INTECH
open science | open minds

OPEN ACCESS
ARTICLE

International Journal of Advanced Robotic Systems

Towards Efficient Multi-Modal Emotion Recognition

Regular Paper

Simon Dobrišek[1], Rok Gajšek[1], France Mihelič[1], Nikola Pavešić[1] and Vitomir Štruc[1,*]

1 Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
* Corresponding author E-mail: vitomir.struc@fe.uni-lj.si

Received 30 Jul 2012; Accepted 2 Oct 2012

DOI: 10.5772/54002

Abstract The paper presents a multi-modal emotion recognition system exploiting audio and video (i.e., facial expression) information. The system first processes both sources of information individually to produce corresponding matching scores and then combines the computed matching scores to obtain a classification decision. For the video part of the system, a novel approach to emotion recognition, relying on image-set matching, is developed. The proposed approach avoids the need for detecting and tracking specific facial landmarks throughout the given video sequence, which represents a common source of error in video-based emotion recognition systems, and, therefore, adds robustness to the video processing chain. The audio part of the system, on the other hand, relies on utterance-specific Gaussian Mixture Models (GMMs) adapted from a Universal Background Model (UBM) via the maximum a posteriori probability (MAP) estimation. It improves upon the standard UBM-MAP procedure by exploiting gender information when building the utterance-specific GMMs, thus ensuring enhanced emotion recognition performance. Both the uni-modal parts as well as the combined system are assessed on the challenging multi-modal eNTERFACE'05 corpus with highly encouraging results. The developed system represents a feasible solution to emotion recognition that can easily be integrated into various systems, such as humanoid robots, smart surveillance systems and alike.

Keywords Emotion Recognition, Video Processing, Speech Processing, Canonical Correlations, GMM-UBM

1. Introduction

Augmenting humanoid robotic systems with emotion recognition capabilities has recently attracted a lot of attention from both, the speech and computer vision communities. This increased attention resulted in a plethora of methods that can be found in the literature and pertain to the field of emotion recognition.

In this paper we build upon our work presented in [1, 2] and present a novel multi-modal emotion recognition system exploiting video (i.e., facial expression) and audio information. The proposed system processes each source of information separately and then combines the results and the matching score level. Both the video- and audio-processing parts of the system are implemented using novel approaches that improve upon existing methods from the literature.

Existing video-based emotion recognition techniques, for example, are typically grouped into [3]:

- *feature-based techniques* that detect and track specific facial features, such as the corners of the mouth or eyebrows, and use the obtained information to conduct emotion recognition, and
- *region-based approaches*, where facial motion is first measured on certain regions of the face, such as the eye or mouth region, and then exploited for emotion recognition.

Both types of methods require the detection and tracking of specific facial landmarks throughout the entire length of the image- or video-sequence and are, due to the difficulty of this task, also prone to error [1]. In this paper, we take a fundamentally different approach and develop a novel method for emotion recognition from video data that adopts matching of image sets [4, 5]. With the proposed approach, no tracking of individual facial landmarks is needed. Instead, the procedure relies solely on the facial region as a whole, which can be robustly and efficiently extracted from video data using existing face detection techniques, as for example, the Viola-Jones face detector [6].

Similarly as for the video modality, numerous techniques for audio-based emotion recognition can also be found in the literature. Here, the techniques differ mainly in terms of the modeling approach used to represent the given audio features. Schuller et al. [7] classify the existing techniques into two classes:

- *frame-level modeling techniques*, which build statistical models of feature vectors extracted from overlapping frames of a given utterance, and
- *supra-segmental modeling techniques*, where a number of statistical functionals are applied to the frame-level features of one utterance, yielding a single high-dimensional feature vector per utterance.

The low level acoustic features for both types of modeling techniques typically consists of spectral, prosodic and voice quality features [7, 8]. Although, the final recognition performance for both types of modeling techniques depends heavily on the classification method adopted, it was shown by various group evaluations that both types of modeling techniques are capable of yielding state of-the-art recognition results for the task of emotion recognition from audio data [9, 10].

