Emotion Detection in Video Sequences

Shuo Yao and Ishika Prasad
Department of $Computer Science$
Rochester Institute of Technology
Rochester, NY 14623
{sy9663;ip1262}@rit.edu

Abstract

Enabling machine systems to detect various facial emotions and further classify different facial expressions in the video sequences gathers the researcher's attention. This paper presented two approaches to feature extraction. The first approach compares the two adjacent frames and dense optical flow computed using Gunner Farneback's algorithm. The second approach uses OpenFace 2.2.0 for face feature embedding. OpenFace is an implementation of face recognition with deep neural networks. Further, both approaches use Principle Component Analysis (PCA) for facial feature extraction, K Nearest Neighbor (KNN) algorithm for symbolized features and final step uses Hidden Markov Model (HMM) for classification of six different expressions. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) has used for the training and testing of our approach. The model is evaluated using confusion matrix, precision, and recall.

1 Introduction

Emotion is one of the fundamental parts of human to human communication. To build an intelligent and robust model of human to machine interaction, the model should replicate the human to human interaction based on facial expressions. The real-time emotion detection with high accuracy is an important and challenging problem. There are six universal facial expressions: happy, sad, surprised, disgust, fear, and anger. The facial emotions are expressed through the activation of specific sets of facial muscles. The main goal of the paper is the build a system that detects and classifies the universal emotions in the video sequence.

Achieving this model can be used in various fields, namely health care, stores, automotive industry. In health care, it can be used to determine that if the patient needs a consultant with the physician urgently. In other words, physicians can determine which patient should be treated on a priority basis. Stores use surveillance cameras to identify any criminal records. This recorded video can be used to evaluate customer interest and satisfaction. In the automotive industry, it can be used in smart cars where if the driver feels drowsy then the driver can be alerted.

MH et al. [2017] proposed patterns of oriented motion flow (POMF) for feature extraction from the facial video. The information about facial motion in different directions computed by POMF. Sharma and Gupta [2014] proposed Fast PCA for further reducing the emotions from the continuous sequence of images. Mishra and Ratnaparkhi [2018] proposed the emotion detection in games using Hidden Markov Model. Aswin K.M et al. [2016] proposed the method for emotion detection from speech and video using Mel-Frequency Cepstral Coefficients (MFCC) and Support Vector Machines (SVM).

In this paper, the work incorporates different strategies to build a model that recognizes and classifies the facial emotion in the video sequence. Two approaches are used for face tracking and feature extraction. One type of approach includes analyzing the facial emotions through the sequence of

Figure 1: Example of happy expression Rav

Figure 2: Example of sad expression Rav

images used different strategies to meet various requirements. The pre-processing part includes face tracking and extraction to get the frames from the video. If there are n frames the compare the n frames with adjacent two frames to create the optical Flow. The feature extraction is done using PCA (Principle Component Analysis) algorithm as this approach is effective while maintaining accuracy and speed. Another type of approach used OpenFace for tracking of face and feature extraction. This method is efficient with speed and accuracy for the rapid motion of face videos. Above both approaches uses the same method for symbolized features, the approach used is KNN (K-Nearest Neighbor) and K-means algorithm but KNN gives better results as compared to the K-means algorithm. Finally, for detecting and classifying the emotions, we tried different types of HMM (Hidden Markov Model), Gaussian emissions the best result compared to others.

1.1 Data

The data we have used for the model is RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) which contains 2,276 files (total size: 10 GB). The database contains 12 male and 12 female actors. There are two lexically-matched statement vocalized by these 24 professional actors in a neutral North American accent. Video includes surprise, fearful, disgust, sad, angry, and happy expressions. These videos are labeled data and every emotion samples are balanced. Only emotions are considered for training the model. Each of the 2276 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4) Rav.

Rav Filename identifiers are given below:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong).
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Examples of six different emotions of the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database are displayed in Figures 1, 2, 3, 4, 5, and 6.

