Using 3D-CNNs for Emotion Perception in Intelligent Tutoring Systems

Muhammad Zuhayr Raghib   Ajjen Joshi   Margrit Betke
Boston University
{mzraghib, ajjendj, betke}@bu.edu

Abstract

In this paper, we investigate the estimation of human emotion using non-contact computer vision methods. We explore two approaches, one taking a per-frame classification approach using OpenFace; an open source facial recognition tool, and one using a novel approach where we utilize 3D Convolutional Neural Networks to extract spatio-temporal features from a set of concatenated frames for emotion perception. This work focuses on a video dataset of students giving computerized tests, specifically prepared for training and testing our methods. We build on the work by Kensho Hara et al. *[7], to fine-tune a 3D ResNeXt architecture, that achieves an AUC score of 0.79 in the Chalearn dataset, which is equivalent to being ranked 7th in the Chalearn Looking at People First impressions Challenge, and an AUC score of 0.63 on the effort labels from the student dataset, outperforming the use of OpenFace by 3%. The same method is then applied on the emotion labels in the student dataset, achieving an AUC score of 0.63 for 'Interest' and 0.6 for 'Confidence'.*

1. Introduction

The task of emotion detection in the context of student engagement during interaction with computerized tutoring systems is a particularly difficult one. Firstly, there does not exist a large database of training data, which is key for supervised learning using deep neural networks. Facial expressions also vary dynamically and unlike commonly recognized emotions like 'happy' or 'sad', the features for 'engagement', 'confidence' or related emotions are more complex, and less observable and verifiable by human observation (see Figure 1).

To date, a number of attempts have been made to classify facial emotions in images [8], however to our best knowledge, this is the first attempt to approach this problem as an action recognition problem, and use very deep 3D CNNs, leveraging the availability of large action recognition datasets, to pre-train the models beforehand, to gain the advantages of complex features detected by deep models, on the limited data that is collected for our work.

A model with the ability to detect complex emotions from facial data would be advantageous for a wide variety of applications [1] and the effectiveness of the results of this study could be used to further progress related work.

2. Related Work

Kensho Hara *et al.* [7] examine the architectures of multiple CNNs with spatio-temporal 3D convolutional kernels on current video datasets including the UCF-101, HMDB-51 and ActivityNet. They also show a significant improvement in performance by pre-training models on the kinetics dataset prior to fine-tuning. The model with the best performance over all of their experiments for action recognition was the 3D-ResNeXt-101.

Through their two streamed I3D model which utilizes simple two-stream 3D architectures pretrained on Kinetics, J. Carreira and A. Zisserman [3] show that 3D convnets directly learn hierarchical representations of spatio-temporal data, although they have many more parameters than 2D models because of the additional kernel dimension.

D. Tran *et al.* [4] show that their C3D can model appearance and motion information simultaneously and outperforms the 2D ConvNet features on various video analysis tasks.

3. Datasets

In the experiments, we first used the Chalearn dataset [3] to achieve the best possible classification results on the test data. When satisfactory results were achieved, the resulting model was used for classification on our developed student dataset.

3.1. Chalearn Dataset

The ChaLearn Looking at People - first impressions dataset to evaluate our system. This dataset comprises of 10000 clips (average duration 15s) extracted from more

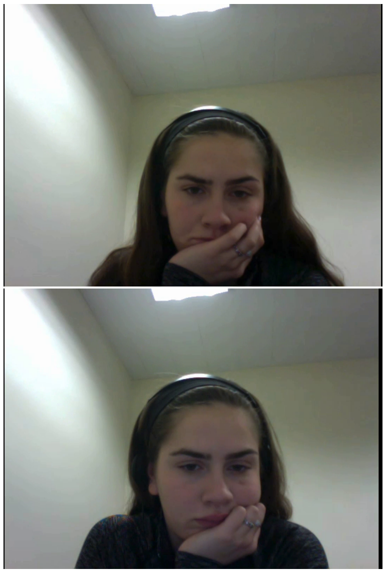

Figure 1. Example frames from student dataset. The subject is in the process of solving mathematics questions on computerized tutoring software. The top picture is an example of an emotion label 'interest', while the picture on the bottom is that of the label 'confidence'. These examples seem very similar, thereby demonstrating the difficulty for humans to decipher complex emotions like those being investigated in our work, and show that individual frames may not be sufficient to describe such emotions. Additionally, these frames reveal difficulties of dealing with 'noisy' data from obstructions by the subjects own hands in front of her face, which may be addressed if temporal features are taken into account.

than 3,000 different YouTube high-definition (HD) videos of people facing and speaking to a camera. The videos are split into training, validation and test sets with a 3:1:1 ratio. People in videos show different gender, age, nationality, and ethnicity, and are labeled with personality traits including: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. Thus each clip has ground truth labels for these five traits represented with a value within the range [0, 1].

