Summarization of Videos from Online Events Based on Multimodal Emotion Recognition

Abstract:

In this paper, we propose a novel video summarization technique for automatic affect analysis of participants of an online event. At first, face verification neural network is used to cluster facial regions that correspond to each participant. Next, emotional features are extracted from each face by using EfficientNet model obtained in the previous paper of the author. The features of several consecutive frames are combined into a single descriptor that is used to classify emotions. In addition, audio features are extracted using wav2vec, and an ensemble of audio and video classifiers predicts emotions for each face. Finally, dependence of these emotions on time is visualized in special color charts. In the experimental study with the AFEW dataset it was demonstrated that the proposed approach makes it possible to obtain the best-known validation accuracy 67.88%. The models were optimized using OpenVINO and gained reasonable performance even if Nvidia GPUs are unavailable. The models and source code are publicly available at https://github.com/amirabdrahimov/multimodal-emotion-recognition.

Published in: 2022 International Russian Automation Conference (RusAutoCon)

Date of Conference: 04-10 September 2022

Date Added to IEEE *Xplore*: 27 September 2022

ISBN Information:

DOI: 10.1109/RusAutoCon54946.2022.9896386

Publisher: IEEE

Conference Location: Sochi, Russian Federation

Funding Agency:

SECTION I.

Introduction

With the onset of the COVID-19 pandemic, video conference tools, e.g., Zoom, MS Teams, etc., play an incredible role in bringing people together. All over the world, many classes, both at universities and schools, have been moved to an online format due to social distancing measures aimed at preventing the spread of the virus. Although these technologies allow to get rid of direct presence and continue classes, in its current

form they have several limitations. Many teachers and students complain about the lack of emotional interaction between each other in such conditions [1]. It was also shown that emotion is an essential part of studying motivation during classroom interactions [2–4]. During the offline lecture, a speaker can get feedback from listeners by observing their reaction, and even though modern video conferencing tools allow to see the faces of the participants, they cannot fully convey the mood of the audience, for example, due to the limited number of faces that can fit on the monitor screen. One of the possible solutions is to equip video conferencing tools with facial emotion recognition instruments [5], but to examine the overall mood of the lesson some brief summaries are required. That is the summarization problem that can be used to find the parts of the lecture that are difficult for most students. Unfortunately, traditional techniques [6, 7] are not suitable for emotions analysis, and the direct use of images can increase the burden on end users while tracking the evolution of emotions in a video. Therefore, for our purposes, it is necessary to tackle summarization from the emotions analysis perspective. The nature of emotions is multimodal itself, although great progress in the face and video analysis has been made [8–10], the facial expression recognition remains a challenging problem due to the expression suffers from the large pose, illumination variance, occlusion, motion blur, etc. Moreover, existing approaches to multimodal emotion recognition use complex ensembles of models that are not suitable for practical use on devices with poor computational capabilities [11, 12].

To move away from the limitations associated with video analysis and to fully handle emotion recognition, we propose an ensemble of audio and video-based classifiers. Our approach allows to fast perform summarization on various devices using the OpenVINO framework [13], regardless of their computing power and GPU availability. Such an expansion of video communication capabilities may help the speaker to feel the mood of the audience better, as well as provide other end users with summary statistics of the video.

This article is organized as follows. Section 2 reviews the recent literature related to emotion recognition on the AFEW dataset, and video summarization. Section 3 contains an overview of the proposed approach (Fig. 1). Experimental results are presented in section 4. Finally, concluding comments and future works are discussed in Section 5.

SECTION II.

Related Works

The video-based emotion recognition techniques are typically compared based on the results on the AFEW (Acted Facial Expressions In The Wild) from the EmotiW (Emotion

Recognition in the Wild) challenges [14]. In this paper, we examine emotion recognition based on two modalities - video and audio. Text information, the position of a person in the frame, physiological data and other signs were not considered.

Fig. 1.