Data downloaded from the link provided below:

`https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio/kernels`

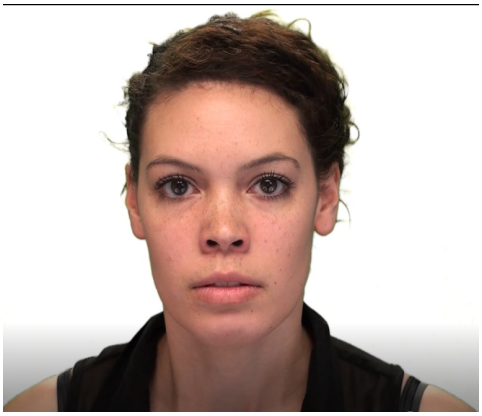

Figure 3: Example of angry expression Rav

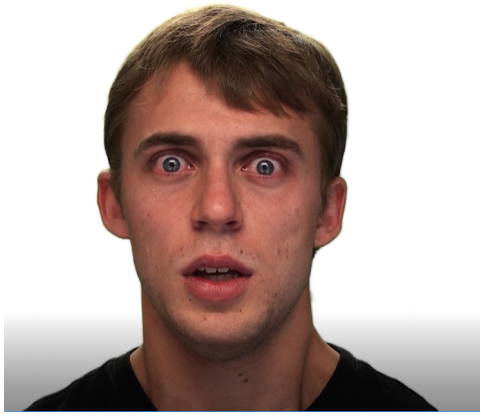

Figure 4: Example of fearful expression Rav

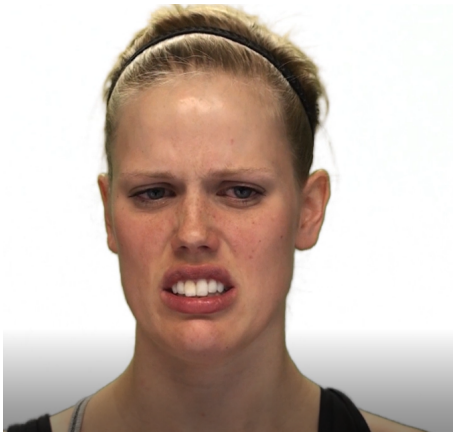

Figure 5: Example of disgust expression Rav

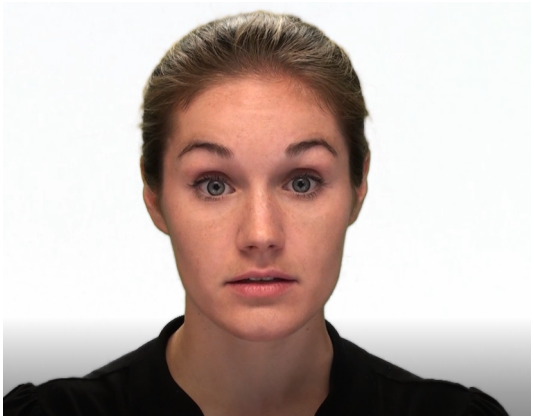

Figure 6: Example of surprised expression Rav

2 Past related work

MH et al. [2017] used the depth database of facial expression. This database is built using a Zcam depth camera. It was developed based on both depth camera-based and RGB image sequences. The approach used to build a model was POMF (Patterns of Oriented Motion Flow) descriptor for feature extraction, K-means clustering technique and HMM (Hidden Markov model) which classify the emotion.

Sharma and Gupta [2014] used the Cohn-Kanade dataset which supports all the six basic expressions namely disgust, anger, sad, surprised, happy and fear. The approach used to create a system was a fast PCA (Principle Component Analysis) algorithm for the feature extraction and SVM (Support Vector Machine) classifier to classify the image sequence and recognize the emotions.

Mishra and Ratnaparkhi [2018] did not mention the dataset used. The approach used to build a model was a PCA (Principle Component Analysis) technique which is used for feature extraction (targeting global details of a face and color reduction) and used HMM (Hidden Markov Model) to classify the expression.

Aswin K.M et al. [2016] used the Cohn-Kanade dataset. The approach used to create a system was a Gabor filter banks for feature extraction and apply SVM to classify the emotions.