3.2. Student Dataset (GRIT Data Collection)

The student dataset for this work was developed using camera recordings of students taking computerized mathematics tests. The camera used for the recordings was positioned to capture the whole face of the subject throughout the entire duration of the test. The recorded video was then segmented into shorter clips for each question, with an average length of approximately 30 seconds. For each clip, a label was assigned for Mastery, emotionAfter, emotionLevel and effort. The labels were assigned by the subject after the completion of each question. A total of 201 clips were collected, which were split into a 70% train and 30% test split.

3.3. Kinetics Dataset

The Kinetics dataset is currently the largest action recognition dataset that includes more than 300,000 trimmed videos covering 400 categories. The number of training, validation, and testing sets are about 240,000, 20,000, and 40,000, respectively.

4. Approach

4.1. Network Architectures

4.1.1 Open Face

OpenFace [2] is an open source tool for face recognition with deep neural networks. For a given image, it returns facial landmark coordinates, head pose and eye gaze coordinates, and facial Action Unit(AU) estimation.

4.1.2 3D-ResNeXt-101

For this study, we selected a 3D-ResNeXt-101 model, since it has been shown to give the best performance on action recognition applications by Kensho Hara *et al.* [7]. This model is an extended version of a residual network (ResNets) [7].

ResNet architectures provide shortcut connections that allow a signal to skip one or more layers and simply be added to the outputs of the next layer in the network (performing identity mappings). These connections enable the training of very deep networks, by solving the vanishing gradient problem.

ResNeXt introduces a new dimension to ResNets called cardinality that has been shown to be more effective (when increased) compared to the existing dimensions of depth and width. The ResNeXt block reshapes the original ResNet block by introducing grouped convolutions, which divide the feature maps into small groups. Cardinality refers to the number of convolutional layer paths or groups in the bottleneck block. We use a cardinality value of 32, suggesting grouped convolutions with 32 groups. Additionally, 3D kernels are used to capture spatio-temporal features across 16 frame segments.

4.2. Fine Tuning and Feature Extraction

4.2.1 Fine Tuning 3D-ResNeXt-101

For the 3D-ResNeXt-101 network, a model pre-trained on the Kinetics dataset using PyTorch and input preprocessing tools developed by Kensho Hara *et al.* [7] were utilized for fine tuning on the Chalearn dataset. Only the conv_5x and fc layers of the model were fine-tuned, with the prior layers kept frozen. A batch size of 10 was used with a learning rate of 0.001, assigned with a weight decay of 1e-5. The

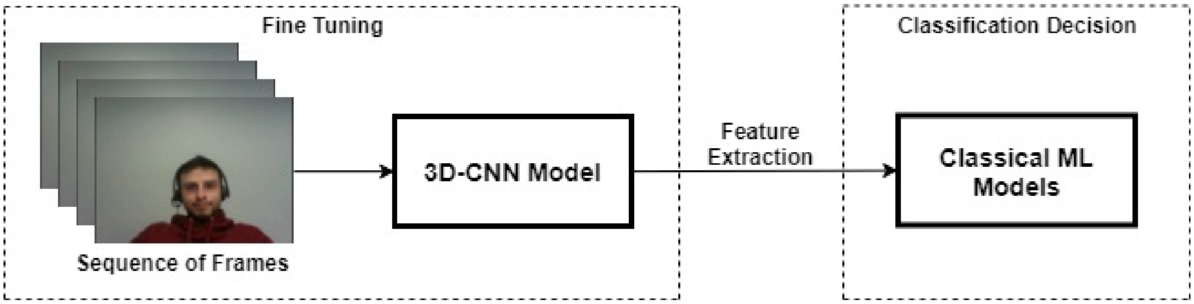

Figure 2. A high level system diagram of the use of 3D-CNNs for emotion estimation

Chalearn dataset was chosen for fine tuning since it is similar to the student dataset in that each video focuses on the face of a subject throughout the entire clip.

After fine tuning, a 2048-dimension feature vector is extracted from the output of the conv_5 layer for each 16 frame segments of an input video clip. The features for all the segments were concatenated, and the mean of all the segments was taken to produce a single 2048 feature vector for an entire video clip.

4.2.2 OpenFace

OpenFace was used to get extract the following for each frame: 1) Action Unit intensities, 2) Action Unit Presence and 3) Eye gaze vectors. The mean, standard deviation, minimum and maximum of these features across all the frames of a clip were calculated and concatenated to give a single 164-dimensional vector for the entire video clip.

5. Experimental Results and Discussion

For our work, where applicable, we converted problems into classification problems by thresholding the target values at 0.5 for regression values, or by considering one class at a time for multi-class classification datasets. We used the Area under the ROC curve (AUC) to estimate the classification accuracy.

For each problem, multiple machine learning models were used for binary classification of emotion and effort labels of the student dataset. These include 1) Random Forests with 400 trees with a minimum number of samples required to be at a leaf node set to 50. 2) 2-Layer Multi Layer Perpectrons (MLPs) with 100 neurons in each hidden layer trained using stochastic gradient descent. 3) Linear Support Vector Machines (SVM).

