTube. Each segment is labeled in the range [-3, 3]. Among all of the 2199 segments, only 116 of them have lengths lasting over 10 seconds. So in our experiment, segments longer than 10 seconds are shortened to 10 seconds by cutting the long tail.

In MOUD experiment, we conduct a binary sentiment classification and use accuracy and F1 score to evaluate performance. We randomly select 64%:16%:20% for training, validation and test set. More specifically, we use 288 samples for training, 72 for validation, and 90 for test.

In MOSI experiment, we conduct binary sentiment classification, five-class sentiment classification and sentiment regression. We use both accuracy and F1 score for binary classification, the accuracy for five-class classification, and the mean absolute error (MAE) for regression tasks. The amount of training, validation and test set is 1376:344:479.

We compare the performance of our model with the following state-of-the-art models for video sentiment analysis:

AU + SVM (Pérez-Rosas et al. 2013) trains an SVM model on Action Unit features and holds the state-of-the-art for video sentiment analysis on MOUD dataset.

TFN-V (Zadeh et al. 2017) is the visual subnet of the state-of-the-art multimodal sentiment analysis model on CMU-MOSI dataset.

Task	MOUD		MOSI			
	Binary		Binary		5-class	Regr.
Metric	Acc.	F1	Acc.	F1	Acc.	MAE
SOTA	67.3	61.2	**65.3**	54.5	29.5	**1.21**
Ours	**71.1**	**62.8**	63.9	**65.9**	**29.8**	1.29

Table 1. Results on visual data of MOUD Dataset and CMU-MOSI Dataset. 'Regr.' stands for Regression.

Table 1 shows that the results of visual sentiment analysis on MOUD and MOSI dataset. It indicates that hvnLBP-TOP is an effective feature. The model performs better than all existing methods on MOUD dataset.

Conclusion

This paper proposes a novel facial expression feature for video sentiment analysis. We use a machine learning architecture to verify our proposed feature. The experiment results indicate the effectiveness of our feature.

Acknowledgements: Project (Grant No: 61673235) supported by National Natural Science Foundation of China.

References

Ahonen, T.; Hadid, A.; and Pietikainen, M. 2006. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12), 2037-2041.

Burkert, P.; Trier, F.; Afzal, M. Z.; Dengel, A.; and Liwicki, M. 2015. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*.

Hong, H.; Neven, H.; and von der Malsburg, C. 1998, April. Online Facial Expression Recognition Based on Personalized Galleries. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition* (p. 354). IEEE Computer Society.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Mistry, K.; Zhang, L.; Neoh, S. C.; Lim, C. P.; and Fielding, B. 2017. A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition. *IEEE transactions on cybernetics*, *47*(6), 1496-1509.

Pérez-Rosas, V.; Mihalcea, R.; and Morency, L. P. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 973-982).

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L. P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, *31*(6), 82-88.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L. P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1103-1114).

Zhao, G., and Pietikainen, M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 915-928.