The contribution of this paper with respect to audio-based emotion recognition stems from an improved approach to frame-level modeling, which relies on Gaussian Mixture Models (GMMs). While in the commonly used approach a single Universal Background Model (UBM) is built first from the extracted acoustic feature vectors and then adapted by Maximum A Posteriori (MAP) adaptation, the possibility of decoupling emotion specific information from other paralinguistic cues, which exist in the UBM-MAP derived GMMs, is examined in this paper. We show that by exploiting gender information, we consistently improve upon the recognition performance of the standard UBM-MAP technique.

We assess both proposed uni-modal approaches using the eNTERFACE'05 [11] database, which is one of the few freely available databases containing multi-modal recordings of various types of emotions. Additionally, we study the applicability of several fusion schemes to further improve upon the results obtained with the individual modalities. Our results show that both uni-modal approaches as well as the proposed combined system compare favorably with state-of-the-art techniques from the literature [9, 12, 13, 14].

The rest of the paper is structured as follows. In Section 2 we elaborate on the proposed multi-modal emotion recognition system and describe in detail the video and audio processing parts of the system as well as the fusion schemes used to combine both parts into a coherent system. In Section 3 we present the experimental database, on which our system was evaluated, and the main findings of the paper. We conclude the paper with some final comments in Section 4.

2. System description

2.1 Overview

The multi-modal emotion recognition system introduced in this paper consists of an audio and a video sub-system. Figure 1 presents the basic structure of the system. Each subsystem first processes its corresponding input and then produces a matching score. The two scores are then fused at the matching-score level to allow for a reliable classification decision.

The video sub-system comprises:

- *a face detection module* that detects the facial region throughout the given video sequence,
- *a subspace creation module*, which constructs a subspace from the extracted facial images to encode the emotional state, and
- *a 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:

- *a feature extraction and modeling module* that calculates the feature vectors from each sample recording and creates a probabilistic model from the computed feature vectors and
- *a matching module*, which produces the scores, based on the support vector models of each class.

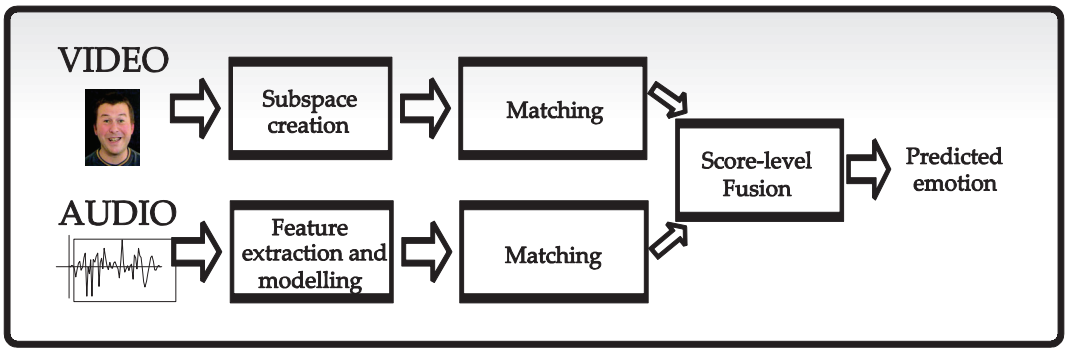

Figure 1. Block diagram of the multi-modal emotion recognition system

Figure 2. Sample facial regions extracted from a video sequence depicting the emotion "anger".

A detailed description of all system parts is presented in remainder of the paper.

2.2 The video sub-system

This section introduces our approach to emotion recognition from video data. It presents all the procedural steps that need to be taken to achieve reliable emotion recognition using holistic (appearance-based) techniques applied to image sets.

Face detection

The first procedural step required for building an emotion recognition system based on video data is the extraction of the region of interest from each frame of a given video sequence.

As our video sub-system relies on facial expression analysis, we adopt the established Viola-Jones face detector [6] for this purpose and employ it to detect the boundaries of facial regions in each frame of the currently processed video. Once the entire video sequence is processed, we resize the detected regions to a fixed size of 64×64 pixels and finally photometrically normalize them using histogram equalization. The result of the described procedure is a set of facial images as shown in Figure 2.

Note here that no geometric alignment of the facial regions based on specific landmarks is performed, which significantly increases the robustness of our approach when compared to existing methods from the literature, as no (error-prone) facial-landmark-localization procedure is needed [1, 2].

Figure 3. The estimated identity specific part of the images in the video sequence (left), channel images (right)

Subspace creation

The extracted and normalized facial regions constructed with the procedure presented in the previous section form the foundation for the second step of our video sub-system, namely, the creation of a subspace that relates to the emotional state expressed in the given video sequence.