5.1. Analysis of classification using OpenFace features

Feature vectors from the OpenFace model were extracted for the student dataset and divided into a 70% Training and 30% Test set. The training set was used to train multiple classifiers with 2-Layer MLPs giving the best results for effort and emotion labels, shown in Table 1 and 2 respectively.

	Effort
OpenFace	0.67
3D-CNN	0.70

Table 1. AUC scores for the effort label SOF (Solved on First Try) on test data from the student dataset. The 3D-CNN performed with a 3% improvement over the OpenFace model.

	Confidence	Interest	EmotionLevel=3
OpenFace	0.60	0.62	0.63
3D-CNN	0.62	0.63	0.83

Table 2. AUC scores for Emotion labels on test data from the student dataset. The use of 3D-CNNs improves the performance of binary classifiers for emotion labels by 2-3 % for 'condfidence' and 'Interest' and more than 30% for EmotionLevel.

5.2. Analysis of classification using 3D-CNN features

Feature vectors from the 3D-ResNeXt-101 model were extracted for the Chalearn dataset to see its viability for the model's use on the student dataset.A 2-layer MLP was trained on the 6000 video training set features and tested on the 2000 video test set features. This achieved an AUC score of 0.79 (also see Figure 3), which would be ranked 7th in the ChaLearn Looking at People 2016 First Impressions challenge [6].

To reduce the time required for training, the dimentionality of the data was reduced using Principal Component Analysis (PCA). The top 20 components that described 95% of the data were selected for training.

We then extracted features for the student dataset and divided the resulting dataset into a 70% Training and 30% Test set. The training set was used to train multiple classifiers with 2-Layer MLPs giving the best results for effort and emotion labels, shown in Table 1 and 2 respectively.

For the label EmotionLevel, the most frequently occur-

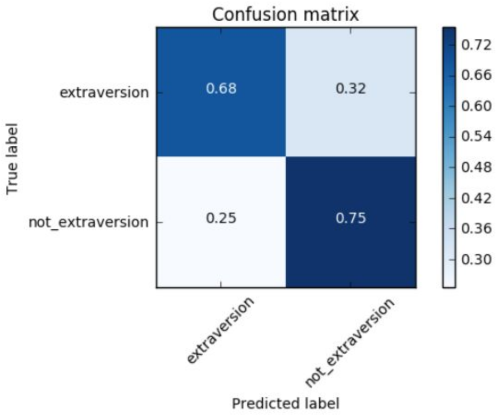

Figure 3. Confusion Matrix for prediction on the label 'extraversion' from Chalearn test dataset. Features were extracted using the 3D-ResNext architecture

ring level in the dataset at 50% was 3 (range 1-5). Therefore a binary classifier was used for this level. The 3D-CNN improved the classification performance by 31%.

Binary classification for the emotion labels 'Confidence' and 'Interest' gave results with improvements of 2-3% when using 3D-CNNs over OpenFace.

6. Conclusion

The results from our work demonstrate that 3D-CNNs can model temporal video data to predict or estimate emotions using only human facial data. Our experiments show that using this approach improves classification performance over the practice of aggregating per-frame predictions from open source tools. Being the first study on the student dataset, our results can also serve as a benchmark for future models and studies.

One strength in our approach is in the simplicity of its implementation, compared to more complex 2 stream methods used for action recognition.

Further work may include fine-tuning the 3D CNN for a larger number of epochs, as well as experimenting with multiple CNN architectures.

Additionally, as part of the student data collection process, high definition video was also collected using a Gopro camera. This could be used for to extract the eye regions, or specifically, the pupil regions of the image, and apply cognitive load estimation methods as seen in [5] to estimate the emotional stress being faced by the subjects in real time.

The student dataset itself can be improved by shortening the duration of each clip. While the current dataset clip duration averages at 30 seconds, this in fact varies from 5 seconds to 2 minutes.

References

[1] M. S. W. S. A. Koakowska, A. Landowska and M. R. Wrobel. Human-Computer Systems Interaction: Backgrounds and Applications 3. *Cham: Springer International Publishing*, pages 51–62, 2014.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. OpenFace: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. *arXiv preprint,arXiv:1705.07750*, 2017.

[4] R. F. L. T. D. Tran, L. Bourdev and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the International Conference on Computer Vision (ICCV)*, page 44894497, 2015.

[5] L. F. et al. Cognitive Load Estimation in the Wild. *10.1145/3173574.3174226*, 1-9, 2018.

[6] P.-L. V. et al. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. *Computer Vision ECCV 2016 Workshops*, 9915, 2016.

[7] K. Hara, H. Kataoka, and Y. Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *arXiv preprint*, arXiv:1711.09577, 2017.

[8] A. A. P. V. Ruiz-Garcia A., Elshaw M. Deep Learning for Emotion Recognition in Faces. *Villa A., Masulli P., Pons Rivero A. (eds) Artificial Neural Networks and Machine Learning*, 9887, 2016.