Bouzakraoui et al. [2016] used the data collected from the video camera. The approach used to build a model was an LBP (Local Binary Pattern) algorithm which is used to extract the features of facial emotions and classify the expressions using SVM.

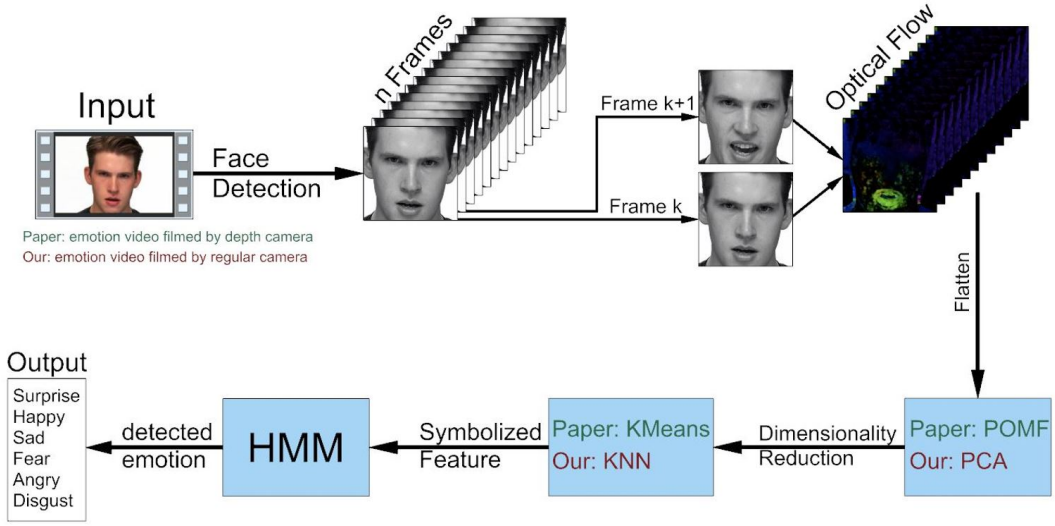

Figure 7: Steps of the proposed first approach.

We have used the RAVDESS dataset which is different from the referenced papers and two different approaches that are not similar to the above-referenced paper. In the first approach, the feature extracted from the video sequence using PCA and then KNN and HMM used to identify the emotion. In the second approach, the feature extraction using OpenFace then used PCA, KNN, and HMM to detect and classify the facial expression.

3 Proposed approach

We proposed two approaches to detect and classify the facial expression in video sequences. Our first approach is inspired by MH et al. [2017] and the second approach is feature extraction variation in the first approach. The rudimentary concept for these two approaches is to take the input video which will go through a feature embedding module and get converted to a sequence of features. Finally, these sequences of features then pass through a classification module for emotion recognition.

3.1 First approach

The steps of the first approach is shown in Figure 7. The model will take a video as input, this video will be processed with face detection and normalization module and converted to a list of frames. For each fame k and frame k+1, dense optical flow will be computed using Gunner Farneback's algorithm G [2003] to embedding the optical motion at each pixel from frame k to k+1. This optical flow sequence contains a large amount of information with an extremely high feature dimension. Principle Component Analysis (PCA) will be used for feature extraction on the optical flow sequence. Then this dimension reduced optical flow sequence will be symbolized using the K-nearest neighbor (KNN) algorithm and input to the classification module for emotion recognition. The classification module contains six Hidden Markov Models (HMMs) for six emotion classes. The input sequence will be tested with all six HMMs and will be assigned to the one with the highest probability value.

3.2 Second approach

For the success of emotion detection on video sequences, one of the most critical part is face feature embedding (especially motion). In our second approach, we decided to use OpenFace 2.2.0 for face feature embedding. OpenFace is an implementation of face recognition with deep neural networks based on F et al. [2015]. Figure 8 shows an example of facial behavior analysis in a video clip from Friends (TV Series) on OpenFace toolkit. The embedded feature will go through Principle Component Analysis (PCA) for dimension reduction and then follow the same process as we described on the first approach. The steps of the second approach is shown in Figure 9.