To facilitate the theoretical derivation of our subspace creation procedure let us consider a set of facial images $\chi_v = \{\boldsymbol{x}_i \in \mathbb{R}^d, \text{for } i = 1, 2, \ldots, n_v\}$ extracted from a given video sequence v. Here, $\boldsymbol{x}_i$ represents the i-th d-dimensional facial image (in vector form) from the sequence v and n_v denotes the total number of frames constituting v. When building a subspace from the facial images in χ_v, we assume that each image $\boldsymbol{x}_i$ can be decomposed into a constant, identity-specific part $\hat{\boldsymbol{x}}_i$ and a variable part $\boldsymbol{c}_i$ (often referred to as the channel part) caused by non-identity related factors, such as illumination, pose and/or facial expression. Thus, we can write:

$$\boldsymbol{x}_i = \hat{\boldsymbol{x}}_i + \boldsymbol{c}_i. \quad (1)$$

Since we presume that illumination changes are satisfactorily compensated for with our histogram equalization procedure (and the exclusion of the first three basis vectors of the created subspace), the remaining variability must inevitably be linked to pose and facial expression changes, which are reasonable indicators of the emotional state of the subject shown in the given video sequence. Clearly, if we were able to estimate the variable part of each image in χ_v, we could estimate an emotion-specific subspace that could serve as the basis for recognition.

Let us assume that the variable part of the images $\boldsymbol{c}_i$, for $i = 1, 2, \ldots, n_v$, represents a random variable drawn from the standardized normal distribution $N(0,1)$. The video-

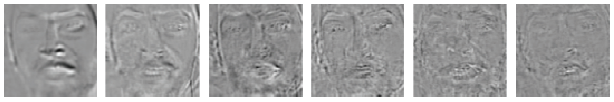

Figure 4. Some examples of the computed subspace basis for the video sequence shown in Figure 2

sequence-conditional mean $\boldsymbol{\mu}_v$ then serves as the (variation-free) estimate of the constant identity-specific part of the images $\boldsymbol{x}_i$ of v, as shown by the following expression:

$$\boldsymbol{\mu}_v = \frac{1}{n_v}\left(\sum_{i=1}^{n_v} \widehat{\boldsymbol{x}}_i + \sum_{i=1}^{n_v} \boldsymbol{c}_i\right) = \frac{1}{n_v}\sum_{i=1}^{n_v} \widehat{\boldsymbol{x}}_i. \quad (2)$$

Considering this observation, we can conclude that removing the sequence-specific mean $\boldsymbol{\mu}_v$ from all images in χ_v results in a new set C_v that encodes only the variable part of the video sequence, i.e.:

$$C_v = \{\boldsymbol{c}_i = \boldsymbol{x}_i - \boldsymbol{\mu}_v; \text{ for } i = 1, 2, \ldots, n_v\}. \quad (3)$$

An example of the estimated identity-specific part as well as some channel images (computed based on the sequence shown in Figure 2) are presented in Figure 3. Note how fairly well the sequence-specific mean captures the identity of the subject shown in the video sequence, while the channel images capture the variability caused by pose and facial expression changes.

To capture the information contained in this set into a subspace useful for emotion recognition, we first compute a scatter matrix $\boldsymbol{\Sigma}$ from all images in C_v. If we arrange the image in C_v into the matrix $\boldsymbol{C} \in \mathbb{R}^{d \times n_v}$, where $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{n_v}]$, then the scatter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, can be computed as

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^T, \quad (4)$$

where ^{T} denotes the transpose operator.

Finally, the subspace encoding the variable part of the facial images (i.e., the maximum variance directions [15]) is characterized by the leading eigenvectors (corresponding to non-zero eigen-values) of the following eigen-problem:

$$\boldsymbol{\Sigma}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i, i = 1, 2, \ldots, \; d' \leq n_v. \quad (5)$$

It should be noted that for classification purposes we discard the first three computed eigenvectors, as these usually correlate heavily with illumination changes. Thus, for a given video sequence v we construct a subspace W_v of the following form:

$$W_v = \{\boldsymbol{w}_i; \text{ for } i = 4, 5, \ldots, \; d' \leq n_v\}. \quad (6)$$

Some examples of the subspace basis (in image form) corresponding to the video sequence in Figure 2 are shown in Figure 4.

