Combining Audio and Video for Detection of Spontaneous Emotions

Rok Gajšek, Vitomir Štruc, Simon Dobrišek, Janez Žibert, France Mihelič, and Nikola Pavešić

Faculty of Electrical Engineering, University of Ljubljana
Tržaška 25, SI-1000 Ljubljana, Slovenia
{rok.gajsek,vitomir.struc,janez.zibert,simon.dobrisek,france.mihelic, nikola.pavesic}@fe.uni-lj.si
http://luks.fe.uni-lj.si/

Abstract. The paper presents our initial attempts in building an audio video emotion recognition system. Both, audio and video sub-systems are discussed, and description of the database of spontaneous emotions is given. The task of labelling the recordings from the database according to different emotions is discussed and the measured agreement between multiple annotators is presented. Instead of focusing on the prosody in audio emotion recognition, we evaluate the possibility of using linear transformations (CMLLR) as features. The classification results from audio and video sub-systems are combined using sum rule fusion and the increase in recognition results, when using both modalities, is presented.

Key words: bimodal emotion database, emotion recognition, linear transformations

1 Introduction

Recent years have seen an increase of analysis of psycho physical state of the users of the biometric systems. Especially, in the fields of audio and video processing. Recognising human emotions represents a complex and difficult task, the more so if the emotions recorded are of spontaneous nature. While human observers might have difficulties to correctly recognise the emotions of another individual in these situations, the task poses an even bigger challenge to automated systems relying on audio and video data. Various techniques have been proposed in the literature for the recognition of emotions from audio-video sequences; however, only a hand full of them were tested on sequences of spontaneous emotions.

In the article we present separately, the video and audio sub-systems for the emotion recognition. For video, a discrete cosine transformation (DCT) was applied, yielding a set of features, which were then used, with their deltas and classified using nearest neighbour classifier.

In emotion recognition from audio, different prosody features are generally used to capture emotion specific properties of the speech signal. Instead, we focused on using maximum likelihood linear transformations, otherwise heavily used in speaker or environment adaptation, as features for emotion classification.

2 Video emotion recognition

Many of the existing techniques make use of the coding system developed by Eckman [4] to encode different facial expressions in terms of so-called action units, which relate to the muscular configuration across the human face. Automated visual (emotion) recognition systems typically track a set of facial features across a video sequence trying to determine the action units based on which facial expressions can be categorised into emotional states. An overview of existing techniques for emotion recognition from video data can be found in [5].

Regardless of the technique used for recognising the emotional states of an individual from a video sequence, a number of issues needs to be solved in the design of a robust emotion recognition system. These issues commonly include: detecting the facial area in each video frame of the video sequence, photometrically and geometrically normalising the facial region, extracting and/or tracking of facial features in each video frame and classification of the final feature vector sequence into an emotional category. As we are concerned only with the detection and recognition of neutral and aroused emotional states in this paper, the final classification procedure represents a two class problem.

2.1 Face detection, tracking and normalisation

To detect the facial region in each video frame of the given video sequence we adopted the popular Viola-Jones face detector [6]. During the face detection stage all artifact not belonging to the facial region are removed from the video frames.

The key element of the employed face detector is an image representation called integral image, which allows visual features of image sub-windows of arbitrary sizes to be computed in constant time. Once these features over a predefined number of sub-windows sizes have been computed, AdaBoost is adopted to select a small set of relevant features that are ultimately fed to a cascade classifier. The classifier then performs the actual face detection by classifying each query face window into either the class of "faces" or the class of "non-faces".

While the face detectors usually results in a good performance, it still exhibits some shortcomings. The most troublesome issue we encountered when testing the face detector was the false assignment of image windows to the class of faces. To overcome this problem and consequently to improve the face detectors performance we further processed the detectors output using a skin-colour filter of the following form:

$$f_{sc}(I(x,y)) = \begin{cases} 1 \text{ , if } A \ \& \ B \ \& \ C \\ 0 \text{ , otherwise} \end{cases}, \qquad (1)$$

where $I(x,y)$ denotes the input image, f_{sc} denotes the skin-colour filter, the operator & denotes the logical AND, and the expressions A, B, and C represent conditions concerning the red (R), green (G) and blue (B) colour components of $I(x,y)$. , i.e., The skin-colour filter f_{sc} produces a binary mask with pixel values equal to one at every position that corresponds to the colour of the human skin. This binary mask is combined with the face detector by discarding all image subwindows, but the one with the highest number of skin-colour pixels. An example of a successful deployment of the employed face detector is shown in Fig. 1.

