Multimodal Emotion Recognition Based on the Decoupling of Emotion and Speaker Information

Rok Gajšek, Vitomir Štruc, and France Mihelič

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

Abstract. The standard features used in emotion recognition carry, besides the emotion related information, also cues about the speaker. This is expected, since the nature of emotionally colored speech is similar to the variations in the speech signal, caused by different speakers. Therefore, we present a gradient descent derived transformation for the decoupling of emotion and speaker information contained in the acoustic features. The Interspeech '09 Emotion Challenge feature set is used as the baseline for the audio part. A similar procedure is employed on the video signal, where the nuisance attribute projection (NAP) is used to derive the transformation matrix, which contains information about the emotional state of the speaker. Ultimately, different NAP transformation matrices are compared using canonical correlations. The audio and video sub-systems are combined at the matching score level using different fusion techniques. The presented system is assessed on the publicly available eNTERFACE'05 database where significant improvements in the recognition performance are observed when compared to the stat-of-the-art baseline.

Key words: speech, video, acoustic features, emotion recognition, multimodal databases

1 Introduction

The focus of the speech recognition and computer vision communities on emotion or affect related topics, has been increasing over the past years. Findings in the field of automatic emotion analysis can benefit other areas of interest in human computer interaction (HCI) such as automatic speech recognition (ASR) systems where performance drops significantly when the speech is emotionally colored, dialog managers, where a detection of a frustration in user's speech could redirect the call to a human operator, etc. Our previous work [1] leads us to believe that the phenomena of emotions in speech is by its nature similar to the way the speaker specific information is conveyed by the speech signal. Therefore, in this work we evaluate the possibility of extracting speaker specific information from the features, thus increasing the accuracy of the emotion recognition performance. This was achieved by estimating a linear transformation that mapped the original feature vector extracted from the audio signal, to the mean vector of the appropriate speaker. The columns of the transformation matrix and the bias vector concatenated together formed a new feature vector.

A similar procedure was adopted for the video subsystem. Here, a subspace encoding the emotional state of the subject in the video sequence was constructed, and compared to the prototypical subspaces of the emotional classes in the database. The two subspaces were compared using an image-set based method relaying on the computation of canonical correlations.

The results from the audio subsystem were fused with the video subsystem in order to evaluate the overall increase in accuracy of our multimodal emotion recognition system. The eNTERFACE'05 multimodal database [2] was used to evaluate the final recognition performance.

2 Audio-Video emotion recognition system architecture

The emotion recognition system consists of the two subsystems, one for each modality. Fig. 1 presents the structure of the system, where the left part represent the video subsystem and the right represents the audio part. Each systems performs the matching on its own and the scores for all classes are combined at the fusion level to produce the final score.

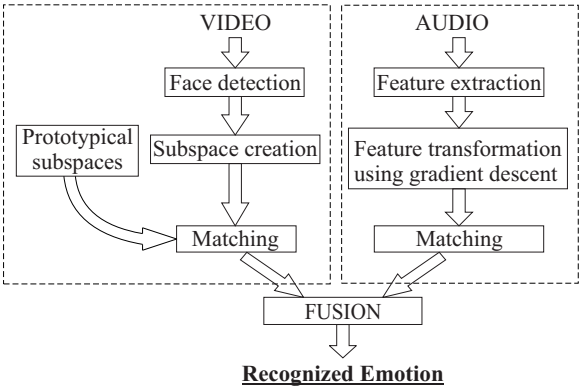

Fig. 1. Overvew of the multimodal emotion recognition system.

3 Audio subsystem

The audio subsystem consists of three major parts, as presented in the right part of the Fig. 1. First, the low level acoustic features are extracted from the audio signal. Then, the procedure of extracting emotion related information from the features and discarding the identity information is performed. Finally, the matching algorithm produces the scores for each sample. The above steps are described in more detail in the following sections.

3.1 Acoustic features

At the Interspeech '09 Emotion Challenge [3] the baseline feature set, consisting of 384 different features, was presented. In one part of the competition, where contestants were asked to produce their own feature sets that would surpass the baseline, there were none officially recognized contestants. This leads us to believe, that the proposed feature set forms the current stat-of-the-art in emotion recognition. In order to speed up the next step in the audio subsystem (Section 3.2) the whole feature set was reduced to 100 best features following the feature selection procedure based on mutual information as described in [4]. The comparison of the reduced and original feature set is presented in the Section 6.