Constructing the templates

To be able to compare the subspaces computed from individual video sequences, we require some prototypical subspaces that serve as templates for our emotional classes. To construct these templates, we follow a similar approach as the one presented in the previous section and compute a subspace for each emotional-class featured in the training data.

Assume that our training data comprises p sets of facial images extracted from p different video sequences, i.e., χ_{v_1}, χ_{v_2}, ..., χ_{v_p}. Furthermore, assume that these sets correspond to N emotional-classes with the corresponding class labels ω_1, ω_2, ..., ω_N. The prototypical subspaces (or templates) W_{ω_i} (for i = 1, 2, ..., N) are then constructed by a simple eigen-decomposition of the emotion-specific scatter matrices $\boldsymbol{\Sigma}_{\omega_i}$, i.e.:

$$\boldsymbol{\Sigma}_{\omega_i} = \boldsymbol{C}_{\omega_i}\boldsymbol{C}_{\omega_i}^T, \quad (7)$$

where $\boldsymbol{C}_{\omega_i}$ denotes the matrix containing the variable part of all image sets $\chi_{v_j} \in \omega_i$ (for $j \in 1, 2, \ldots, p$).

The described procedure results in N subspaces W_{ω_1}, W_{ω_2}, ..., W_{ω_N} that serve as templates for our N emotional classes.

Subspace matching

Consider two d'-dimensional linear subspaces W_v and W_ω. Within the proposed video-based emotion recognition framework we measure the similarity of the two subspaces using canonical correlations, which represent cosines of principal angles $0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_d' \leq (\pi/2)$ and are defined as [4]:

$$\cos\theta_i = \max_{\boldsymbol{w}_{v_i} \in W_v} \max_{\boldsymbol{w}_{\omega_i} \in W_\omega} \boldsymbol{w}_{v_i}^T \boldsymbol{w}_{\omega_i}, \quad (8)$$

subject to $\boldsymbol{w}_{v_i}^T \boldsymbol{w}_{v_i} = \boldsymbol{w}_{\omega_i}^T \boldsymbol{w}_{\omega_i} = 1$, $\boldsymbol{w}_{v_j}^T \boldsymbol{w}_{v_i} = \boldsymbol{w}_{\omega_j}^T \boldsymbol{w}_{\omega_i} = 0$, for $i \neq j$ [4], where the vectors $\boldsymbol{w}_{v_i}$ and $\boldsymbol{w}_{\omega_i}$ represent the i-th basis vectors of the subspaces W_v and W_ω, respectively.

The canonical correlations can be computed via Singular Value Decomposition (SVD) of the correlation matrix of the two subspaces. Let $\boldsymbol{W}_v$ and $\boldsymbol{W}_\omega$ stand for the matrices containing in their columns the orthonormal basis vectors of the subspaces W_v and W_ω. Then the SVD of the correlation matrix canbe computed as follows [4]:

$$\boldsymbol{W}_v^T \boldsymbol{W}_\omega = \boldsymbol{Q}_{v\omega} \boldsymbol{\Lambda} \boldsymbol{Q}_{\omega v}, \quad (9)$$

where $\boldsymbol{\Lambda}$ stands for the diagonal matrix of canonical correlations, i.e., $\boldsymbol{\Lambda} = diag(\cos\theta_1, \cos\theta_2, \ldots, \cos\theta_d')$ and $\boldsymbol{Q}_{v\omega}$, $\boldsymbol{Q}_{\omega v}$ represent orthogonal matrices.

As we have emphasized above, the canonical correlations measure the similarity between 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, 5]. For classification purposes we use only the first (the maximum) canonical correlation and define the similarity between two subspaces as $\delta(W_v, W_\omega) = \cos\theta_1$. Thus, we formulate the classification problem as follows:

$$\delta(W_v, W_{\omega_k}) = max_{i=1}^{N} \delta(W_v, W_{\omega_i}) \rightarrow W_v \in \omega_k. \quad (10)$$

The above expression postulates that in case the similarity between the subspaces W_v and W_{ω_k} is the highest among the similarities to all N subspaces, then the subspace W_{ω_k} is assigned to the k-th class.

