LONG SHORT TERM MEMORY RECURRENT NEURAL NETWORK BASED ENCODING METHOD FOR EMOTION RECOGNITION IN VIDEO

Linlin Chao[1], *Jianhua Tao*[1, 2], *Minghao Yang*[1], *Ya Li*[1] *and Zhengqi Wen*[1]

[1]National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences
[2]Institute of Neuroscience, State Key Laboratory of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, CAS
{linlin.chao, jhtao, mhyang, yli, zqwen}@nlpr.ia.ac.cn

ABSTRACT

Human emotion is a temporally dynamic event which can be inferred from both audio and video feature sequences. In this paper we investigate the long short term memory recurrent neural network (LSTM-RNN) based encoding method for category emotion recognition in the video. LSTM-RNN is able to incorporate knowledge about how emotion evolves over long range successive frames and emotion clues from isolated frame. After encoding, each video clip can be represented by a vector for each input feature sequence. The vectors contain both frame level and sequence level emotion information. These vectors are then concatenated and fed into support vector machine (SVM) to get the final prediction result. Extensive evaluations on Emotion Challenge in the Wild (EmotiW2015) dataset show the efficiency of the proposed encoding method and competitive results are obtained. The final recognition accuracy achieves 46.38% for audio-video emotion recognition sub-challenge, where the challenge baseline is 39.33%.

Index Terms—Emotion Recognition, Facial Expression, Multimodal, Long Short Term Memory Recurrent Neutral Network

1. INTRODUCTION

Emotion recognition plays an important role in human machine interaction. It has gained increasingly attention [1]. Early researches mainly focus on utterance level speech emotion recognition or static image level facial expression recognition. However, emotion is a temporally dynamic event which can be better inferred from both audio and video feature sequences. This point of view is proved by cognitive researchers, where they argue that the dynamics of human behaviors are crucial for their interpretation [4]. Moreover, a number of recent studies [5-7] in affective computing demonstrate this point of view.

Previous works on video based category emotion recognition for temporal modeling can be broadly categorized into two groups. One is extracting spatial temporal features. These spatial temporal features include Local Binary Patterns from three Orthogonal Panels (LBP-TOP) features [5], Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) [28] and so on. This kind of modeling is more suitable to short range modeling, often several successive frames. Also, the temporal length of the spatial temporal features is fixed. After dynamic feature extraction, sequence modeling, often chosen to be the pooling based methods, is still needed. The other one is encoding features extracted from each image in an image sequence to a fixed length feature representation. Among these encoding methods, pooling among all the frames is one of the widely utilized methods. However, the pooling based encoding methods lose information from successive frames inevitably [29]. Besides, Riemannian manifolds are introduced into video based emotion recognition by Liu et al. [7]. They utilize Riemannian manifolds to represent the similarity distance measurement for the image based features from video. Although discriminative emotion clues among frames are utilized, the single frame's emotion clues are not fully considered.

In this paper, we investigate the sequence encoding method for emotion recognition with LSTM-RNN. Emotion clues from single frames and successive frames can be encoded into the hidden representation of LSTM-RNN layer together through its gate mechanism. After LSRN-RNN encoding, the variable-length feature sequence is mapped to a fixed length vector for further emotion classification.

2. AUDIO VIDEO ENCODING METHODS

Two kinds of encoding methods are introduced. The first one is LSTM-RNN based method. The other one is temporal pooling [21], which is utilized as one of the baselines.

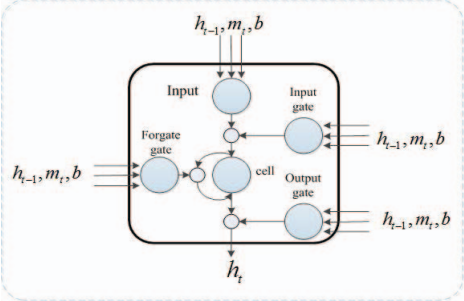

Fig.1. Architecture of the memory block utilized in LSTM.