3.2 Decoupling of emotion and speaker specific information

The idea of decoupling the emotional and speaker specific information is similar to the constrained version of MLLR (CMLLR) transformation [5] used for both speaker recognition and speech recognition (Eq. (1) shows the transformation of the mean vector in CMLLR).

$$\hat{\boldsymbol{\mu}} = \mathbf{A}'\boldsymbol{\mu} - \mathbf{b}' \tag{1}$$

Here, the matrix $\mathbf{A}$ and bias vector $\mathbf{b}$ are estimated for each speaker by increasing the likelihood of the acoustic model. This way, the speaker specific information is "moved" from the acoustic models representing base units (phones, triphones, etc.) to the transformation matrix and bias vector. Hence, the matrix $\mathbf{A}$ and vector $\mathbf{b}$ present an effective source of information for discriminating between different speakers. We propose a similar procedure of splitting the speaker specific information. The goal is to find a linear transformation which maps each speaker's sample to the mean value of all the samples from that particular speaker. Eq. (2) formulates the transformation, where $\hat{\boldsymbol{\mu}}_i$ is the average of all the samples from the i-th speaker, $\mathbf{x}_{i,j}$ is the j-th sample from the speaker i, and matrix $\mathbf{A}_j$ and vector $\mathbf{b}_j$ represent the transformation for the sample j.

$$\hat{\boldsymbol{\mu}}_i = \mathbf{A}_j\mathbf{x}_{i,j} + \mathbf{b}_j \tag{2}$$

Estimation of $\mathbf{A}$ and $\mathbf{b}$ is done employing a gradient descent based method. The cost function $J(\mathbf{A}_j, \mathbf{b}_j)$ for sample j is defined as shown in Eq. (3), where $\mathbf{x}_{i,j}$ is the j-th sample from the speaker i, and $\hat{\boldsymbol{\mu}}_i$ is the average for speaker i.

$$J(\mathbf{A}_j, \mathbf{b}_j) = \sum_{d=1}^{D}(\mathbf{A}_j\mathbf{x}_{i,j}^d + \mathbf{b}_j - \boldsymbol{\mu}_i)^2 \tag{3}$$

The partial derivatives of the cost function with respect to both variables $\mathbf{A}_j$ and $\mathbf{b}_j$ are shown in Eq. (5) and (4).

$$\frac{\partial J(\mathbf{A}_j, \mathbf{b}_j)}{\partial \mathbf{A}_j} = 2 * \sum_{d=1}^{D}(\mathbf{A}_j\mathbf{x}_{i,j}^d + \mathbf{b}_j - \hat{\boldsymbol{\mu}}_i) * \mathbf{x}_{i,j} \tag{4}$$

$$\frac{\partial J(\mathbf{A}_j, \mathbf{b}_j)}{\partial \mathbf{b}_j} = 2 * \sum_{d=1}^{D}(\mathbf{A}_j\mathbf{x}_{i,j}^d + \mathbf{b}_j - \hat{\boldsymbol{\mu}}_i) \tag{5}$$

In every iteration the new matrix $\hat{\mathbf{A}}_j$ and vector $\hat{\mathbf{b}}_j$ are estimated according to the Eq. (6) and (7).

$$\hat{\mathbf{A}}_j = \mathbf{A}_j - \delta_A \frac{\partial J(\mathbf{A}_j, \mathbf{b}_j)}{\partial \mathbf{A}_j} \tag{6}$$

$$\hat{\mathbf{b}}_j = \mathbf{b}_j - \delta_b \frac{\partial J(\mathbf{A}_j, \mathbf{b}_j)}{\partial \mathbf{b}_j} \tag{7}$$

When the cost function converges below the minimum threshold the final estimates of $\mathbf{A}_j$ and $\mathbf{b}_j$ are produced. Since the linear transformation converts each sample into the average feature vector for a specific features, the matrix $\mathbf{A}$ and vector $\mathbf{b}$ are believed to hold only the information about the speaker's emotional state. Therefore, the columns from the matrix $\mathbf{A}_j$ and the bias vector $\mathbf{b}_j$ are concatenated to form the new feature vector.

3.3 Matching of the acoustic features

For the task of producing the recognition scores, in order to enable the future fusion with video, support vector machines classifier (SVM) with a linear kernel, was employed. The one-versus-one approach was used for dealing with the five class recognition task. The setup used to produce the acoustic scores is described in more detail in Section 5.

4 Video subsystem

The video subsystem comprises three main modules, as depict on the left hand side of Fig. 1: the face detection module, which detects and extracts the facial regions from the individual frames of the video sequence, the subspace creation module, which constructs a subspace from the given video sequence encoding mostly the emotion specific information, and finally, the matching module which compares the constructed subspace with the emotion-specific subspaces stored in the systems database. The described modules are presented in more detail in the remainder of this section.