2.3 The audio sub-system

The audio part of our emotion recognition system builds on the traditional UBM-MAP technique of representing acoustic feature vectors. In this section we elaborate on our approach and describe the entire procedure of emotion recognition based on audio data.

Acoustic features

The acoustic feature vectors used in our experiments comprise of the standard set of 1-12 Mel-frequency Cepstral Coefficients (MFCC) plus energy. The MFCC features are first smoothed with a moving average filter of length 3 and then normalized using Cepstral Mean Normalization (CMN). In order to include temporal information as well, the first order delta coefficients are also generated and added to the feature vector. Thus, the final length ofthe feature vector equals 26. The described procedure is implemented using the open SMILE feature extractor [16].

GMM-UBM modeling

The frame-level features presented in the previous section are used to construct Gaussian mixture models (GMMs), which represent generative statistical models capable of characterizing arbitrary data distributions [17, 18].

Formally, a GMM is defined as a linear combination of several multivariate Gaussian probability density functions (PDFs), i.e.,

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\boldsymbol{x}), \quad (11)$$

where w_i denotes the weight associated with the i-th Gaussian PDF $p_i(\boldsymbol{x})$:

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}. \quad (12)$$

In the above equations $\boldsymbol{\mu}_i$ denotes the mean vector of the i-th Gaussian PDF, $\boldsymbol{\Sigma}_i$ denotes the covariance matrix of the i-th Gaussian PDF, d stands for the dimensionality of the PDF and $\lambda = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, w_i\}$ (for i = 1, 2, ..., M), represents the set of GMM parameters. Note that a M-component GMM is fully characterized by the values of its parameters λ.

The concept of universal background models (UBM) was first introduced for the problem of speaker verification [18]. In general, a UBM represents a Gaussian mixture model, which is trained on some generic training data (usually all available training samples). The parameters of the UBM, i.e., λ_{UBM}, are estimated based on the maximum likelihood (ML) criterion via the expectation-maximization (EM) algorithm [19]. The model is typically initialized using either k-means clustering or the Linde-Buzo-Gray algorithm.

Once the UBM is computed, the maximum a posteriori (MAP) estimation criterion (as described in [19]) is used to adapt the UBM to an utterance-specific GMM. The means of the utterance-specific GMM represent a new feature vector. While the GMM for a given test utterance could also be calculated directly from the set of feature vectors extracted from the utterance, adapting the UBM to the data in the given utterance has three important advantages:

- it ensures that the ordering of the GMM parameters in λ is the same as in the UBM for each computed GMM;
- it compensates for the insufficient amount of data in the given utterance; and
- it incorporates domain specific knowledge into the computed GMM.

When computing a GMM from the UBM, the first step is to determine the probabilistic alignment of a particular sample $Pr(i|\boldsymbol{x}_j)$ against all M UBM components as follows:

$$Pr(i|\boldsymbol{x}_j) = \frac{w_i p_i(\boldsymbol{x}_j)}{\sum_{j=1}^{M} w_j p_j(\boldsymbol{x}_j)}, \quad (13)$$

where $p_i(\boldsymbol{x}_j)$ again denotes the Gaussian probability density function of the feature vector $\boldsymbol{x}_j$ for the i-th component of the GMM, j denotes the feature vector index with j = 1, 2, ..., N, N stands for the total number of feature vectors extracted from the given image, and w_i represents the weight associated with the i-th GMM component.

In the second step, the sufficient statistics for updating the mean feature vectors are computed. In general, the MAP estimation procedure updates the means, variances and weights of the GMM, but commonly the focus is only on updating the GMM's means. The statistics required for the MAP adaptation are:

$$n_i = \sum_{t=1}^{T} Pr(i|\boldsymbol{x}_t), \quad (14)$$

and

$$E_i(x) = \frac{1}{n_i} \Sigma \sum_{t=1}^{T} Pr(i \mid x_t)\, x_t, \quad (15)$$

where n_i and E_i stand for the null and first order sufficient statistics.

So far the presented adaptation procedure is identical to the Expectation step when using the ML criterion in the EM algorithm. The difference to the ML-based procedure is shown in the Maximization step, where the update rule becomes:

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\boldsymbol{x}) + (1 - \alpha_i^m)\boldsymbol{\mu}_i, \quad (16)$$

as postulated by the MAP criterion.