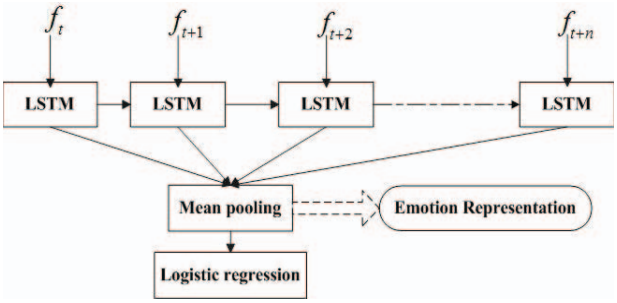

Fig.2. Training and evaluation diagram of the LSTM-mean encoding method.

2.1 LSTM-RNN based encoding methods

LSTM-RNN is a kind of recurrent neural network which has the ability to learn long-term dynamic while avoiding the vanishing and exploding gradients problems. In a standard recurrent neural network, given the input sequence **m**= $(m_1, \dots, m_T)$, the hidden vectors **h** = $(h_1, \dots, h_T)$ and output vectors **y** = $(y_1, \dots, y_T)$ are computed as:

$$h_t = f_{act}(W_{mh} m_t + W_{hh} h_{t-1} + b_h) \quad (1)$$

$$y_t = f_{out}(W_{hy} h_t + b_y) \quad (2)$$

where f_{act} is the activation function of hidden layer, often chosen to be the tanh function. f_{out} represents the activation function of the output layer.

In LSTM-RNN, f_{act}in equation (1) is replaced by the specially designed memory block. Each block contains one or more recurrently connected memory cells and three multiplicative units, the input, output, and forget gates, which control the information flow inside the memory block. The surrounding network can only interact with the memory cells via the gates. As research going on, several kinds of memory blocks are put forward. We use the memory block as described in [18] (Fig.1).

Two kinds of LSTM-RNN based encoding methods are investigated in this paper. One is the

$$f_{encod1}(h_1, \dots, h_T) = h_T \quad (3)$$

For each sequence, the last hidden vector is chosen as the final representation for post processing. We call this encoding way as LSTM-last encoding. This is an N to one mapping way. Only the hidden vector of the last time step is fed into the output layer.

The other one is represented as

$$f_{encod2}(h_1, \dots, h_T) = \sum_{i=1}^{T} h_i \quad (4)$$

All the hidden vectors from different time steps are averaged for the output layer. We call this encoding way as LSTM-mean encoding. Fig.2. depicts the training and evaluation diagram of this encoding method.

2.2 Temporal pooling encoding

Temporal pooling is investigated to compare with LSTM-RNN based encoding methods. It is put forward by [21], which is used for music classification. It has been introduced into dimensional emotion recognition for temporal modeling [6]. Different from traditional pooling encoding methods, it is utilized as a pooling layer in the neural network after one or several hidden layers. Through the end-to-end training way, better features can be pooled. In this context, feature sequence from one video is fed into the neural network, and after several hidden layers, the temporal pooling layer pools all the hidden representations to one single vector. This vector is the encoding vector for emotion representation.

2.3 Audio video fusion

Both feature level fusion and decision level fusion are investigated. The decision level fusion is implemented by averaging the predicted probabilities from each feature set. Feature level fusion is achieved by concatenating the encoded feature vectors, which are corresponding to different feature sets, and feeding the concatenated vector to the final classifier.

3. DATASET AND FEATURE SET

The EmotiW2015 [27] provides the common benchmarks for emotion recognition researchers, which mimics real-world conditions. There are two sub-challenges: audio-video based emotion recognition challenge based on the Acted Facial Expression in the Wild (AFEW) database [8] and image based static facial expression recognition challenge based on SFEW database [9]. The emotion recognition challenge contains audio-video short clips labeled using a semi-automatic approach defined in [8]. The task is to assign a single emotion label to the video clip from the six universal emotions (Anger, Disgust, Fear, Happiness, Sadness and Surprise) and Neutral. The databases (AFEW and SFEW) are divided into three sets for the challenge: training, validation and testing. The training and validation sets are utilized to train the emotion recognizer. Prediction results on testing set are utilized to rank participants.

3.1 Face shape feature

For video features, we mainly focus on the face part. As the face shape provides import clues for facial expression [11], we use the landmarks' location of the face as face shape

feature. After feature normalization for each clip, these features can also reflect the head movement and head pose.

