Facial Expression Recognition in Videos

An CNN-LSTM based Model for Video Classification

Muhammad Abdullah
Vision and Image Processing Lab,
Department of Computer Engineering,
Sejong University
Seoul, South Korea
m.abdullah.rise@gmail.com

Mobeen Ahmad
Vision and Image Processing Lab,
Department of Computer Engineering,
Sejong University
Seoul, South Korea
ehmedmobeen@gmail.com

Dongil Han*
Vision and Image Processing Lab,
Department of Computer Engineering,
Sejong University
Seoul, South Korea
dihan@sejong.ac.kr

Abstract—**Facial Expressions are an integral part of human communication. Therefore, correct classification of facial expression in image and video data has been an important quest for researchers and software development industry. In this paper we propose the video classification method using Recurrent Neural Networks (RNN) in addition to Convolution Neural Networks (CNN) to capture temporal as well spatial features of a video sequence. The methodology is tested on The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Since no other results were available on this dataset using only visual analysis, the proposed method provides the first benchmark of 61% test accuracy on given dataset.**

Keywords—facial epxression recognition; video classification; recurrent neural networks; temporal feautres; HRI

I. Introduction

Facial Expression not only play a vital role in daily life communication like showing anger, happiness and sorrow but they also provide a lot of hidden and non-verbal information for important tasks like video surveillance and monitoring, video summarization for movies trailer generation and creation of highlights of an event and last but not the least human robot interaction / human machine interaction. Lots of work has been done on facial expression recognition in images [1] but when it comes to the real time perception or video analysis, we are supposed to guess the expression based on multiple frames. It is commonly observed individual frames are not enough to provide the correct classification of a person's expressions such as during a conversation or performing a task many frames pass through the overlapped regions of different facial expression and it's hard to guess the expression by looking at those individual frames even for a human observer until they are shown the frames before and after the frame under observation e.g. expression of disgust, anger and fear. Moreover, facial expressions might be our criteria to search but we usually need to include a continuous sequence of frames in our output. For example, in order summarize all the sad dialogues in a movie, it will not suitable to include abruptly changing chunks of frames, rather we might want to include the whole dialogue even if the sadness was visible on face only for last a few frames. There have been several research publications made on video content understanding. Most of these publications provided methods to use handcrafted features along with machine learning for action classification in the videos [2-4]. While latest trends suggest using state-of-the-art deep learning techniques for video classification. Some of these approaches are discussed further in related work. In the end, an end to end system using a CNN followed by an RNN is proposed to Video Facial Expression Recognition (VFER).

II. Related Work

A. Support Vector Machine (SVM) Based Classification

The features obtained from deep learning proved to be highly successful in image classification. This inspired many researchers to utilize these trained features for video classification. In this approach feature representation for each video frame is extracted by passing it through a state-of-the-art deep learning model till an intermediate fully connected layer. Collective features for all the frames are then classified at video-level using a well-known classifier such as SVMs [5]. Some later works also used the advanced feature encoding strategies such as vector of locally aggregated descriptors (VLAD) encoding [6] and Fisher Vector encoding with Variational AutoEncoder (FV-VAE) [7].

B. 3-Dimentional CNN Architectures

The ability of CNNs to learn features from raw data as an end-to-end pipeline make it suitable to used CNN based deep learning models on the video data to learn spatio-temporal patterns. For this purpose, traditional 2D CNNs were extended to introduce 3D CNNs that utilize 3-dimnetional kernels for convolution on stacked video frames [8]. However, their performance was worse than the state-of-the-art handcrafted features.

C. Bi-stream CNNs Systems

The basic idea behind the first two stream method was to divide the learning of video features into two parts [9]. First a 2D CNN is used to extract spatial features from individual video frames. Then the optical flow is computed from adjacent video frames using displacement vector fields via temporal CNN. The final prediction is generated as the weighted sum of

Corresponding Author*: Dongil Han (Professor, Sejong University) probability scores from both CNN streams. The method was successfully used to achieve benchmark performance on action recognition task in videos. Various extensions have been made to this approach since then e.g. replacing the optical flow with motion vectors [10], fusion of classification score via convolutional layer [11] and linear combination of classification score [12].