4.1 Face Detection

Extraction and tracking of the facial region during the entire length of the given video sequence is done using the established Viola Jones face detector. The description of the detector would exceed the scope of this paper; however, the interested reader is referred to [6] for more information. An example of the output of the face detection module when applied on a sample video sequence is shown in Fig. 2

Fig. 2. An example of the output of the face detection module.

4.2 Decoupling of Emotion and Speaker Information

The second module in the video subsystem represent the subspace creation module. Here, a subspace is created from the output images of the face detector in such a way that the emotion specific information in the subspace is enhanced, while the subject (or better said video sequence) specific information is decreased.

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:

$$\mathbf{x}_i = \hat{\mathbf{x}}_i + \mathbf{c}_i, \tag{8}$$

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, changes in the emotional state of the subject shown in the image.

Let us now assume that the variable part $\mathbf{c}_i$ of the image represents a random variable drawn from the normal distribution $\mathcal{N}(0, 1)$. It is possible to show that the video-sequence-conditional mean $\boldsymbol{\mu}_\mathcal{Z}$ represents an estimate of the constant identity-specific part of the images $\mathbf{x}_i$ (see [7] for details). Based on this 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/emotion) part of the video sequence, i.e.:

$$\mathcal{C}_\mathcal{Z} = \{\mathbf{c}_i = \mathbf{x}_i - \boldsymbol{\mu}_\mathcal{Z}; \text{for } i = 1, 2, ..., n_\mathcal{Z}\}. \tag{9}$$

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 $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$:

$$\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^T, \tag{10}$$

where T represents the transpose operator.

The subspace encoding the channel variations is finally determined by the leading eigenvectors (that correspond to non-zero eigenvalues) of the following eigenproblem:

$$\boldsymbol{\Sigma}\mathbf{w}_i = \lambda_i \mathbf{w}_i, \quad i = 1, 2, ..., d' \leq n. \tag{11}$$

If the scatter matrix is computed only from one test video sequence we obtain a subspace, that needs to be classified into one of the emotion classes. On the other hand, if the subspace is computed from all training video sequences of a given emotional state, we obtain the class prototypes (subspaces) for the specific emotion.

4.3 Matching the subspaces

The last module in the video processing chain is the matching module, where the test subspace and prototype subspace are compared. Let us consider two d'-dimensional

linear subspaces $\mathcal{W}_\mathcal{Z}$ and $\mathcal{W}_\omega$, where the subspace $\mathcal{W}_\mathcal{Z}$ can be thought of as a subspace extracted from a test video sequence and the subspace $\mathcal{W}_\omega$ represents the prototype subspace for the class labeled ω. We can measure the similarity of the two subspaces in terms of the canonical correlations, which are defined as cosines of principal angles $0 \leq \theta_1 \leq \theta_2 \leq ... \leq \theta_{d'} \leq (\pi/2)$, i.e.:

$$cos\theta_i = \max_{\mathbf{w}_{\mathcal{Z}i} \in \mathcal{W}_\mathcal{Z}} \max_{\mathbf{w}_{\omega i} \in \mathcal{W}_\omega} \mathbf{w}_{\mathcal{Z}i}^T \mathbf{w}_{\omega i}, \tag{12}$$

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 $i \neq j$ [8], where the vectors $\mathbf{w}_{\mathcal{Z}i}$ and $\mathbf{w}_{\omega i}$ denote 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:

$$\mathbf{W}_\mathcal{Z}^T \mathbf{W}_\omega = \mathbf{Q}_{\mathcal{Z}\omega} \boldsymbol{\Lambda} \mathbf{Q}_{\omega\mathcal{Z}}, \tag{13}$$

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

The first canonical correlation accounts for the similarity of the closest two basis vectors of the two subspaces $\mathcal{W}_\mathcal{Z}$ and $\mathcal{W}_\omega$, while the remaining ones hold information about the proximity of the basis vectors in other dimensions [8], [9]. 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:

$$\delta(\mathcal{W}_\mathcal{Z}, \mathcal{W}_{\omega_k}) = max_{i=1}^N \delta(\mathcal{W}_\mathcal{Z}, \mathcal{W}_{\omega_i}) \mapsto \mathcal{W}_\mathcal{Z} \in \omega_k. \tag{14}$$

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.

5 Experimental setup

The tests were conducted using the eNTERFACE'05 multimodal database [2], which consists of 6 different emotion classes. The five fold cross validation protocol was used, where in each fold 80% of samples were used for training, 10% comprised the development set for training the parameters of fusion, and 10% were used for testing.