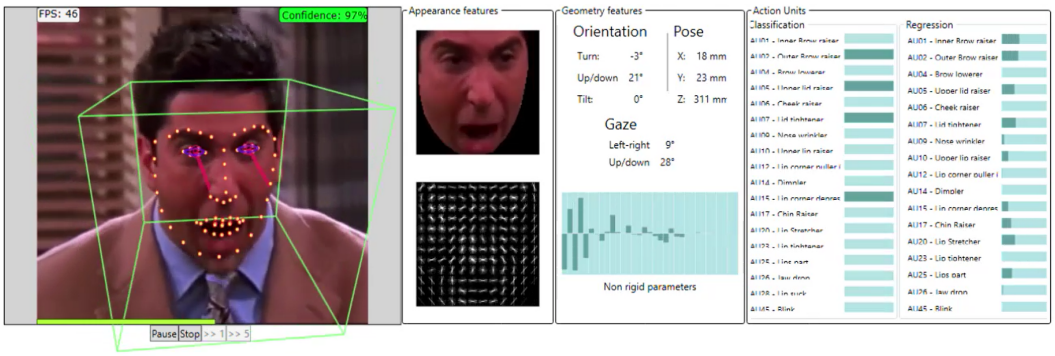

Figure 8: Example of facial behavior analysis on OpenFace toolkit

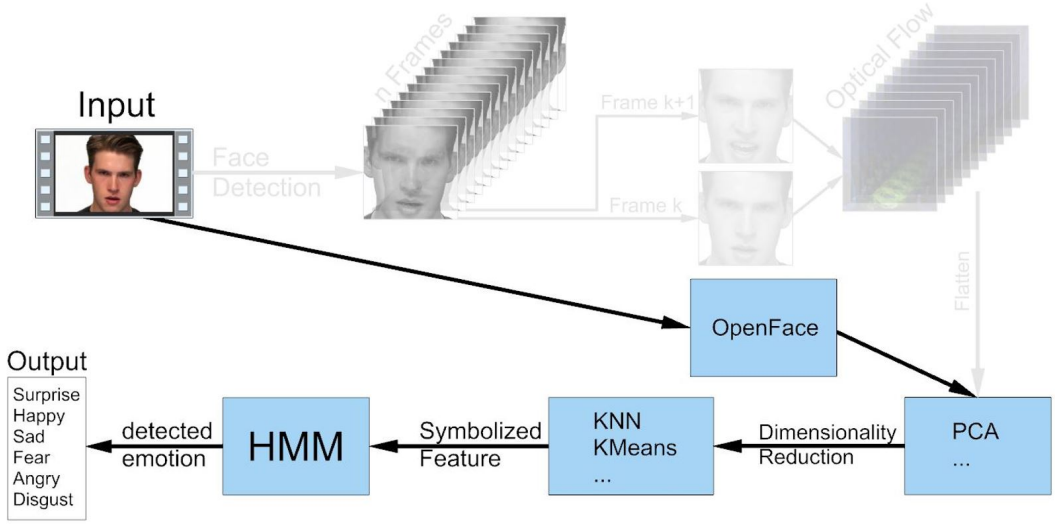

Figure 9: Steps of the proposed second approach.

4 Experiments and Results

In the following section, we will present the data, experiments, and results for our approaches.

4.1 Data

In this paper, we performed the training and testing of the model using RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database. Video includes surprise, fearful, disgust, sad, angry, and happy expressions. These videos are labeled data and every emotion samples are balanced. Only emotions are considered for training the model. We have considered 2,276 video files for our implementation. A Unique filename is provided for each of the video files. The filename contains a 7-part numerical identifier (e.g., 03-01-04-02-01-02-15.mp4). These filename identifiers have 7 different meanings to the labeled unique filename which is described in section 1.1. Further, we divided the video files for training and testing datasets. 1818 video files were randomly selected as training data and the rest 458 video files were selected for testing. To be mentioned, the number of each emotion sample is balanced in both training and testing data.