3.2 Face appearance feature

For face appearance feature, we utilize the features extracted from a convolutional neural network (CNN) model. Previous work [12] shows that CNN model trained via ImageNet dataset [20] can be generalized well in many other tasks. Liu et al. [7] utilize the CNN model trained via face recognition dataset to extract face representation. And this representation can be generalized to facial expression recognition problem. We employ the same strategy from [7] to train a CNN model from Celebrity Faces in the Wild (CFW) [13] and Facescurb [14] dataset, which are designed for face recognition tasks. Over 110,000 face images from 1032 people are used for training and the labels are their identities. The architecture is the same with [15]. There are three fully connected layers and five convolutional layers. Compared to the three fully connected layers, convolutional layers have better generation performance [12]. The neurons of different convolutional layers have different mapped ranges in the original image, the higher the larger. The higher layers extract more abstract features [16]. Recently, study [17] also shows that the features extract from lower layers can also be useful. Thus, we extract the feature from the 5th pooling layer, the 4th convolutional layer and the 3rd convolutional layers as three different face appearance feature sets, which are represented as pool5, conv4 and conv3 respectively. With features extracted from lower layers, more detailed description from part of face region can be extracted for fine-grained facial expression classification.

3.3 Audio feature

We utilize the YAAFE toolbox [10] to extract audio features. All the waves are resampled to 16 kHz and 27 features are extracted. More details of these features can be found in [10]. There are also four feature transforms used on MFCC [30] features. The feature transforms include the first and second derivate, cepstrum and auto correlation peaks integrator. All the features are extracted in the default parameters. Finally, 939 dimensions features are extracted for each frame and the frame length is 1024. The audio features are then PCA whitened [26], with the final 50 dimensions are kept.

4. EXPERIMENTS

4.1 Experiment setup

We follow the challenge criterion of EmotiW2015 to utilize training set, validation set and testing set. We utilize the landmarks provided by the organizers for the shape feature. Caffe [24] implementation of CNN is utilized to extract face appearance features, where the face image is provided by the organizers.

The dimension of face appearance features is high. The dimension numbers of pool5, conv4, and conv3 are 9216, 43264 and 69984 respectively. Meanwhile, the training data is relatively small. Thus we employ random forest algorithm implemented by scikit-learn [25] for feature selection and 2048 features are kept for each feature set. The random forest classifier is trained via the SFEW database, where one of the seven emotion labels is assigned to a single static face image.

For LSTM-RNN encoding and temporal pooling encoding, we use the implementation from Theano [22] [23]. The architectures of the networks keep the same except the number of nodes in the input layer, with one hidden layer, one LSTM layer or temporal pooling layer and softmax regression layer. There are 64 memory cells are utilized in the LSTM layer and 64 nodes in the first hidden layer. The maximum training epoch is 100 with dropout regularization technique utilized in all layers except the LSTM layer. The drop rate is 0.5. Weight decay in the softmax regression layer with the parameter 0.0005 is applied to prevent over fitting. The best results are chosen by the prediction accuracy in the validation set.

As this dataset is collected in the wild environment and video signal is the mainly collection modality, the background of audio signal is very noisy and many subjects in the video speak nothing. It is difficult for LSTM-RNN to learn the dynamic evolving of successive frames from audio. Thus we use temporal max-pooling instead. The architecture is one hidden layer, one temporal max-pooling layer and softmax regression layer. There are 128 nodes in the hidden layer and the temporal pooled hidden representation is the encoded vector for further processing.

SVM is the classifier utilized for feature level fusion. It classifies the concatenated vectors obtained from each feature set. LibLinear [19] implementation of SVM is utilized. We employ linear kernel and one verse the rest strategy in all experiments. The penalty parameter C is chosen by grid search method.

Table 1 Experiment results on validation set with the pool5 feature set.