The adaptation parameter α_i^m, which controls the balance between the old values of the means and the new estimate is computed as:

$$\alpha_i^m = \frac{n_i}{n_i + \tau}, \quad (17)$$

where τ is the relevance factor, which is the same for all components of the GMM. The value of the relevance factor is chosen experimentally and usually falls in the interval between 8 and 16.

After sufficient iterations of the described procedure, the algorithm stops, if the change in the component means is sufficiently small or a predefined number of iterations is reached. The size of this final GMM vector equals the dimension of the original feature vector multiplied by the number of components of the GMM and, thus, increases with the increase of GMM components, i.e., *M*.

The result of the described procedure is a UBM model and a separate super-vector (comprised of the mean vectors of the *M* GMM components) for each available utterance. The super-vector of means is taken as a feature representing the utterance and typically Support Vector Machines (SVM) are used for classification.

UBM-MAP derived super-vectors for emotion recognition

As already emphasized in the previous section, the use of Universal Background Models (UBM) in combination with the maximum a posteriori (MAP) adaptation criterion was initially introduced to the field of speaker recognition by Reynolds et. al [19], but has since been successfully applied for the recognition of other paralinguistic information in speech as well [10, 20].

Clearly, if the information not related to the task at hand (in our case emotion recognition) could be excluded from the recognition procedure, it would improve the performance of the recognition task. Due to the limited amount of data (per speaker) available in most corpora commonly used in the field of emotion recognition, there is not enough statistical information for decoupling the speaker-specific information. We, therefore, take a different approach and exploit the possibility of excluding gender-specific information. As we show in Section 3, MAP derived models reliably distinguish between genders, empowering the system to take this information into account when making predictions about the emotional class of test utterances.

The illustrated procedure can be more thoroughly described as follows. During training, a single UBM is build using all of the available training data. This UBM is then adapted via the MAP criterion to produce two gender-specific UBMs. Note that in practice only the mean vectors of all Gaussian mixtures are adapted, while the covariance matrices and weights of the initial UBM are left unchanged. Once the two gender-specific UBMs are computed, the training utterances are partitioned into two disjoint sets in accordance with their gender labels (which for the training data are known in advance).

Next, a super-vector comprised of the mean vector of the estimated Gaussian mixtures is constructed for each training utterance by transforming the appropriate gender-specific UBM via the MAP rule. Using this procedure, we arrive at two sets of super-vectors, one for males and one for females, with each super-vector corresponding to a given emotional class.

In the final stage of the training procedure, a *pairwise* SVM classification scheme is trained based on the constructed super-vectors to discriminate between the different emotional-classes.

While the gender labels of the training utterances are known in advance, this is not the case for the test utterances. Hence, to be able to exploit gender-specific information for the emotion recognition task, the gender of the speaker a given test utterance belongs to has to be determined first. This can be done efficiently by a likelihood comparison against the male and female UBMs. Once the gender is known, a super-vector is constructed for the given test utterance by MAP adaptation of the predicted gender's UBM and concatenation of the means of the Gaussian mixtures. The resulting super-vector is ultimately classified into an emotional-class using the trained SVM classification scheme.

2.4 Information fusion

To combine the information from the two sub-systems presented in Sections 2.2 and 2.3 we assess two fusion schemes in this paper. The first is a weighted sum-rule, while the second is a weighted product-rule [21].

Figure 5. Sample frames extracted from video sequences of a random subject from the eNTERFACE'05 database depicting (from top to bottom): anger, disgust, fear, happiness, sadness and surprise. Note that the presented frames are not sampled in equal intervals from the video sequences and that they are processed by the face detection module of our system.

Assume that for a given test recording and a given emotional class our video sub-system has produced a matching score δ_v and, similarly, that our audio sub-system has produced a matching score of δ_a for the same recording and the same emotional class. Then the weighted sum-rule generates a new matching score δ_{sum} based on the following expression:

$$\delta_{sum} = \vartheta\delta_v + (1-\vartheta)\delta_a, \tag{18}$$

where $\vartheta \in [0,1]$ is a weighting factor that needs to be set based on some training/development data.

Similarly, the weighted product-rule produces a matching score of:

$$\delta_{prod} = \delta_v^{\vartheta}\delta_a^{1-\vartheta}, \tag{19}$$

where $\vartheta \in [0,1]$ is again a weighting factor that needs to be set in advance.