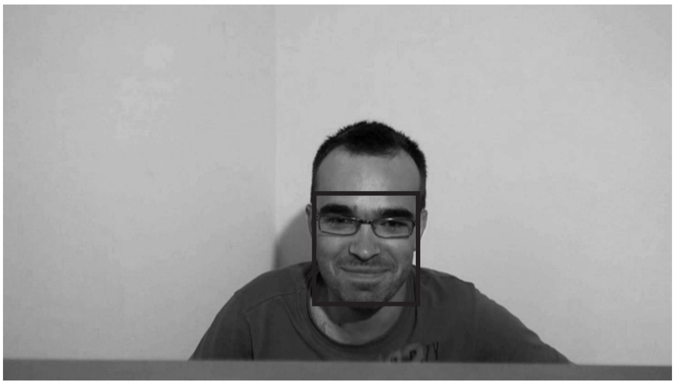

Fig. 1. An example of the Viola-Jones face detector output

The presented face detector was applied to each frame of the video sequence, resulting in a sequence of detected facial regions that formed the foundation for the normalisation step. During the normalisation procedure, the facial regions of each frame were first aligned (in such a way that the faces were in an upright position), their histogram was equalised to enhance the contrast and finally they were cropped to standard size of 128 × 128 pixels.

2.2 Feature extraction and classification

There are several options on which features to extract from a video sequence for the task of emotion recognition. Since this work describes our initial attempts in the design of a emotion recognition system, we decided to use simple holistic features extracted from each video frame by means of the discrete cosine transform (DCT).

The DCT is a popular image processing tool commonly used for image compression. When adopted as a feature extraction approach it is used to reduce the dimensionality of input frames from N to d, where $d << N$ and N stands for the number of pixels in the normalised facial region, i.e., $N = n \times n = 128 \times 128$. In our case, the selected value of d was 300.

Formally, the 1-dimensional DCT transform on a n-dimensional sequence $u(i)$, where $i = 1, 2, ..., n$, is defined as follows:

$$v(k) = \alpha(k) \sum_{i=0}^{n-1} u(i) cos(\frac{(2i+1)\pi k}{2n}), \tag{2}$$

where

$$\alpha(0) = \sqrt{\frac{1}{n}}, \text{ and } \alpha(k) = \sqrt{\frac{2}{n}}, \text{ for } 1 \leq k \leq n-1. \tag{3}$$

In the above expressions $v(k)$ denotes the DCT transformed sequence $u(i)$. Since the DCT transform represents a separable transform its 2-dimensional variant is obtained by first applying the 1-dimensional DCT to all image rows (which act as the sequences $u(i)$) and then to all image columns.

To encode some of the dynamic information contained in the video frames the feature vector of each frame computed via the 2-dimensional DCT was augmented with delta features defined as the difference of DCT feature vectors of two consecutive frames. The presented procedure resulted in a sequence of feature vectors of length $2d$. The sequence was finally classified into one of the two emotional classes using a variant of the nearest neighbour classifier.

3 Audio emotion recognition

Analysis of prosody has been the main focus of research in the field of emotion recognition from speech. Features, such as pitch and energy with their means, medians, standard deviations, minimum and maximum values [3], are generally combined with some higher level features, such as speaking rate, phone or word duration, etc. Language model based features, have also been studied [7]. In our work we evaluated the possibility of using linear transformations as emotion features.

3.1 Linear transformations of HMMs

Linear transformations of Hidden Markov Models are widely used in the fields of environment or speaker adaptation. A speaker or environment specific information is hidden in the linear transformation, which is then used in combination with the global acoustical model. Although, there are other ways of calculating the transformation matrix, in our work we concentrated on applying maximum likelihood estimation for determining the parameters of the transformation. This type of linear transformation is named Maximum Likelihood Linear Regression (MLLR) [8]. Even though, all parameters of a HMM model can be transformed using linear transformation, usually only the means and the variances of the Gaussian distributions are considered. A constrained version of MLLR (CMLLR), where the same transformation matrix is used for both, means and variances, was used. The transformation of a vector of means μ and the covariance matrix Σ is presented by the following equation.

$$\hat{\mu} = \mathbf{A}'\mu - \mathbf{b}' \quad , \quad \hat{\mathbf{\Sigma}} = \mathbf{A}'\mathbf{\Sigma}\mathbf{A}'^{\mathbf{T}} \tag{4}$$

The above equations present the CMLLR transformation in model space, but by applying equation (5), the same transformation can be applied to feature space as well.

$$\hat{\mathbf{o}(\tau)} = \mathbf{A}'^{\mathbf{-1}}\mathbf{o}(\tau) + \mathbf{A}'^{\mathbf{-1}}\mathbf{b}' = \mathbf{A}\mathbf{o}(\tau) + \mathbf{b}. \tag{5}$$

The matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are usually combined in one matrix $\mathbf{W} = [\mathbf{A}\ \mathbf{b}]$, that represents the CMLLR transformation.

