Multi-modal Emotion Recognition using Canonical Correlations and Acoustic Features

Rok Gajšek, Vitomir Štruc, France Mihelič
Faculty of Electrical Engineering, University of Ljubljana, Slovenia
{*rok.gajsek, vitomir.struc, france.mihelic*}*@fe.uni-lj.si*

Abstract

The information of the psycho-physical state of the subject is becoming a valuable addition to the modern audio or video recognition systems. As well as enabling a better user experience, it can also assist in superior recognition accuracy of the base system. In the article, we present our approach to multi-modal (audio-video) emotion recognition system. For audio sub-system, a feature set comprised of prosodic, spectral and cepstrum features is selected and support vector classifier is used to produce the scores for each emotional category. For video sub-system a novel approach is presented, which does not rely on the tracking of specific facial landmarks and thus, eliminates the problems usually caused, if the tracking algorithm fails at detecting the correct area. The system is evaluated on the eNTERFACE database and the recognition accuracy of our audio-video fusion is compared to the published results in the literature.

1. Introduction

The field of audio and video processing has seen over the past years, an increase in research of emotion-related analysis and modeling. The automatically detected emotional state of the subject promises to assist in increasing the performance of the systems dealing with the main challenges in audio and video processing such as identification and verification (audio and video) or speech recognition. The added information about the subject's emotional state can also provide crucial information, e.g. to dialog systems (if a user frustration is detected, the system could switch to a human operator), user modeling systems or can even enable a more secure and credible exchange of information [8].

Prosodic features usually form the basis for emotion recognition from audio signals [1], but other types of features have also been considered. In [3] we presented a successful application of linear transformations of Hidden Markov Models (HMM) as a feature for emotion recognition. Recently, the highest recognition rates are reported when spectral, cepstral and voice quality features are added to the "standard" prosodic features set [10]. Hence, the use of such a broad set of features for our audio sub-system is the obvious choice.

Most of the existing emotion recognition schemes relying on facial expression analysis use feature- or region-based approaches to determine the emotional state of the subject in the given image or video sequence. Sebe et al. [11] categorize the feature-based approaches as techniques that detect and track specific features, such as the corners of the mouth or eyebrows, and the region-based approaches as methods in which facial motion is measured in certain regions on the face such as the eye or mouth region. These techniques require detecting and tracking of specific facial landmarks throughout the entire length of the image- or video-sequence. We, however, take a different approach and perform emotion recognition from video data based on matching of image sets [4], [13]. This way, no tracking of individual facial landmarks is needed, instead it relies solely on the facial region as a whole, which can efficiently be extracted from the video sequence using existing face detection techniques [12].

The eNTERFACE emotion database [6] is used in development and evaluation of our system since it has been extensively tested in the past [5], [7], [9], and hence, the already reported results present the reference for our system.

2. Multi-modal emotion recognition system

The multi-modal emotion recognition system presented in this paper consist of two sub-systems relying on audio in video input. The video sub-system comprises several modules: *(i)* the face detection module that detects the facial region of the given video se-

quence, *(ii)* the subspace creation module, which constructs a subspace from the extracted facial images to encode the emotional state, and *(iii)* the matching module that compares the subspace constructed from the video sequence to the prototypical subspaces of the emotional classes using canonical correlations.

Similarly, the audio sub-system comprises of: *(i)* the feature extraction module that calculates the feature vector for each sample file and *(ii)* a matching module, which produces the scores, based on the support vector models of each class. Finally, both sub-systems are combined at the matching score level using weighted sum-rule fusion to make a multi-modal decision about the emotional state of the subject in the given video sequence.

3. Emotion recognition from video

3.1. Face detection

The first step in facial expressions analysis is the detection and tracking of the facial region throughout the entire video sequence. To this end we adopt the popular Viola-Jones face detector [12] and apply it to all frames of the given video sequence. The result of this procedure represents a set of facial images which are resized to a fixed size of 64×64 pixels and ultimately subjected to a histogram equalization procedure to account for any potential illumination variations present during the recording of the video sequence. Some sample images are shown in Fig. 1.

Figure 1. Sample facial regions extracted from a video sequence (Anger).

No geometric alignment of the facial region is performed on the extracted images, which adds additional robustness to our approach, as no localization of the specific facial landmarks is necessary.

3.2. Subspace creation

The second step in our video-processing-chain represents the creation of a subspace that encodes the emotion expressed in the given video sequence and is employed in the classification step to determine the best matching emotional state.