Since two different classification techniques are used for the video and audio modality, the matching scores δ_v and δ_a need to be normalized to balance their impact. Towards this end, we use rank normalization on the matching scores prior to the fusion process [22].

3. Experiments

3.1 Database and experimental protocol

For the experiments presented in the remainder of this section, we adopt the publicly available eNTERFACE'05 [11] corpus. The corpus contains recordings of 44 subjects of 14 different nationalities, uttering 5 sentences per each of the 6 emotional classes. These 6 classes correspond to the 'big six' archetypal emotions, as proposed by Ekman in [23]. They are also adopted for the MPEG-4 standard and represent anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA) and surprise (SU).Some frames (after the face detection step) extracted from video sequences of all six emotional classes of a random subject from the eNTERFACE'05 database are shown in Figure 5.

In our experiments 43 subjects are used, subject 6 is omitted as only one recording exists for each emotion. Furthermore, only 2 sentences, portraying happiness, can be found in the database for subject 23. Hence, the total number of utterances available for our experiments sum up to 1287. Since we experimented with the exclusion of gender information from our emotion recognition task, we annotated all data with gender labels prior to the experiments.

For a robust estimate of the recognition performance of the proposed system a 5 fold cross validation protocol is employed (1030 samples were used for training and 257 for testing). The folds are randomly selected without the attention to the distribution of speakers. The evaluation measure for all tests is unweighted (UW) class-wise recall (averaged over 5 folds), which is the predominant way of measuring emotion recognition accuracy [9]. We also report weighted average (WA) recalls of our experiments, even though these are not as reliable, since emotional databases tend to have miss-balanced emotional classes.

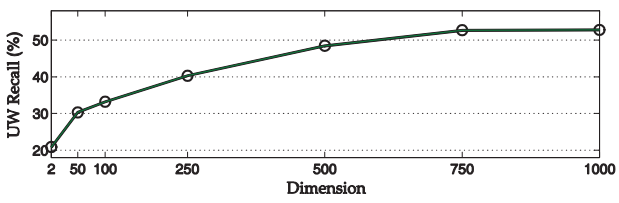

Figure 6. Video recognition results in the form of average unweighted recalls for different subspace dimensionalities

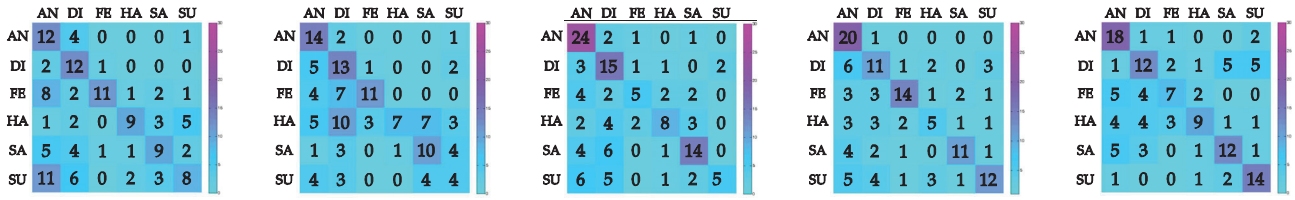

Figure 7. Confusion matrices for all 5 folds of our cross validation procedure generated using the presented video sub-system.

3.2 Results

Assessing the video sub-system

Our first series of experiments aims at assessing the recognition performance of the proposed video sub-system. Specifically, we are interested in the recognition results obtained with respect to the dimensionality of the linear subspace, which is denoted with d' in Section 2.2. Hence, we vary the dimensionality of the subspaces from $d' = 2$ to $d' = 1000$ and observe the average unweighted recall of the experiments (see Figure 6).

Note that with the increase in the dimensionality the recognition performance steadily improves, of course, at the expense of computational complexity. Note that the recognition performance is improved only by a little when the dimensionality of the subspace is increased from $d' = 750$ to $d' = 1000$. Thus, we select a dimensionality of $d' = 750$ for our subspace and use this value for our following experiments. Figure 7 shows more detailed results for this subspace dimensionality, as confusion matrices for all 5 folds of our cross validation procedure are presented there.