3.2 Training procedure for the estimation of CMLLR transformations

Our goal was to evaluate the usage of CMLLR transformations as a feature for emotion recognition. In order to estimate these transformations for a particular emotion or arousal state, the global speaker independent acoustical model is required. A Voicetran database [9] composed of broadcast weather news and read speech, was used to build the monophone acoustic model. The following steps describe the procedure used to train the acoustical model.

- Calculation of the global means $\mu_o(s)$ and covariance matrices $\Sigma_o(s)$, for each speaker s in the database.
- Initialisation of the matrix $\mathbf{W_0(s)}$ for each speaker s using equation 6 and $\mu_0(s)$ and $\Sigma_0(s)$ from the first step.
- Training of the first acoustical model $\mathbf{AM_1}$ using $\mathbf{W_0(s)}$ transformation.
- Estimation of the of the new MLLR transformations $\mathbf{W_1(s)}$.
- The above two steps were then repeated four times, finally yielding the speaker independent model $\mathbf{AM_5}$
- Five alternations between the training of the acoustical model $\mathbf{AC_i}$ and the estimation of the $\mathbf{W_i}$ transformation matrix.
- After the fifth iteration the final acoustic model $\mathbf{AC_5}$ and $\mathbf{W_5(s)}$ are calculated

$$\mathbf{A(s)} = \mathbf{L_0(s)^T}, \mathbf{b(s)} = -\mathbf{L_0(s)^T}\mu_{\mathbf{o}}\mathbf{(s)}. \tag{6}$$

In equation (6), the $\mathbf{L_0(s)}$ is a Cholesky decomposition matrix of the inverse of the global covariance matrix $\mathbf{\Sigma_0(s)^{-1}}$.

The result from the above procedure relevant to the calculation of the CMLLR matrices for emotions, is the set of five acoustical models. These are used unchanged in a similar procedure for estimation of CMLLRs for emotion classification. The s parameter describing a particular speaker, now represent an emotional class. Again, a five-iteration process described above is applied, except for the training of the acoustical model in each step. The final set of emotion specific CMLLR transformations is acquired after the fifth iteration.

4 Bimodal emotional database

The basis, for the evaluation of proposed procedures, forms the audio-video database of spontaneous emotions - AvID [2]. A description of the process of labelling the data into different emotion groups is given.

4.1 Emotion labelling

When working with databases of spontaneous emotions, the task of labelling the emotions, presents a big challenge, as oppose to the databases with acted emotions. In the case of acted emotions, the person labelling the data has an a-priori knowledge about which emotional state is represented in the recording, whereas in the case of spontaneous database the decision needs to be based solely on the recording. Furthermore, the amount of different emotions represented in the acted databases can be controlled and normally they are equally distributed. In databases of spontaneous nature, this can be controlled to some extent (the second session) or can not be influenced directly at all (the first session).

For the task of labelling the AvID database we employed five students of psychology. Due to the time consuming nature of the labelling task, which is currently under way, not all recordings will be labelled by all five students. Still, the task of evaluating the agreement between the labellers presents the first obstacle we had to overcome. In the case of only two people labelling emotions, only the recordings were they both agreed on the label were initially used. Recordings labelled by three or more people provide other options that can be evaluated. For the initial tests we used the majority vote, meaning that, if the majority of the labellers agreed on the particular label, their decision was set as a reference. Since the recordings were transcribed in sentences, this formed the basis for evaluation of the agreement between labellers. Here, it should be noted that spontaneous speech, especially emotionally coloured, can not always be segmented into proper sentences. Thus, some sentences can be very short or can only contain just few syllables. Matching time between the labellers were analysed and are presented in the section 5.

5 Experimental results

From the AvID database, six sessions, fully transcribed and labelled as normal or aroused, were used as the basis for the evaluation. From this recordings a set of 668 MLLR matrices were calculated. Due to the spontaneous nature of the recordings, there was a strong prevalence of normal speech (529 MLLRs) over aroused speech (39 MLLRs). At least 15 seconds of audio was used for the estimation of each $39x39$ element MLLR matrix. The MLLR matrices were converted into vectors and linear and non-linear versions of Principal Component Analysis (PCA) were evaluated. Since this paper represents our initial attempts in audio-visual emotion recognition, a simple nearest neighbour classifier was selected for the classification phase. The full data set was divided into training set (80%) and testing set (20%) during five fold cross validation. Using the training data a class prototype was build for both normal and emotional state by taking the mean feature vector over all subjects in the class. For each MLLR derived feature vector, an Euclidean distance is calculated to both, neutral and emotional class prototypes. Based on the ratio between the two distances, the feature vector was classified. A comparison of linear and non-linear PCA (polynomial kernel of the fourth order was used) is presented using Detection Error Trade-off curves