Let us assume that we have a set of facial images $\mathcal{X}_\mathcal{Z} = \{\mathbf{x}_i \in \mathbb{R}^d; \text{for } i = 1, 2, ..., n_\mathcal{Z}\}$ extracted from the given video sequence $\mathcal{Z}$. Here, $\mathbf{x}_i$ denotes the i-th d-dimensional facial image (in vector form) from the video sequence and $n_\mathcal{Z}$ stands for the number of frames in the sequence $\mathcal{Z}$. We assume that each of the $n_\mathcal{Z}$ facial images $\mathbf{x}_i$ can be decomposed into the following form:

$$\mathbf{x}_i = \hat{\mathbf{x}}_i + \mathbf{c}_i, \tag{1}$$

where $\hat{\mathbf{x}}_i$ represents the identity-specific (constant) part of the image $\mathbf{x}_i$, and $\mathbf{c}_i$ stands for the variable part of the image caused, for example, by variations in illumination, pose or facial expression.

While illumination changes are caused by external conditions, changes in pose and most of all facial expression can be linked to the emotional state of the subject. If we assume that the variable part $\mathbf{c}_i$ of the image represents a random variable drawn from the standardized normal distribution $\mathcal{N}(0, 1)$, then we can show that the video-sequence-conditional mean $\boldsymbol{\mu}_\mathcal{Z}$ represents an estimate of the constant identity-specific part of the images $\mathbf{x}_i$ given by Eq. (1), i.e.:

$$\boldsymbol{\mu}_\mathcal{Z} = \frac{1}{n_\mathcal{Z}}(\sum_{i=1}^{n_\mathcal{Z}} \hat{\mathbf{x}}_i + \sum_{i=1}^{n_\mathcal{Z}} \mathbf{c}_i) = \frac{1}{n_\mathcal{Z}} \sum_{i=1}^{n_\mathcal{Z}} \hat{\mathbf{x}}_i. \tag{2}$$

Based on the presented observation we can conclude that if we remove the mean $\boldsymbol{\mu}_\mathcal{Z}$ from all facial images $\mathbf{x}_i$ comprising the set $\mathcal{X}_\mathcal{Z}$, we arrive at a new image set encoding only the variable (or channel) part of the video sequence, i.e.: $\mathcal{C}_\mathcal{Z} = \{\mathbf{c}_i = \mathbf{x}_i - \boldsymbol{\mu}_\mathcal{Z}; \text{for } i = 1, 2, ..., n_\mathcal{Z}\}$. An example of the estimated identity-specific part as well as some channel images (corresponding to the Fig. 1) are shown in Fig. 2.

Figure 2. The estimated identity specific part of the images in the video sequence (left), channel images (right).

To capture the variability of the channel images into a subspace that can be used for classification, we compute a scatter matrix $\boldsymbol{\Sigma}$ from the set of channel images $\mathcal{C}_\mathcal{Z}$. The first step here is the construction of the channel matrix $\mathbf{C} \in \mathbb{R}^{d \times n}$, i.e., $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n]$. This matrix is then employed for computation of the scatter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$: $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^T$, where T denotes the transpose operator.

The subspace encoding the channel variations is ultimately determined by the leading eigenvectors (that

correspond to non-zero eigenvalues) of the following eigenproblem: $\mathbf{\Sigma w}_i = \lambda_i \mathbf{w}_i, \quad i = 1, 2, ..., d' \leq n$. To ensure that the computed subspace encodes mostly information related to the emotional state of the subject in the given video sequence, we discard the first three eigenvectors, which are usually linked to illumination changes. Thus, we obtain the following subspace from the given video sequence $\mathcal{Z}$: $\mathcal{W}_\mathcal{Z} = \{\mathbf{w}_i; \text{for } i = 4, 5, ..., d' \leq n\}$.

3.3. Building the emotional-class models

Here, we follow a similar approach as the one described in the previous section and compute a subspace for each emotional-class featured in the training data.

Let us assume that our training data consists of p sets of facial images (extracted from p video sequences) $\mathcal{X}_{\mathcal{Z}_1}, \mathcal{X}_{\mathcal{Z}_2}, ..., \mathcal{X}_{\mathcal{Z}_p}$ that correspond to N emotional classes labeled as $\omega_1, \omega_2, ..., \omega_N$. We then build a subspace $\mathcal{W}_{\omega_i}$ (for $i = 1, 2, ..., N$) for each emotional class by a simple eigen-decomposition of the emotion-specific scatter matrix $\mathbf{\Sigma}_{\omega_i}$, i.e.: $\mathbf{\Sigma}_{\omega_i} = \mathbf{C}_{\omega_i}\mathbf{C}_{\omega_i}^T$,, where $\mathbf{C}_{\omega_i}$ stands for the emotion-specific channel matrix, which is constructed by simply concatenating the channel matrices corresponding to the image sets $\mathcal{X}_{\mathcal{Z}_j \in \omega_i}$ of the emotional class ω_i.

