Two Stage Emotion Recognition using Frame-level and Video-level Features

Carla Viegas[1],[2]
[1] Language Technologies Institute, Computer Science Department, Carnegie Mellon University, Pittsburgh, USA
[2] TalkMeUp Inc., Pittsburgh, USA

Abstract— **This paper compares a seven class classifier with a two stage classification for categorical emotion recognition. We use the Multimodal Emotion Recognition (MER) Dataset of the FG2020 competition which apart from video contains skeleton data collected with a Microsoft Kinect. We compare the performance of different unimodal features as well as various combinations of multimodal features. We also compare frame-level features with video-level features. We achieved 50% accuracy using multimodal video-level features and two stage classification on one hand, 49% accuracy is achieved with the seven class classifier on the other hand.**

I. INTRODUCTION

Emotion recognition has a variety of interesting applications such as human computer interaction, advertising [1] and communication. As we can express emotions through the way we speak (prosody), facial expressions, gestures, and the speech content the use of several modalities for classification has been in major focus in the last decade [2].

In the last years the main focus for multimodal emotion recognition has been the use of deep neural networks. CNNs have been applied to generate facial descriptors [3] as well as sound representations [4]. LSTMs have been used for sentiment analysis on text [5]. However, in order to use deep neural networks sufficient data is necessary.

Given the size of the MER dataset (45 recordings per emotion in the training set) and the challenge restriction of using pretrained models, we compared the performance of traditional classifiers. We performed an extensive parameter search for different classifiers such as Random Forest (RF), Support Vector Machines (SVM) and K-Nearest-Neighbor (KNN). Then, we compared the performance of a seven class classification with a two stage classification. In the first stage we train a classifier to distinguish between positive, neutral and negative emotions. In the second stage we train two classifiers: one for positive emotions (happy, surprise) and another for negative emotions (disgust, fear, anger, sad). We also compare the performance on frame and video-level features using unimodal and multimodal features.

In the end, we were able to achieve 50% accuracy using the two stage classifier on video level features. The best results were obtained by concatenating Facial Action Units (FAUs), Mel Frequency Cepstral Coefficients (MFCC) and statistical measures from the motion data.

II. RELATED WORK

In recent years several approaches using deep neural networks (DNNs) were presented for emotion recognition. However, using DNNs on small datasets is challenging.

Fine tuning of previously trained neural networks on other face datasets has been commonly appield in order to use smaller emotion datasets. [6] used video features by finetuning the VGG16-Face model [7] with the FER2013 face emotion database. Also useful when working with small datasets is cross-modal transfer. [8] assumed that the emotional content of speech correlates with the facial expression of the speaker. They use a teacher network to learn emotion recognition from speech by using facial emotion recognition as the “teacher”.

As for the challenge it is not permitted to use pretrained models. Therefore we focused on using hand-crafted features. In previous EmotiW challenges Gabor and Local Binary Patterns (LBP) have been used as video descriptors [9]. Also Facial Action Units (FAUs) have shown to be suitable in emotion recognition and stress detection [10]. For emotion recognition from audio MFCC [11] have been used as well as pitch for prosody detection [12].

As emotions evolve over time, previous works have approached the problem either by classifying emotions frame by frame (frame-level) or by classifying frame sequences (video-level). In the 2015 Emotion Recognition in the Wild (EmotiW) Challenge [13] both approaches are combined by using CNNs on static images and RNNs on frame sequences.

III. METHODOLOGY

A. Data

The dataset consists of video and Microsoft Kinect skeleton recordings of 16 professional actors (8 male and 8 female). Each actor performs all seven emotions: Neutral state (Ne), Sadness (Sa), Surprise (Su), Fear (Fe), Anger (An), Disgust (Di), Happiness (Ha). Each emotion is performed five times during a recording. In each repetition the actors say in Polish “Kady z nas odczuwa emocje na swj sposb” (en. “Each of us perceives emotions in a different manner”). The videos are recorded with 1920x1080 resolution and 29.97 fps. Audio is provided through the videos. The Kinect data provides position and rotation of 25 joints per frame with a sampling rate of 25 fps. The training set contains recordings of 5 female and 4 male actors, the validation set contains 2

This work was supported by TalkMeUp Inc.

978-1-7281-3079-8/20/$31.00 ©2020 IEEE
DOI 10.1109/FG47880.2020.00143

female and 2 male actors respectively. The actors are strictly separated in training and validation set. For each emotion 45 performances are available in the training set and 20 performances are available in the validation set.

B. Features

We extract features on two different levels: video and frame. As the use of pretrained networks was not permitted for the challenge, we extracted Gabor, LBP and FAUs from the video recordings and MFCC features from audio respectively. We also detected the duration of voice activation and the silence length between utterances. From the 25 joints recorded by the Kinect we only used the coordinates of 5. In the following we will explain the used features in more detail.

a) Video-level features: To obtain video-level features, we used statistical information of the features over all frames.

Video features: We extracted Gabor and LBP features per keyframe and calculated the mean for each feature during the video. The FAUs were extracted using OpenFace [14]. The mean, standard deviation, minimum and maximum values were computed for each AU and used as feature.