D. *Recurrent Neural Networks*

The bi-stream approach captures the motion features to represent the action for a very small duration. Additionally, it fails to preserve the frames order information, during the training [13]. However, most of the video analysis requires understanding the events or actions happening over a longer period. These actions/events may even be further divided into several sequential actions or events e.g. a cooking show teaching a recipe. Recurrent neural networks such as LSTM on the other hand demonstrated good performance on a variety of tasks such as speech and text analysis where temporal information is critical. In addition to that LSTM does not suffer from gradient vanishing. Even a simple 2-layer LSTM demonstrated good results for action recognition task [14].

III. Methodoloy

As discussed earlier in related work, CNNs perform very well on image classification tasks. So, it is very intuitive to use the power CNNs to detect spatial features, followed by an RNN to learn the temporal features from sequences of these spatial features. Effectively making a CNN-RNN or CRNN system. We decided to test a simple, single layer LSTM to learn the temporal features from stacked spatial features extracted by a CNN for each video.

A. *CNN Training*

Since the focus of our research to classify videos based on facial expressions. Therefore, it was intuitive to train the CNN model to extract emotion features rather than using general object classifiers such as pretrained CNNs on ImageNet [16]. So FER2013 [17] was used as CNN training dataset. In order to achieve better results the Xception Net [15] was pretrained on VGG-Face dataset [18] and later it was trained on FER2013 dataset leveraging the transfer learning.

B. *Spatial Feature Extraction*

- Since our CNN is trained to detect expressions from human faces, it is important to extract frame only where face is visible. Therefore, all the frames are pre-processed to detect faces in the video sequence.

- Frames are then cropped to contain detected faces only.

- These cropped images are resized and converted to grayscale to match the shape and channel specification as of FER2013.

- Output of previous step is fed to a feed forward pass of CNN until an intermediate fully connected layer.

- Output of this fully connected layer represents the emotion features at that layer. Effectively chopping off the top classification part of the network so that we end up with a vector of features.

- Now the those extracted features are converted into sequences of features. For this purpose, features of each individual frames are stacked together for every video and save to the disk.

C. *Long-Short Term Memory*

LSTM is a special RNN that allows the network to learn long-term dependencies. We used a single LSTM layer followed by a Dropout with some dropout in between. The sequences of extracted features in previous step are then fed to this RNN along with the ground truth labels for training purpose. This simple and shallow network outperformed all the methods we took into consideration at in related work study.

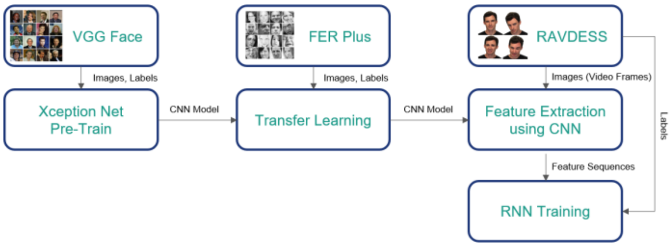


Fig. 1. Pipeline for Video Classification Model Training

IV. Dataset

For validation of the proposed method we used The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [19]. The dataset is a dynamic, multimodal set of facial and vocal expressions in North American English. Though this dataset provides multimodal information using visual expressions and audio features. It contains 4904 video files (720p) from 7 facial expression classes named calm, happy, sad, angry, fearful, surprise, and disgust expressions. Since our goal is to validate our system only for visual facial expression analysis, so we used the data in video-only mode to validate our proposed method. To follow a standard 80:10:10 training, validation and test ratio respectively 3923 videos were used for training, 490 videos for validation while 491 videos were used for test purpose.

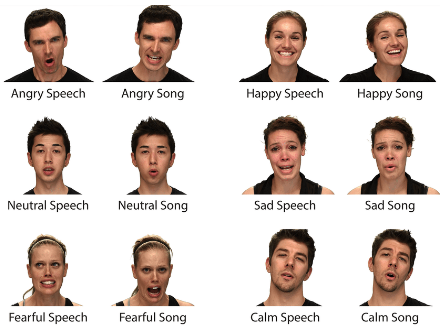
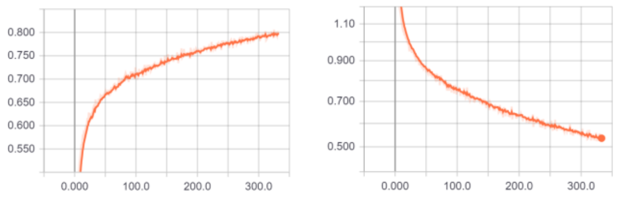


Fig. 2. Sample frames from videos of different expression classes. (Image courtesy Original Dataset Publication)

V. Results

The Xception Net used to extract features from the frames achieved 80% validation accuracy in FER2013 dataset. The validation accuracy and validation loss for 330 epochs is plotted in the graphs below. It can be observed that accuracy was still increasing with time and we believe with further parameter tuning and training ever higher accuracy can be achieved. However currently trained model is good enough to validate our proposed method.