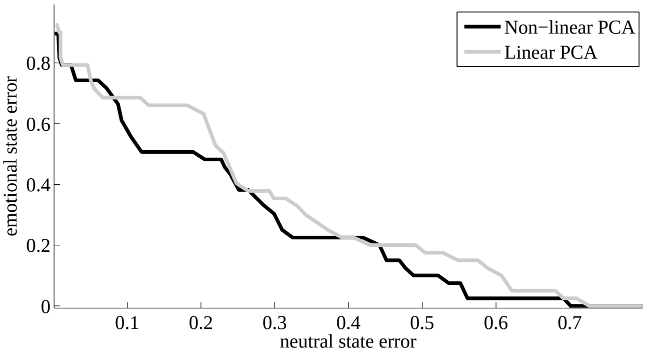

Fig. 2. Comparison of linear and non-linear PCA

(DET) in Fig. 2, which shows the average values of the five fold cross validation. The superior results are achieved with the non-linear version, which is reasonable, since the non-linear dependencies in data are also considered.

The video features, extracted as described in Sec. 2.2 were analysed using a similar procedure as above and classified using a variant of the nearest neighbour classifier. The Fig. 3 presents, in a form of a DET plot, an improvement after

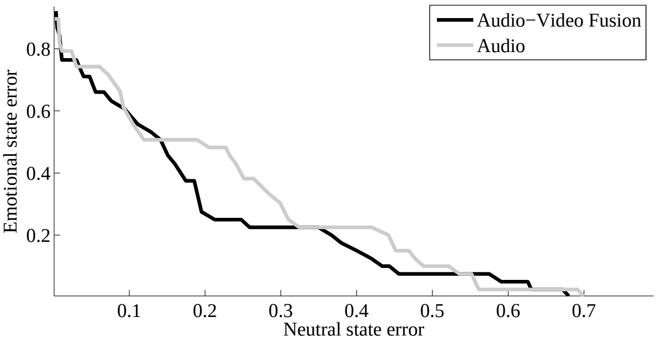

Fig. 3. Increase in classification accuracy of audio-video Fusion.

a sum rule fusion of audio and video. For more realistic assessment of the classification results, an agreement between the labellers should be discussed. One session, out of the six in question, was labelled individually by five people and the agreement reached between all five labellers was only in 59.1% of utterances. Agreement increases if only two people are labelling the data, which was the case for the rest of the sessions evaluated here, yielding an average of 89.3%.

The lowest equal error rate (ERR) that can be achieved by our current system, as shown in Fig. 3, is 24.2%

6 Conclusion

Our initial attempts in building an audio video emotion recognition system were presented. The task of labelling the AvID database of spontaneous emotions, which forms the basis for the evaluation of the emotion recognition system, was discussed. The issue of emotional labelling of spontaneous recordings was discussed and the agreement between the labellers was presented. The performance of our system was presented and compared to the agreement achieved between different labellers.

Acknowledgement

A three
line
acknowledgement.

References

1. Song, M., Chen, C., You, M.: Audio-visual based emotion recognition using tripled hidden Markov model. Acoustics, Speech, and Signal Processing, Proceedings, vol. 5, (ICASSP '04) , pp. 877–880 (2004)
2. Gajšek, R., et al.: Multi-Modal Emotional Database: AvID. Informatica 33 (2009), pp. 101–106, (2009)
3. Busso, C. et al.: Analysis of emotion recognition using facial expressions, speech and multimodal information. ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces, pp. 205–211. ACM, New York (2004)
4. Eckman, P.: Strong Evidence for universals in facial expressions. Psychol. Bull. 115(2), 268–287 (1994)
5. Pantic, M., Rothkrantz, L.J.M.: automatic analysis of facial expressions: the state of the art. IEEE TPAMI 22(12), 1424–1445 (2000)
6. Viola, P. and Jones M.: Robust real-time object detection. In: Proc. of the Second Intenrnational Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing and Sampling, Vancouver, Canada, (2001)
7. Ang, J., et al.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In Proc. ICSLP 2002, vol. 3, pp. 2037–2040 (2002)
8. Gales, M. J. F.: Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language, vol. 12 (2), pp. 75–98 (1998)
9. Mihelič, F., et al.: Spoken language resources at LUKS of the University of Ljubljana. Int. J. of Speech Technology, vol. 6 (3), pp. 221–232 (2006)