Encoding method	Network topology	Accuracy
LSTM-mean	2048/64/64/7	0.4420
LSTM-last	2048/64/64/7	0.3909
temporal mean pooling	2048/64/7	0.4394
temporal max-pooling	2048/64/7	0.4474
temporal max-pooling	2048/256/7	0.4528
max-pooling	2048/64/7	0.4339
mean pooling	2048/64/7	0.4367

4.2 Encoding methods comparison

We compare the encoding methods based on the pool5 feature set. Comparison results are shown in table 1. The results show that LSTM-mean encoding performs much better than LSTM-last encoding. This can be explained that every frame of the feature sequence has contribution to final classification and the hidden representation of last timestep has limitation to fully contain the total sequence's

emotion clues. When compare LSTM-mean encoding with temporal mean pooling encoding, both of this two methods contain pooling operation. The difference is the pooled features. One contains dynamic changes among successive frames, while the other is single frame based. The improved performance demonstrates that LSTM layer has the ability to contain emotion clues form dynamic changes among successive frames, which is ignored by pooling based encoding method. The experiments results also show that temporal max-pooling performs better than LSTM-mean and temporal mean pooling. Increasing the number of nodes in hidden layer improves the performance further. The results of traditional mean pooling and max-pooling method are also compared. Both temporal pooling method and LSTM-mean show superior performance. Thus, we utilize the LSTM-mean encoding and temporal max-pooling method for video feature sets to submit result on testing set.

Table 2 Experiment results on validation set and testing set for Audio-Video Emotion Recognition sub-challenge

Method	Accuracy	
	Val	Test
Audio-temporal max-pooling	0.3864	
Landmarks-LSTM-mean	0.4286	
Pool5-LSTM-mean	0.4420	
Conv4-LSTM-mean	0.4420	
Conv3-LSTM-mean	0.4420	
Decision level fusion (LSTM-mean)	**0.4933**	0.4434
Feature level fusion (LSTM-mean)	0.4852	**0.4638**
Decision level fusion (temporal max-pooling)	0.4798	0.3915
Feature level fusion (temporal max-pooling)	0.4367	
Challenge Baseline [27]	0.3608	0.3933

4.2 Final results and analysis

Experiments results on both validation set and testing set for audio-video emotion sub-challenge are shown in Table 2. Among the five feature sets, face appearance feature sets get the best performance. The three appearance based feature sets get the same accuracy on the validation set. The face shape based feature set performs better than the audio modality.

Combining the five feature sets, the results improve significantly. Decision level fusion, which averages the prediction results of the LSTM-mean method (video feature sets) and temporal max-pooling method (audio feature set) shows the best performance on validation set and over fits on the testing set. The best result is obtained by feature level fusion with LSTM-mean encoding method applied for video feature sets. The results prove that the emotion representations from different feature sets are discriminative for emotion classification, although these emotion representations only have 64 dimensions for each video feature sequence. This implies the effectiveness of the proposed LSTM-RNN encoding method. More details of the best submitted result are shown in Fig.3. The confusion matrix shows that "Angry", "Happy" and "Neutral" are easier to classify. The "Disgust" and "Fear" are easy to be misclassified to "Angry". "Surprise" is easy to confuse with "Sad". The reason may lie in the data set distribution is not balance. And fine grained classification among "Angry", "Disgust" and "Fear" needs more effort.

The results of purely temporal max-pooling based encoding method do not show comparative performance as LSTM-mean encoding method with both decision level fusion and feature level fusion. Although the decision level fusion result is closer to LSTM-mean encoding method on validation set, it over fits on testing set.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**59.49%**	0.0%	2.53%	3.80%	12.66%	12.66%	8.86%
Disgust	**24.14%**	10.34%	0.0%	17.24%	17.24%	20.69%	10.34%
Fear	**33.33%**	0.0%	6.06%	3.03%	13.64%	12.12%	31.82%
Happy	9.26%	0.0%	0.93%	**65.74%**	4.63%	16.67%	2.78%
Neutral	10.06%	1.26%	1.89%	7.55%	**52.83%**	16.98%	9.43%
Sad	11.27%	0.0%	0.0%	19.72%	11.27%	**45.07%**	12.68%
Surprise	14.81%	3.70%	3.70%	3.70%	3.70%	**37.04%**	33.33%

Fig.3. Confusion matrix of the testing set result from feature level fusion (audio: temporal max-pooling, video: LSTM-mean).

5. CONCLUSIONS

In this paper, LSTM-RNN based encoding method for audio video emotion recognition is proposed. Each emotion video sequence can be encoded into a single low dimensional vector. This vector combines the discriminative power from single image and temporal context from image sequence together, which is highly discriminative for emotion classification. Comparison results with the widely utilized pooling method and temporal pooling results show the proposed method has better modeling capacity.