Audio features: MFCC features were extracted using OpenSMILE [15]. To reduce dimensions we used K-Means clustering (clusters=200) to obtain a 200 x 39 feature for each audio recording. To obtain a video-level feature the mean of the K-Means representation was used. Given that in each emotion performance the same sentence was said, we used a voice activity detector[1] to extract the duration of the utterances and the silence in between utterances. As the articulations detected varied in each recording, we created a vector of 1 x 30 per video with zero-padding in order to be used in our model training.

Motion features: After viewing some video recordings, we decided to focus on joints that are most used in gestures during speech. We chose to use the x and y position of the left hand tip, the right hand tip, and the head and only the y positions of the left and right shoulder as there is small movement in x direction. From each of the listed joints the mean and standard deviation of their positions were computed. We also computed the acceleration per joint by calculating the distance of each joint in two consecutive frames. We added the mean, standard deviation, minimum and maximum values of the acceleration to the feature.

Multimodal features: To create multimodal features we performed early fusion by concatenating features from different modalities. We tested different combinations of features.

b) Frame-level features: We used the gollowing selection of the features mentioned above without computing statistical measurements: Gabor, LBP, FAUs and MFCC features. From the skeleton data we computed distances between the selected joints and included the accelerations of each joint per frame.

C. Classification

We defined two separate classification problems: a) a 7 class problem (one for each emotion category) and a b) two stage classification. In the two stage classification we first detect whether the emotion is positive, neutral or negative (3 class classification). In the second stage we use two different classifiers to detect one of the 7 categories.

a) 7 class classification: We perform an extensive parameter search for different classifiers: AdaBoost , Naive Bayes (NB), Nearest Cluster (NC), KNN, SVM, Random Forest (RF), and Decision Tree (DT). We use 5-fold cross-validation on the training set and evaluate prediction results on the validation set. In the frame-level case we compute the majority vote to obtain one prediction from all frames. The actor recordings are strictly separated between the training set and the validation set.

b) 2 stage classification: **Stage 1**: In the first stage, we classify each video into 3 categories: positive emotions (happy, surprised), neutral, negative (sad, disgust, fear, anger). For this purpose, we train the same classifiers as in the 7 class problem by performing 5-fold cross-validation on the training set and evaluate on the validation set.

Stage 2: In the second stage, we train two different models. One for positive emotions and one for negative emotions. We train each model with the entire training set using 5-fold cross-validation. For evaluation we use the predictions from stage one and forward the features to the predicted stage 2 classifier.

IV. RESULTS

Table I shows the results of the 7 class classification problem on video-level features. For each feature the best classifier and its accuracy as well as F1 score, precision and recall are shown. The highest accuracy results were obtained using multimodal features. The combination of FAUs with the audio features obtained the highest accuracy of 49%, as well as the combination of FAUs, audio, motion and voice activity. Both features obtained best results using SVM.

In Table II we see the results of the first stage classification. The reduction of classes improved the accuracy values as expected. However, the highest accuracy is only 73% using the mutlimodal feature of the combination of FAUs, audio, motion and voice activity.

Table III shows the final prediction results of the 2 stage classification. The best result was obtained by combining FAUs with the motion features. The highest accuracy in this case was 50%.

Frame-level classification results are shown in Table IV. The highest accuracies were achieved by the FAUs and MFCC features, respectively 40% and 39%.

V. DISCUSSION

In this work we compared the results of a 7 class classification problem with a two stage classification. During stage 1 we reduced the seven emotion classes to three (positive, neutral, negative). By doing so we expected to facilitate classification in the second stage.

[1] https://github.com/marsbroshok/VAD-python

TABLE I

RESULTS FROM 7 CLASS CLASSIFICATION USING VIDEO-LEVEL FEATURES.

Feature	Classifier	Accuracy	Precision	Recall	F1-score
Audio	NB	0.35	0.37	0.35	0.34
Motion	KNN	0.31	0.32	0.31	0.27
Voice Activity	KNN	0.27	0.34	0.27	0.26
FAUs	RF	0.35	0.38	0.35	0.35
Gabor	RF	0.20	0.19	0.20	0.17
LBP	Adaboost	0.24	0.26	0.24	0.20
Gabor + LBP + Audio + Motion	SVM	0.42	0.45	0.42	0.42
Gabor + Audio + Motion	SVM	0.39	0.40	0.39	0.38
LBP + Audio + Motion	SVM	0.38	0.40	0.38	0.38
FAUs + Audio + Motion	SVM	0.48	0.49	0.48	0.48
FAUs + Audio + Motion + Voice Activity	SVM	0.49	0.50	0.49	0.49
FAUs + Audio	SVM	0.49	0.50	0.49	0.49
FAUs + Motion	SVM	0.46	0.46	0.46	0.46
FAUs + Voice Activity	SVM	0.44	0.42	0.44	0.42

TABLE II

RESULTS FROM STAGE 1 CLASSIFICATION (3 CLASS PROBLEM) USING VIDEO-LEVEL FEATURES.