4.2 Experiments and Evaluation methods

Even though the training process seems to be straight-forward, lots of experiments have been done on parameter tuning and model selection. For the symbolization module, k-nearest neighbors (KNN) and k-means are both good candidates. We have conducted a set of tests to evaluate their perfor-

mance and turns out KNN gives us the best result. We tried a different type of HMMs and the one with Gaussian emissions gives us the best result. Their test results will be presented in the next section. Confusion matrix, precision, and recall of the test result will be used for evaluating the performance.

	Surprise	Happy	Sad	Fear	Angry	Disgust	Precision
Surprise	17	2	8	6	0	2	0.486
Happy	7	74	4	3	4	3	0.779
Sad	1	0	7	9	7	1	0.28
Fear	37	1	37	56	51	8	0.295
Angry	3	0	0	0	2	2	0.286
Disgust	11	0	21	2	12	60	0.566
Recall	0.234	0.961	0.909	0.737	0.263	0.789	

Table 1: Test result for the proposed first approach

	Surprise	Happy	Sad	Fear	Angry	Disgust	Precision
Surprise	66	0	0	0	0	0	1.0
Happy	0	75	0	0	0	0	1.0
Sad	7	1	72	1	3	0	0.857
Fear	1	1	3	75	2	0	0.915
Angry	1	0	0	0	71	0	0.986
Disgust	1	0	2	0	0	76	0.962
Recall	0.868	0.974	0.935	0.987	0.934	1.0	

Table 2: Test result for the proposed second approach

4.3 Results and discussions

From Table 1 we can see the result of our first approach. Even though a lot of experiments have been done for parameter tuning, the model's performance is still limited (accuracy 0.472). One possible reason is that the input video has a distinct difference between a human face and background. When people are talking, they will move their heads, which will cause strong optical flow responses around their faces. These responses are noises and need to be removed. A solution is to implement an edge detection model to detect the silhouette of the human face and remove the optical flow around it. From Table 2 we can see the result of our second approach. There is a huge improvement in the model's performance (accuracy 0.950). This result shows that OpenFace 2.2.0 can be a great choice for feature embedding in video emotion detection projects. This module can also work as a benchmark for us to refine our optical flow feature embedding module.

5 Conclusion and future work

To conclude, we have made two approaches for emotion detection on video sequences. The first one uses optical flow and the second one uses OpenFace 2.2.0 toolkit for feature embedding. KNN was selected for symbolization and HMM was selected for classification. For the first approach, the performance of optical flow is limited and a noise reduction module, as well as a better feature extraction module, needs to be developed. For the second approach, the next thing that should do is to improve the model's reaction speed. Currently, it takes almost 30 seconds to process a 5 seconds long video sequence, more work should be done to improve the model's speed to better fit a real-life scenario. Furthermore, other than HMM, LSTM could also be a good selection for the classification module. Lastly, this data set also includes audio data, which gives the possibility to develop a model that can use both video and audio for a better performance.

References

https://zenodo.org/record/1188976.xchgjkvkil9. URL `https://zenodo.org/record/1188976#.XcHgjkVKiL9`.

Aswin K.M, K. Vasudev, K. Shanty, and Sreekutty I.K. Hers:human emotion recognition system. *IEEE Access*, pages 176–179, 2016.

M. S. Bouzakraoui, A. Sadiq, and N. Enneya. Towards a framework for customer emotion detection. *IEEE Access*, pages 1–6, 2016.

Schroff F, Kalenichenko D, and Philbin J. Facenet. A unified embedding for face recognition and clustering. *IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Farnebäck G. Two-frame motion estimation based on polynomial expansion. *Scandinavian conference on Image analysis*, pages 363–370, 2003.

Kabir MH, Salekin MS, Uddin MZ, and Abdullah-Al-Wadud M. Facial expression recognition from depth video with patterns of oriented motion flow. *IEEE Access*, 5:8880–8889, 2017.

P. Mishra and S. Ratnaparkhi. Hmm based emotion detection in games. *IEEE Access*, pages 1–4, 2018.

G. Sharma and S. Gupta. Emotion detection in sequence of images using advanced pca with svm. *IEEE Access*, pages 686–690, 2014.