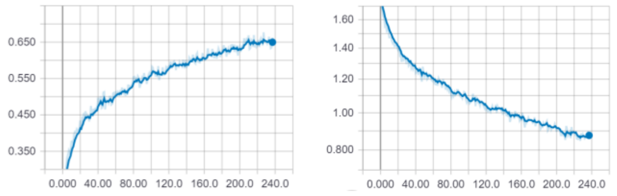


a. Validation Accuracy

b. Validation Loss

Fig. 3. Results of XceptionNet model training on FER2013 using Keras.

The single layer LSTM network also demonstrated very good performance on extracted feature sequences and achieved 65% validation accuracy just after training for 240 epoches.



a. Validation Accuracy

b. Validation Loss

Fig. 4. Training results of LSTM based RNN.

The trained model was able to classify test videos with 61% accuracy. Since no previous publication has demonstrated video classification performance on this dataset using only video frames, this can be considered as the first benchmark on RAVDESS video dataset using only visual features.

VI. Conclusion

The proposed method proves that using a well-trained CNN followed by RNN is equally effective for Video Facial Expression Recognition as for other similar tasks e.g. Action Recognition. The effects of attention-based networks and implications in continuous video analysis will be studied in future. Videos provide a great insight in a lot of domains and provide endless possibilities to develop smart technologies based on visual perception. The work provides the base ground to extend its application in various important tasks such as suspicious behavior detection, video summarization and emotionally aware interactive e-learning solutions.

Acknowledgment

This work was supported by a National Research Foundation of Korea Grant funded by the Korean Government (NRF2017R1D1A1B03028394).

References

[1] Sharma, Archana Kumari, Umesh Kumar, Sandeep K. Gupta, Uma Sharma, and Shubh LakshmiAgrwal. "A Survey on Feature Extraction Technique for Facial Expression Recognition System." In 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1-6. IEEE, 2018.

[2] Aggarwal, Jake K., and Michael S. Ryoo. "Human activity analysis: A review." ACM Computing Surveys (CSUR) 43, no. 3 (2011): 16.

[3] Poppe, Ronald. "A survey on vision-based human action recognition." Image and vision computing 28, no. 6 (2010): 976-990.

[4] Jiang, Bihan, Michel Valstar, Brais Martinez, and Maja Pantic. "A dynamic appearance descriptor approach to facial actions temporal modeling." IEEE transactions on cybernetics 44, no. 2 (2013): 161-174.

[5] Littlewort, Gwen, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. "Dynamics of facial expression extracted automatically from video." In 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 80-80. IEEE, 2004.

[6] Reddy, Mopuri K., Sahil Arora, and R. Venkatesh Babu. "Spatio-temporal feature based vlad for efficient video retrieval." In 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp. 1-4. IEEE, 2013.

[7] Sun, Chen, and Ram Nevatia. "Large-scale web video event classification by use of fisher vectors." In 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 15-22. IEEE, 2013.

[8] Ji, Shuiwang, Wei Xu, Ming Yang, and Kai Yu. "3D convolutional neural networks for automatic human action recognition." U.S. Patent 8,345,984, issued January 1, 2013.

[9] Girdhar, Rohit, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. "Actionvlad: Learning spatio-temporal aggregation for action classification." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 971-980. 2017.

[10] Zhang, Bowen, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. "Real-time action recognition with enhanced motion vector CNNs." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2718-2726. 2016.

[11] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1933-1941. 2016.

[12] Wu, Zuxuan, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification." In Proceedings of the 23rd ACM international conference on Multimedia, pp. 461-470. ACM, 2015.

[13] Singh, Bharat, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao. "A multi-stream bi-directional recurrent neural network for fine-grained action detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1961-1970. 2016.

[14] Wang, Limin, Wei Li, Wen Li, and Luc Van Gool. "Appearance-and-relation networks for video classification." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1430-1439. 2018.

[15] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258. 2017.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database.” IEEE Computer Vision and Pattern Recognition (CVPR), 2009

[17] Carrier, Pierre-Luc, Aaron Courville, Ian J. Goodfellow, Medhi Mirza, and Yoshua Bengio. "FER-2013 face database." Technical report (2013).

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference, 2015

[19] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391.

.