As a result of the presented procedure, we obtain N subspaces $\mathcal{W}_{\omega_1}, \mathcal{W}_{\omega_2}, ..., \mathcal{W}_{\omega_N}$ that serve as prototypes for our N emotional classes.

3.4. Matching the subspaces

Let us consider two d'-dimensional linear subspaces $\mathcal{W}_\mathcal{Z}$ and $\mathcal{W}_\omega$. We can measure the similarity of the two subspaces in terms of the so-called canonical correlations, which represent cosines of principal angles $0 \leq \theta_1 \leq \theta_2 \leq ... \leq \theta_{d'} \leq (\pi/2)$ and are defined as follows:

$$cos\theta_i = \max_{\mathbf{w}_{\mathcal{Z}i} \in \mathcal{W}_\mathcal{Z}} \max_{\mathbf{w}_{\omega i} \in \mathcal{W}_\omega} \mathbf{w}_{\mathcal{Z}i}^T \mathbf{w}_{\omega i}, \tag{3}$$

subject to $\mathbf{w}_{\mathcal{Z}i}^T \mathbf{w}_{\mathcal{Z}i} = \mathbf{w}_{\omega i}^T \mathbf{w}_{\omega i} = 1$, $\mathbf{w}_{\mathcal{Z}j}^T \mathbf{w}_{\mathcal{Z}i} = \mathbf{w}_{\omega j}^T \mathbf{w}_{\omega i} = 0$, for $i \neq j$ [4], where the vectors $\mathbf{w}_{\mathcal{Z}i}$ and $\mathbf{w}_{\omega i}$ represent the i-th basis vectors of the subspaces $\mathcal{W}_\mathcal{Z}$ and $\mathcal{W}_\omega$, respectively.

The canonical correlations can be computed via Singular Value Decomposition (SVD) of the correlation matrix of the two subspaces. Let $\mathbf{W}_\mathcal{Z}$ and $\mathbf{W}_\omega$ stand for the matrices containing in their columns the orthonormal basis vectors of the subspaces $\mathcal{W}_\mathcal{Z}$ and $\mathcal{W}_\omega$. Then the SVD of the correlation matrix can be written as:

$$\mathbf{W}_\mathcal{Z}^T \mathbf{W}_\omega = \mathbf{Q}_{\mathcal{Z}\omega} \mathbf{\Lambda} \mathbf{Q}_{\omega\mathcal{Z}}, \tag{4}$$

where $\mathbf{\Lambda}$ stands for the diagonal matrix of canonical correlations, i.e., $\mathbf{\Lambda} = diag(cos\theta_1, cos\theta_2, ..., cos\theta'_d)$ and $\mathbf{Q}_{\mathcal{Z}\omega}$, $\mathbf{Q}_{\omega\mathcal{Z}}$ represent orthogonal matrices.

The canonical correlations measure the similarity of two subspaces. The first canonical correlation accounts for the similarity of the closest two basis vectors of the two subspaces, while the remaining ones carry information about the proximity of the basis vectors in other dimensions [4], [13]. For classification purposes we use only the first (the maximum) canonical correlation and define the similarity between two subspaces as $\delta(\mathcal{W}_\mathcal{Z}, \mathcal{W}_\omega) = cos\theta_1$. Thus, we formulate the classification problem as follows:

$$\delta(\mathcal{W}_\mathcal{Z}, \mathcal{W}_{\omega_k}) = max_{i=1}^N \delta(\mathcal{W}_\mathcal{Z}, \mathcal{W}_{\omega_i}) \mapsto \mathcal{W}_\mathcal{Z} \in \omega_k. \tag{5}$$

The above expression postulates that if the similarity between the subspaces $\mathcal{W}_\mathcal{Z}$ and $\mathcal{W}_{\omega_k}$ is the highest among the similarities to all N subspaces then the subspace $\mathcal{W}_\mathcal{Z}$ is assigned to the k-th class.

4. Emotion recognition from audio

Following the Interspeech 2009 Emotion Challenge [10], where a state-of-the-art emotion recognition feature set was presented, we based our audio sub-system on prosodic, cepstral and voice quality features. The selected low level descriptors (LLD) are: pitch frequency, root-mean-square value of energy, 12 MFCCs, harmonics-to-noise ratio (HNR) and zero crossing rate (ZRC). After delta coefficients are added to the LLDs, 12 functionals (maximum; minimum; range; max position; min position; mean; std. deviation; linear regression: offset, slope and mean-square-error; kurtosis and skewness) were applied. The open source utility openSMILE [2] is used to extract the described feature set.

Following the pairwise (one-vs-one) classification strategy, a sequential minimal optimization algorithm is used to train a support vector (SVM) classifier for each pair of emotions. After the training, a test sample is classified by each SVM and each time the wining emotion class receives a vote. The final emotion is determined by max-wins voting strategy, where the class with most votes is assigned to the test sample.