Proposed video summarization approach based on emotion classification

Show All

The current best result on the validation part of the AFEW is equal to 65.5% [15]. In this solution, an ensemble of three models, namely, VGGFace, ResNet18, and IR50, was used for feature extraction from visual part, and AlexNet was used for processing of an audio spectrogram. To combine the video and audio parts, the transformer-attention was chosen. However, this is not the best approach for the video part of AFEW. The DenseNet-161 was used to extract multiple facial features from each frame [16]. One of the best single models for the AFEW is obtained via the noisy student training using body language [17]. Finally, the EfficientNet-B0 from the previous papers of one of the authors [8], [18], let us reach the best-known accuracy of 59.27% for single model and facial modality. This approach is also notable for that such a result was obtained using a lightweight model. Indeed, the size of EfficientNet-B0 is 15 MB, which makes it a great option for applications where the weight and speed of the system are critical, such as video summarization.

Video lectures, video blogs, video messages on social networks and videos in many other areas are becoming the main ways to exchange information. One of the possible uses of classified emotions is the summarization of the general state of the person observed in the video. The goal is to provide the user with a brief overview of the content of the video. Depending on whether video frames are used for demonstration or not, summarization techniques can be divided into two categories: image-based techniques and abstract techniques [19]. In techniques based on images, representative or key frames are usually chosen. Key frames should convey most of the basic information contained in the original video. In [6], [7], neural networks with an attention mechanism are used for this task to pay attention to the most relevant frames and “remember” the most important information from previous frames. The video compression with a space-time video montage is presented in [20], in which the informative parts of the video are selected and combined. Abstract methods do not use

objects directly for visualization, but their state over time, for example, using graphs. In [21], the EmotionCues tool was developed that visualizes the emotional state of students. Such summarization combines the general emotion of the video, the evolution of emotion, as well as the emotions of a particular person. To show the mood of a social network user in PEARL [22], emotions are visualized as a stream, where height is responsible for the proportion of emotion.

SECTION III.

Proposed Approach

The video summarization task can be formulated as follows: given an observed video V, it is necessary to select sequences of frames for a certain person with the emotions of interest. The visual modality of a video is typically represented as a sequence of N frames in which the facial regions are detected, and one of C>1 expressions is predicted for each face. Details of the proposed pipeline are considered in this section.

A. Summarization

When it comes to summarization based on emotion recognition, it is not so obvious which fragments to choose that would convey the details of the video. One of the ways is to select episodes (tracks) of a video where a person demonstrates clearly distinguishable emotions for a long time and make a montage out of it. Utilizing the idea of summarizing video frames, we developed an algorithm that creates short videos (gif images) from a sequence of consecutive frames on which the same emotion is classified. The fragments for such videos were selected as follows: the frames should follow each other, the emotion on the current frame should match the emotion on the last frame, and the emotion should be in the list of interesting states. In our case, sequences of frames longer than 10 were used, as it proved to look great during testing.

Just as in EmotionCues [21], in our work, attention is also drawn to the generalization of the emotional state and tracking of dynamical emotions. One of the difficulties that arises when trying to visualize emotions is the occurrence of overload due to variety of emotions, and many faces on the video. For example, Fig. 2 shows what an attempt to animate emotions during the video looked like.

Fig. 2.

Attempt of annimating emotions.

Show All

As a result, it is difficult to distinguish the necessary information from such graphs, as emotions-lines overlap each other. Eventually, we decided to move from animated visualizations to interactive ones, where the user can choose the emotions and the person of interest in the video. After considering various well-established visualization techniques, it was decided to use a bar chart to compare the number of emotions present in different people, a pie chart for a more detailed study of a particular person's emotions, and a color video track on which emotions are visualized during the video. An interactive application with an html interface using Dash has been developed, which can be run in a browser or in a development environment.

The proposed algorithm (Fig. 1) includes six steps:

1.

Frames are extracted from videos by using OpenCV.

2.

Facial regions are detected in each frame using MTCNN [23].

3.

Faces are fed into the face recognition model and a feature vector is extracted for each face. Distances between vectors of facial features are calculated, the closest vectors are organized into clusters.

4.

Emotions are detected on images of faces and assigned to the person to which they belong.

5.

A gif image is created from a sequence of frames on which the same emotion is classified for one person.

6.

Emotion data is sent to the app and used for visualization.

Although our application can work with videos of any content on which people are present, for videos with participants of a video conference we suggest adding an extra step. Optical character recognition tool Pytesseract can be used to additionally identify

a person by name. Further, the approaches used for emotion recognition are explained in more detail.