Feature	Classifier	Accuracy	Precision	Recall	F1-score
Audio	SVM	0.61	0.59	0.61	0.58
Motion	KNN	0.59	0.64	0.59	0.58
Voice Activity	Adaboost	0.57	0.33	0.57	0.42
FAUs	RF	0.64	0.71	0.64	0.64
Gabor	Adaboost	0.62	0.57	0.62	0.51
LBP	RF	0.62	0.61	0.62	0.61
Gabor + LBP + Audio + Motion	SVM	0.66	0.65	0.66	0.65
Gabor + Audio + Motion	RF	0.62	0.54	0.62	0.54
LBP + Audio + Motion	SVM	0.65	0.64	0.65	0.64
FAUs + Audio + Motion	SVM	0.70	0.70	0.70	0.70
FAUs + Audio + Motion + Voice Activity	SVM	0.71	0.70	0.71	0.70
FAUs + Audio	RF	0.73	0.76	0.73	0.71
FAUs + Motion	RF	0.67	0.57	0.67	0.62
FAUs + Voice Activity	SVM	0.66	0.68	0.66	0.66

TABLE III

FINAL RESULTS FROM STAGE 2 CLASSIFICATION AFTER COMBINING THE PREDICTIONS ON THE VALIDATION SET FROM THE CLASSIFIER FOR NEGATIVE EMOTIONS AND THE CLASSIFIER FOR POSITIVE EMOTIONS. FEATURES USED ON VIDEO-LEVEL.

Feature	Classifier	Accuracy	Precision	Recall	F1-score
Audio	RF	0.42	0.53	0.42	0.43
Motion	DT	0.32	0.41	0.32	0.34
Voice Activity	DT	0.35	0.43	0.35	0.34
FAUs	KNN	0.40	0.49	0.40	0.40
Gabor	DT	0.31	0.42	0.31	0.32
LBP	RF	0.30	0.43	0.30	0.30
Gabor + LBP + Audio + Motion	RF	0.44	0.54	0.44	0.45
Gabor + Audio + Motion	SVM	0.41	0.53	0.41	0.42
LBP + Audio + Motion	RF	0.41	0.52	0.41	0.42
FAUs + Audio + Motion	RF	0.50	0.56	0.50	0.49
FAUs + Audio + Motion + Voice Activity	RF	0.49	0.55	0.49	0.47
FAUs + Audio	RF	0.49	0.52	0.49	0.47
FAUs + Motion	SVM	0.50	0.55	0.50	0.49
FAUs + Voice Activity	NC	0.44	0.47	0.44	0.42

TABLE IV

RESULTS FROM 7 CLASS CLASSIFICATION USING FRAME-LEVEL FEATURES.

Feature	Classifier	Accuracy	Precision	Recall	F1-score
FAUs	RF	0.40	0.39	0.40	0.39
MFCC	Adaboost	0.39	0.36	0.39	0.32
Motion	KNN	0.15	0.05	0.15	0.05
Gabor	KNN	0.22	0.22	0.22	0.19
LBP	RF	0.23	0.24	0.23	0.21

However, the final results of the two stage classification were not very different from the seven class classification. Both methods achieved similar results of 49% and 50% accuracy using video-level features. In both cases multimodal features that combine FAUs with audio and motion features achieved the best results. In order to obtain better classification results from the two stage classification the accuracy of the stage 1 classifier needs to be higher.

In the 7 class classification the results of frame-level and video-level feature did not differ much. For future work, pretrained models or transfer learning techniques should help to increase the accuracy.

REFERENCES

[1] N. Hamelin, O. El Moujahid, and P. Thaichon, “Emotion and advertising effectiveness: A novel facial expression analysis approach,” *Journal of Retailing and Consumer Services*, vol. 36, pp. 103–111, 2017.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[3] V. Vielzeuf, S. Pateux, and F. Jurie, “Temporal multimodal fusion for video emotion classification in the wild,” in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 569–576, 2017.

[4] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1d & 2d cnn lstm networks,” *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[5] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, “Deep learning for aspect-based sentiment analysis: a comparative review,” *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019.

[6] 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*, pp. 445–450, 2016.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” 2015.

[8] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, “Emotion recognition in speech using cross-modal transfer in the wild,” in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 292–301, 2018.

[9] M. Kächele, M. Schels, S. Meudt, G. Palm, and F. Schwenker, “Revisiting the emotiw challenge: how wild is it really?,” *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 151–162, 2016.

[10] C. Viegas, S.-H. Lau, R. Maxion, and A. Hauptmann, “Towards independent stress detection: A dependent model using facial action units,” in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, IEEE, 2018.

[11] P. P. Dahake, K. Shaw, and P. Malathi, “Speaker dependent speech emotion recognition using mfcc and support vector machine,” in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pp. 1080–1084, IEEE, 2016.

[12] Q. Zhang, N. An, K. Wang, F. Ren, and L. Li, “Speech emotion recognition using combination of features,” in *2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*, pp. 523–528, IEEE, 2013.

[13] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, “Recurrent neural networks for emotion recognition in video,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 467–474, 2015.

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: an open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, 2016.

[15] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.