5. Experiments and results

A stratified 5-fold cross-validation (80% for training and 20 % for testing) is employed. A third degree polynomial kernel is used for SVM classification in audio sub-system. The highest recognition rate in the video sub-system is achieved using a 500-dimensional subspace (described in Section 3).

Table 1. Confusion matrices for Audio, Video and Fusion (from left to right)

	an	di	fe	ha	sa	su
an	**66**	3	6	5	2	1
di	7	**54**	9	11	2	3
fe	8	14	**45**	6	3	5
ha	11	10	6	**48**	1	10
sa	4	2	9	2	**64**	7
su	4	10	16	10	7	**44**

	an	di	fe	ha	sa	su
an	**50**	9	0	13	5	6
di	13	**47**	6	17	0	3
fe	16	8	**31**	6	11	9
ha	5	3	0	**70**	2	6
sa	10	1	8	14	**46**	9
su	18	2	11	11	16	**33**

	an	di	fe	ha	sa	su
an	**70**	3	4	4	1	1
di	6	**54**	12	8	2	4
fe	10	9	**49**	3	6	4
ha	4	4	5	**68**	0	5
sa	1	3	7	1	**68**	8
su	3	5	6	12	10	**55**

Table 2. Average recalls (%)

	An	Di	Fe	Ha	Sa	Su	All
Audio	79	64	57	58	73	49	**62.9**
Video	61	54	41	81	54	37	**54.7**
Fusion	84	63	63	80	78	60	**71.3**

Half of the test samples in each fold, form the evaluation data for estimating the fusion parameters, and the other half is used for the actual testing. To enable the fusion procedure, the min-max normalization is applied to both, audio and video set of scores. Confusion matrices, showing the distribution of the errors, for both subsystems are shown in Table 1., as well as the confusion matrix of the fusion. Emotion labels follow the labeling in the database: anger (an), disguise (di), fear (fe), happiness (ha), sadness (sa) and surprise (su). Comparing individual sub-system's average recall in Table 2. we can see that the audio sub-system has a superior recognition rate (62.9%), as oppose to the video sub-system (55%). However, a video recognition does extremely well in recognizing happiness (above 80%). The increase in accuracy, gained by audio-video fusion, can be observed in the last column in Table 2. The average recalls (over all emotions) for both sub-systems are superior to the one reported in [5] (37% for video and 33% for audio), plus our system does not relay on any facial landmark tracking, and hence is not that susceptible to tracking errors.

6. Conclusion

We presented our multi-modal emotion recognition system based on canonical correlations for video subsystem, and combination of prosody, spectral and cepstrum features for audio sub-system. The system was tested using the eNTERFACE database and by using a more robust approach for video sub-system, without the need to track certain areas of the face, a similar recognition rates as already reported, were achieved. In the future, we will focus on increasing the accuracy of the audio sub-system by evaluating the possibility of using linear transformations of HMMs as an emotion feature and testing the proposed method of emotion recognition from video on other databases.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18 (1)(1):32 – 80, January 2001.

[2] F. Eyben, M. Wllmer, and B. Schuller. openear - introducing the munich open-source emotion and affect recognition toolkit. In *Proc. of ACII 2009*, Amsterdam, pages 576–581., 2009.

[3] R. Gajšek, V. Štruc, S. Dobrišek, and F. Mihelič. Emotion recognition using linear transformations in combination with video. In *Proc. of Interspeech 2009*.

[4] T. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *TPAMI*, 29(6):1005–1018, June 2007.

[5] M. Mansoorizadeh and N. M. Charkari. Multimodal information fusion application to human emotion recognition from face and speech. Multi. Tools and App, 2009.

[6] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface'05 audio-visual emotion database. In *ICDEW '06*, Washington, DC, USA.

[7] M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory-based multimodal emotion recognition. In *MMM '09*, pages 435–446, Berlin, 2008.

[8] J. Pittermann, A. Pittermann, and W. Minker. *Handling Emotions in Human-Comp. Dialog*. Springer, Dordrecht (The Netherlands), 2009.

[9] B. Schuller. Speaker, noise, and acoustic space adaptation for emotion recognition in the automotive environment. In *Proc. 8th ITG conf. on Speech Comm.*, 2008.

[10] B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *Proc. Interspeech 2009*.

[11] N. Sebe, I. Cohen, and T. G. T. Huang. Multimodal approaches for emotion recognition : A survey. In *Proc. of SPIE*, volume 5670, pages 56–67, January 2005.

[12] P. Viola and M. Jones. Robust real-time face detection. *Int. J. of Comp. Vision*, 57(2):137 – 154, 2004.

[13] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proc. of AFGR*, pages 318–323, 1998.