STeW: Real-time Video Facial Emotion Classification via a Compact Sliding Temporal Windowed Deep Neural Network

James Lee
Alexander Wong
Email: {jrhlee, a28wong}@uwaterloo.ca

University of Waterloo
University of Waterloo, Waterloo AI Institute

Abstract

The real-time classification of human facial expressions presents a challenging task, even for humans. Individuals with Autism Spectrum Disorder (ASD) have an even greater difficulty in detecting and interpreting these facial expressions, which can lead to an increased risk of depression and loneliness due to a disconnect with society. This study explores a compact Sliding Temporal Windowed (STeW) deep neural network architecture for real-time video facial emotion classification. The proposed STeW architecture is designed to provide a balance between speed and the leveraging of temporal characteristics to capture transient nuances of facial expressions. A more difficult dataset (which we call BigFaceX) is proposed by combining and modifying the extended Cohn-Kanade (CK+), BAUM-1, and the eNTERFACE public datasets, and used to evaluate the proposed STeW network. Experimental results show that the proposed STeW network architecture can achieve noticeably higher accuracy when compared to the highly compact mini-Xception network, thus illustrating the potential for leveraging this approach to achieve real-time video facial emotion classification.

1 Introduction

Understanding facial expressions is an important skill to have when interacting with other individuals, not just to understand current emotional states, but also to recognize conversational cues such as level of interest, speaking turns, and level of information understanding [1]. Research has shown that 55% of the information behind a spoken message is attributed to facial expressions [1]. However, individuals with Autism Spectrum Disorder (ASD) have difficulties detecting and interpreting these facial expressions. This can lead to difficulties communicating and interacting with other individuals, which often cause a higher prevalence of loneliness and depression. With a real-time emotion classification system in place to support these individuals with ASD, we can help alleviate some of their difficulties and assist them with societal interaction.

Much work has been done in the field of facial emotion classification, employing a variety of techniques such as CNNs [2], RNNs [3], DNNs [2], and SVMs [1]. These techniques generally fall under one of two categories: i) static image classification, and ii) dynamic video classification. In the first case, a single frame is provided, and a label is produced, usually one from the six basic universal emotions (anger, disgust, fear, happiness, sadness, surprise) [1, 2] plus neutral. For video classification, an entire video clip is passed in as input, and a *single* label is produced.

Real-time video classification generally uses a frame-by-frame approach, where the most recent frame is passed as a single image to produce a label [1]. However, facial expressions are of a dynamic nature [2], and thus the onset and offset phases can easily be confused for other emotions, with only the peak of the expression having a relatively high classification accuracy. Due to this, temporal information becomes important for capturing the smooth transition between expressions as well as the transient nuances of facial expressions. In this study, we explore a compact Sliding Temporal Windowed (STeW) deep neural network architecture in order to better capture these temporal characteristics, for the purpose of real-time facial emotion video classification.

2 Methodology

To achieve a compact network architecture, the proposed STeW network architecture extends upon the mini-Xception architecture [4], which was designed to achieve reasonable accuracy while having a low parameter count, reporting a 66% accuracy on the FER2013 [5] dataset with just ∼60K parameters. In order to take advantage of transient nuances in facial expressions via temporal video information, the input convolutional layer is designed such that a sliding temporal window consisting of multiple adjacent video frames are taken as input and mixed to produce new spatiotemporal features for the subsequent layer. In this study, the sliding temporal window consists of 5 frames in order to decrease inference time while maintaining temporal information. Naturally, each window of frames must be sequential. Famous static image datasets such as FER2013 are not well-suited for this task, as each image is unrelated to the others. However, datasets such as the CK+ dataset [6] consists of labelled expression videos saved as sequential face images. To produce a more difficult dataset for evaluating real-time video facial emotion classification, we propose a custom dataset called BigFaceX, where we combine the CK+, BAUM-1 [7], and eNTERFACE [8] video datasets. To create data samples from these datasets, we extract temporal sliding windows consisting of 5 frames each across each sequence in these datasets. The label for each sliding temporal window is the original label of the video it was taken from. For BAUM-1 and eNTERFACE, the first 5 frames of each video were ignored, as the subject does not start the expression until a few frames in, and a stride of 2 was used, in order to increase temporal width while maintaining a small window size. Each image in each window was then cropped to the facial bounding box using a facial bounding box algorithm, and resized to 48x48 pixels. In total, 69,373 samples were extracted to create BigFaceX, with 80% for the training set and 20% for the test set. For comparison, mini-Xception was trained with BigFaceX and FER2013.

3 Results and Discussion

A number of observations can be made based on the experimental results. First, the proposed STeW network achieved an accuracy of 88.2% on the BigFaceX test dataset. This is noticeably higher than the 75.7% accuracy achieved by the mini-Xception network. Second, leveraging the proposed BigFaceX dataset enabled both networks to achieve strong accuracies, as the BigFaceX dataset more accurately simulates inputs from a real world video setting, since it includes the onset, peak, and offset phases of an expression whereas static datasets such as FER2013 only include the peak, thus providing a better reflection of real-world video facial emotion classification performance. Further work includes an extension of the STeW architecture for improved accuracy as well as improved mechanisms for training using video information.

Acknowledgements

The authors would like to thank NSERC, Canada Research Chairs program, and Microsoft for their support.

References

[1] P. Michel and R. El Kaliouby, “Real time facial expression recognition in video using support vector machines,” in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 258–264.

[2] B. Hasani and M. H. Mahoor, “Facial expression recognition using enhanced deep 3D convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 30–40.

[3] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using CNN-RNN and C3D hybrid networks,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.

[4] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, “Real-time convolutional neural networks for emotion and gender classification,” *arXiv preprint arXiv:1710.07557*, 2017.

[5] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, “Challenges in representation learning: A report on three machine learning contests,” in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

[7] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “BAUM-1: A spontaneous audio-visual face database of affective and mental states,” *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[8] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 audio-visual emotion database,” in *22nd International Conference on Data Engineering Workshops (ICDEW’06)*. IEEE, 2006, pp. 8–8.