Although emotion clues from successive frames are encoded and better performances are achieved, the proposed method still needs a pooling operation for the hidden representations of LSTM-RNN. In the future, we will try to utilize pyramid LSTM-RNN architecture to reduce the influence of pooling operation.

ACKNOWLEDGMENTS

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61332017, No.61375027, No.61305003, No.61203258, 61273288), and the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES

[1] J. Tao and T. Tan, "Affective Computing: A Review", Proc. First Int'l Conf. Affective Computing and Intelligent Interaction, J. Tao, T. Tan, and R.W. Picard, eds., pp. 981-995, 2005.

[2] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58. doi:10.1109/TPAMI.2008.52.

[3] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition", Int. J. Synthetic Emotions, vol. 1, no. 1, pp. 68-99, 2010.

[4] A. Mehrabian, and J. Russell, "An approach to environmental psychology". Cambridge, MA: MIT Press.

[5] G. Zhao and M. Pietikainen. "Dynamic texture recognition using local binary patterns with an application to facial expressions". Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(6):915–928, 2007.

[6] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen, "Multi-scale Temporal Modeling for Dimensional Emotion Recognition in Video", AVEC@MM 2014: 11-18.

[7] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang and X. Chen, "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild", ICMI 2014.

[8] A. Dhall, R. Goecke, S. Lucey and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies", IEEE Multimedia 2012.

[9] A. Dhall, R. Goecke, S. Lucey and T. Gedeon, "Static Facial Expression in Tough Conditions: Data, Evaluation Protocol and Benchmark", IEEE ICCV BEFIT Workshop 2011.

[10] B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software", proceedings of the 11th ISMIR conference, Utrecht, Netherlands, 2010.

[11] H. Gunes, M. Piccardi, and M. Pantic, From the Lab to the Real World: Affect Recognition Usng, Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition. I-Tech Education and Publishing, Vienna, Austria, pp. 185 - 218, 2008.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". arXiv preprint arXiv:1311.2524,2013

[13] X. Zhang, L. Zhang, X.-J. Wang and H.-Y. Shum. "Finding celebrities in billions of web images". Multimedia, IEEE Transactions on, 14 (4):995–1007, 2012.

[14] H.-W. Ng, S. Winkler. "A data-driven approach to cleaning large face datasets". Proc. IEEE International Conference on Image Processing (ICIP), Paris, France, Oct. 27-30, 2014.

[15] Krizhevsky, A., Sutskever, I., Hinton. G., "ImageNet Classification with Deep. Convolutional Neural Networks", NIPS 2012.

[16] M.D. Zeiler, R. Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014, Arxiv 311.2901,2013

[17] Sun Y, Wang X, Tang X. "Deeply learned face representations are sparse, selective, and robust" [J]. arXiv preprint arXiv:1412.1265, 2014.

[18] Zaremba W, Sutskever I. "Learning to execute" [J]. arXiv preprint arXiv:1410.4615, 2014.

[19] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin. "LIBLINEAR: A library for large linear classification", Journal of Machine Learning Research (2008), 1871-1874.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", IJCV, 2015

[21] P. Hamel, S. Lemieux, Y. Bengio, and D. Eck, "Temporal pooling and multiscale learning for automatic annotation and ranking of music audio". In ISMIR (pp. 729-734), 2011.

[22] Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Bergstra, James, Goodfellow, Ian, Bergeron, Arnaud, Bouchard, Nicolas, and Bengio, Yoshua. "Theano: new features and speed improvements". NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2012.

[23] Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. "Theano: a CPU and GPU math expression compiler". In Proceedings of the Python for Scientific Computing Conference (SciPy), June 2010.

[24] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv: 1408. 5093, 2015.

[25] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[26] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning", ICML Unsupervised and Transfer Learning, 2012: 17-3

[27] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi and Tom Gedeon, "Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015", ICMI 2015.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa. "Multi resolution gray scale and rotation invariant texture classification with local binary patterns". Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(7):971–987, 2002.

[29] Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G. and Bartlett, M. (2013). "Multiple Kernel Learning for Emotion Recognition in the Wild". Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI13).

[30] S.B. Davis and P.Mermelstrin, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". IEEE Transactions on Acoustics, Speech and Signal Processing, 28 :357-366, 1980.