The openSMILE toolkit [10] was used to produce the 384 features as described in Section 3.1. The feature set consists of spectral features (1–12 MFCCs), prosodic features (F0, energy), voice quality features (harmonics to noise ratio) and zero-crossing-rate (ZRC). To this low level descriptors, 12 functionals are applied, thus producing the starting feature vector for each sample recording. Next, the number of features was reduced to only 100 most discriminative ones, using the algorithm described in [4]. This step was undertaken not to improve the classification scores, but to enable faster computation times for the estimation of linear transformations in the next step. As described in Section 3.2, the matrix $\mathbf{A}$ and vector $\mathbf{b}$ are estimated in order to transform

each sample in to the mean feature vector for the corresponding speaker. Hence, the transformation contains more concentrated information about the emotional state, and the speaker specifics are discard. Since the dimension of a feature vector at this step is of size 100, the transformation matrix **A** is of size 100 times 100, and bias vector **b** is of size 100. Concatenating the columns of **A** and the vector **b**, thus comprise a new feature vector of size 10100. For the gradient descent procedure, an identity matrix of size 100 x 100 was used for the initial value of **A**, and a vector of all ones for the starting value of **b**.

In the work, presented in this article we did not use a speaker identification system, which would first predict who the speaker in the sample is in order to select the appropriate speaker's mean vector. Instead, we presumed that we know the identity of the speaker and automatically selected the corresponding speaker's mean vector. The described approach was selected in order to evaluate the assumption of speaker specific information being correlated with the emotional state.

A version of sequential minimal optimization SVM was used to produce a model for each pair of emotion classes, following a one-versus-one protocol. Counting the number of wins for each emotion, the score for each sample is produced. Both, the video and the audio scores were normalized using min-max normalization [11]. A product rule fusion was used to combine the matching scores from both subsystems. The parameters of fusion were first determined on the development set, and the final recognition scores were produced on the test set.

6 Results

The standard measure of accuracy in emotion recognition systems has become the unweighted average recall, since it is useful in systems where the emotional classes are not balanced, which is usually the case in databases of spontaneous emotions. In our case the database is balanced, but in order to make our results comparable to the others in the literature, all the results presented in the Table 1 are unweighted average recalls over all emotion classes.

AUDIO subsystem			VIDEO	FUSION
original features (384)	reduced features (100)	decoupled features	subsystem	
61.2%	57.01%	**66.03%**	**54.61%**	**74.33%**

Table 1. Comparison of average recalls for audio, video and multimodal emotion recognition.

First, we evaluated the recognition performance of the original acoustic feature set where a recognition accuracy of 61% was achieved. With the reduction of the number of features from 384 to only 100 most discriminative ones, the recognition rate deteriorated, and the average recall over all folds dropped for approximately 4% absolute. But with the proposed decoupling of speaker specific information using gradient descent method the recognition accuracy jumped to 66.03%, which is a relative increase

of 15%. After the fusion with the scores from the video subsystem, which achieves on its own an accuracy of 54.61%, the final recognition performance of our system climbs to 74.33%.

7 Conclusion

In the paper we presented a method of decoupling the emotion and speaker specific information form the acoustic features, usually used in an emotion recognition systems. The proposed method was evaluated on a popular multimodal database eNTERFACE'05, which enabled the fusion of audio and video. We have shown that if we use a smaller set of features (reducing from 384 to 100) in combination with the proposed method of extracting emotion specific information and discard the speaker information, we can achieve an increase in recognition performance of 15% over the reduced feature set. Over the baseline system we increased the recognition performance by 8% relative. In the future we will focus on modifications of the proposed algorithm in order to be able to use it on a larger set of baseline features.

References

1. Gajšek, R., Štruc, V., Dobrišek, S., Mihelič, F.: Emotion recognition using linear transformations in combination with video. In: Proceedings of Interspeech 2009. (2009)
2. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The enterface'05 audio-visual emotion database. In: ICDEW '06: Proceedings of the 22nd International Conference on Data Engineering Workshops, Washington, DC, USA, IEEE Computer Society (2006)
3. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In ISCA, ed.: Proceedings of Interspeech 2009. (2009) 312–315
4. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1226–1238
5. Gales, M.J.F.: Maximum likelihood linear transformations for hmm-based speech recognition. Computer Speech and Language **12** (1997) 75–98
6. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (May 2004) 137 – 154
7. Štruc, V., Vesnicer, B., Mihelič, F., Pavešić, N.: Removing illumination artefact from face images using the nuisance attribute projection. In: ICASSP 2010. (2010) 846 – 849
8. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. TPAMI **29**(6) (June 2007) 1005–1018
9. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: Proc. of AFGR. (1998) 318–323
10. Eyben, F., Wllmer, M., Schuller, B.: openear - introducing the munich open-source emotion and affect recognition toolkit. In: Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009), Amsterdam, The Netherlands. Volume I., IEEE (2009) 576–581
11. Jain, A.K., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition **38**(12) (2005) 2270–2285