B. Video Modality

For extracting deep facial features, it is typical to use the networks pre-trained for face recognition task, such as ResNet50 (VGGFace2), VGG13 (VGGFace), etc., but these models have rather high computational time and complexity. To speed up this step, we used the lightweight model based on EfficientNet-B0 [8] as the main model for emotion recognition. In fact, it has the best result for the video modality of AFEW [14], near-state-of-the-art facial expression recognition accuracy on AffectNet [24], and has great speed and weight. For face recognition task we also used lightweight EfficientNet-B0. This model was pre-trained on the VGGFace2 dataset and later fine-tuned on the training set from AffectNet to classify 8 emotions. This model was used to extract a 1280-dimensional feature vector from the penultimate layer for every video frame. The video pipeline is adapted from the previous paper [8].

To solve the problem of emotion recognition on the video, each facial image is fed into EfficientNet model to extract emotional D-dimensional features (embeddings). Statistical functions (mean, maximum, minimum, and standard deviation) are calculated for extracted features of a single video and concatenated into one descriptor. As a result, a 4D-dimensional vector is obtained for each video, $4 \times 1280 = 5120$. Finally, several conventional classifiers from scikit-learn, such as Random Forest and Linear SVC, are trained on the L2-normed video descriptors to predict the emotional state for each video.

C. Audio Modality

Similar to working with video, the first stage for audio was the preparation of data. The FFmpeg framework was used to extract audio modality from video with the desired sampling rate of 16000 Hz. Every audio was padded to the length of 50000 and further used to train models.

During our research, we considered two different options: QuartzNet [25] or wav2vec 2.0 [26]. Since the latter model was trained using a self-supervised approach, it demonstrates more powerful representation and better accuracy for a small portion of fine-tuning data. Moreover, our main goal was to get the highest possible accuracy, so we decided to experiment with two different wav2vec 2.0 models, namely, base and large. We fine-tuned the audio model with the training set of AFEW dataset and used it to predict emotions for every clip. Since some emotions can be hard to separate, for instance, contempt and disgust, during the training of the audio model we used label smoothing technique, with value equal to 0.7, which allows the model to divide emotion

classes more easily.

D. Audio-visual Ensemble

The audio and video modalities were combined into a simple blending ensemble. Since LinearSVC from scikit-learn does not have the predict_proba method, we first used decision_function method to predict confidence scores for samples and then applied softmax. After that, the output of these models can be considered as “probabilities” of one of the seven emotions, so we were able to use soft voting to combine the results of fine-tuned wav2vec audio model and classification of facial features.

Finally, since one of the goals was to build an application, we needed to prepare models for CPU execution. While the video model was designed to be small and performative from the beginning, wav2vec 2.0 is a large audio model and needed to be optimized to be deployable. For this purpose, we used OpenVINO toolkit.

SECTION IV.

Experimental Study

A. Quantitative Results

In the experimental study, the AFEW dataset [14] was used. The training and validation sets provided by the organizers of the EmotiW challenges contain 773 and 383 video files, respectively, which were collected from movies and TV serials. Every sample belongs to one of the C=7 emotions (Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral).

For video part, all videos were divided into frames using the FFmpeg framework, and the facial regions in each frame were detected using the MTCNN. If it detects multiple faces in a frame, the face with the largest bounding box is selected. If no face is detected in the picture, the entire frame is passed to the network.

The best accuracy for audio modality of the base wav2vec 2.0 model [26] was equal to 37%, which is a decent result. After training the large wav2vec model, the best accuracy was 41.5%, which is, to the best of our knowledge, the state-of-the-art result for audio modality for the AFEW dataset.

The confusion matrices for the video-only, audio-only, and audiovisual ensemble are shown in Fig. 3. As one can notice, the video model does its best with the “Happy” class, predicting more than 90% of the examples correctly. While the audio model does not perform so well, the combination allowed models to complement each other.

An ensemble of wav2vec and Linear SVC for facial video features with weights 0.11 and

0.06, respectively, achieved accuracy of 64.75%. During the experiments, we found out that the result can be improved even more by adding extra models to the ensemble's predictions. An ensemble of fine-tuned wav2vec 2.0 for audio, Linear SVC and Random Forest for video features, and XGBoost for face descriptors extracted using OpenFace tool, with weights [0.25, 0.11, 0.04, 0.01] turned out to be the best one. We even outperformed the best-known accuracy for AFEW validation set, getting 67.885%. Results for different modalities are presented in Table 1.

TABLE I Results On AFEW

Fig. 3.

Confusion matrices of the proposed approach.

Show All

Results of models' optimization with ONNX and OpenVINO are shown in Table 2. The running time to predict emotion of one facial image was measured on the Asus VivoBook S15 X530UF laptop (CPU: Intel(R) Core (TM) i5-8250U CPU @ 1.60GHz, RAM 6Gb). As one can notice, the conversion made it possible to halve the size of the models, and significantly improved the speed. At the same time, we get a slight drop in accuracy, which, in our opinion, is acceptable. Although the weight of the audio model is still large, especially in comparison with the lightweight video model, its use can add flexibility to the system in situations when, for example, a face is not detected on the video. This is one of the advantages of multimodal approaches - they do not depend on absence of one modality.

TABLE II Models Perfomances

B. Qualitative Results

To demonstrate summarization, the video "Sykkuno & Toast Gets Dating Advice from

xQc" was taken from the YouTube, since people on it demonstrate vivid emotions.

The frames running in a row are quite similar to each other. To reduce computational costs without losing too much information, not every frame can be selected. During experiments we found out that selecting every 10 frames uniformly for a video at 30 frames per second provides good result. The following generally accepted palette of colors is used for the universal designation of emotions: anger - red, contempt - black, disgust - purple, fear - cyan, joy - yellow, neutrality - gray, sadness - blue, surprise - green.

The bar chart provides basic information for all people and the proportions of each type of emotion. By default, the columns are sorted by the total number of emotions found. As soon as the user sorts by a certain type of emotion, the columns of this type will align to the left for ease of comparison. For example, a chart sorted by fear in Fig. 4.

The pie chart allows to estimate the proportion of the presence of various emotions during the video for a particular person (Fig. 5). The user can target a specific emotion and get information about the number of frames on which this emotion was detected. The interface supports the option to select a specific person found during the video.

However, bar charts and pie charts do not convey information about moments of the video when and how long a person experienced a particular emotion, or how emotions replaced each other. To do this, we suggest a design in the form of a video track, on which one can assess when certain emotions appeared (Fig. 6a).

Fig. 4.

Stacked bar chart sorted by fear. Each horizontal bar represents one person, and the length of each sector shows the number of appearances of the corresponding emotion.

Show All

Fig. 5.

Pie chart of the total number of different emotions for a certain person.

Show All

To comply with the ratio of the length of the video in this visualization, an additional white color is used to indicate fragments when the person has not been detected. For some frames, the model may mistakenly predict an emotion and because of this, the video track will be full of colors. To eliminate this effect, we propose to do smoothing, that is, take the average prediction of emotions in a window of a certain size M (Fig. 6).

Fig. 6.

Color video track: (a) without smoothing of sequential frames; (b) smoothing with $M = 1$ (aggregation of 3 sequential frames); (c) smoothing with $M = 3$ (aggregation of 7 sequential frames); (d) smoothing with $M = 5$ (aggregation of 11 sequential frames).

Show All

SECTION V.

Conclusion and Future Work

In this paper, the novel video summarization pipeline (Fig. 1) for online conference tools based on emotion recognition was proposed. Several ways of summarization were explored, through interactive visualizations and by highlighting a sequence of frames with an emotional state of interest. Many applications can be found for such summaries, one of them is to show parents child's satisfaction during classes. To make used models easily and universally deployable for different devices, we converted them to FP16 format using OpenVINO. As a result, not only high accuracy (Table 1), but also great speed and model size (Table 2) are observed. Although, the resulting model for audio analysis is quite slow for real-time emotion recognition, its use for pre-recorded videos allows not to depend on poor face detection quality and other problems that occur during the analysis of visual content.

There are several possibilities for future work. First of all, our audio wav2vec 2.0 neural network is too heavy for practical applications. The model like QuartzNet [25] can be considered for faster real-time summarization. Moreover, it is necessary to try more advanced fusion techniques to improve the result, for example, transformers and frame/channel-level attention [15].

ACKNOWLEDGMENT

The authors would like to thank Lev Evtodienko and Sergey Vakhrameev for their help with training the audio model and conducting experiments.