Assessing the audio sub-system

The second series of experiments evaluates the performance of the audio sub-system. Throughout all presented experiments different numbers of GMM components were assessed. Generally, the recognition performance increases with the increase in the number of Gaussian mixtures, but with a limited amount of data, one can quickly either over-train the models, or singularities can occur during covariance calculations.

Gender recognition	Number of GMM components				
	8	16	32	64	128
UW Recall	94.2	96.3	97.1	98.2	98.4

Table 1. Gender recognition results for the audio sub-system

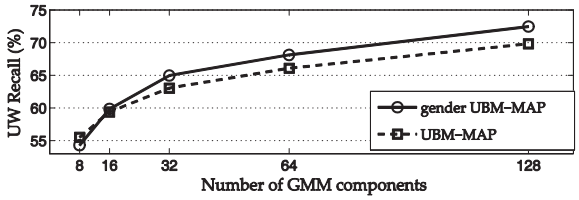

Figure 8. Emotion recognition based on audio data results

In Section 2.3. we stated that gender-specific UBMs can be used via likelihood calculations to recognize gender. Table 1 presents the gender recognition results based on likelihood calculations against male and female UBMs with respect to GMM complexity. As expected, the recognition rate increases with the number of GMM components. Even with 8 components the results are above 94%.

While for the gender recognition task, a simple likelihood comparison is sufficient to obtain "good" recognition results, the emotion recognition experiments require the use of more advanced approaches. Thus, following the gender detection step, an utterance specific vector of means is produced based on the MAP criterion and the UBM of the predicted gender. This super-vector is finally subjected to our SVM classification scheme.

As shown in Figure 9, the proposed gender-specific UBM-MAP method outperforms the standard UBM-MAP approach, except in the case of 8 GMM components where higher gender detection errors cause a slight decrease in emotion recognition performance. With the increase of the number of GMM components the emotion recognition performance increases for both systems, but since the gender predictions become more accurate, the efficiency of our procedure becomes more evident.

Similar as for the video sub-system, we also present detailed results of our experiments with the audio sub-system in the form of confusion matrices for all 5 folds of our cross validation procedure. The matrices are presented in Figure 9.

Audio-Video Fusion

Our third series of experiments assessed the performance of the combined multi-modal system with different fusion techniques. In order to train the fusion parameters (i.e., the weighting factors ϑ), the test samples are randomly split into two parts. The first half is used for the estimation of the fusion parameters, and the second half for evaluation. The scores, produced during classification from both modalities, are combined in order to give the final prediction for each test utterance. The results of the weighted sum and weighted product fusion are presented in the lower part of Table 2.

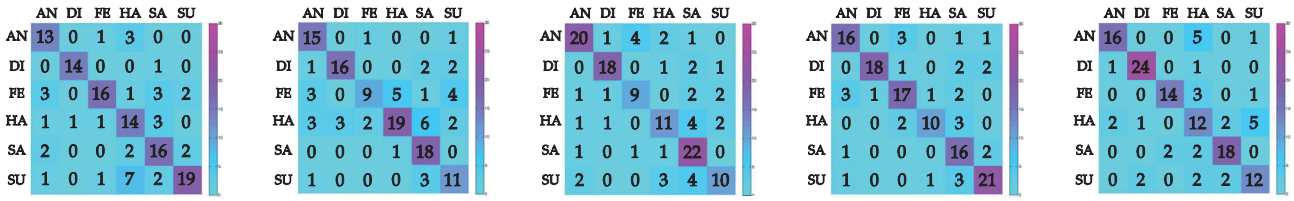

Figure 9. Confusion matrices for all 5 folds of our cross validation procedure generated using the presented audio sub-system

	System description		Emotion recognition	
	Description	# feat	UW	WA
Audio	MFCCs & HMMs [14]	13	55.9	/
	MFCC & GMMs [7]	13	67.1	67.0
	Supra-segmental modeling [7]	56	72.5	72.4
	UBM-MAP (ours)	13	69.6	69.7
	Gender UBM-MAP (ours)	*13*	*72.5*	*72.3*
Video	SAMMI framework [12, 24]		28.0	/
	Video sub-system [13]		37.0	/
	LBPs+HMMs [14]		37.7	/
	Canonical correlations (ours)		*52.8*	*52.2*
Fusion	Audio Video HMM [14]		56.3	/
	SAMMI framework [12, 24]		67.0	/
	Async. feature fusion [13]		71.0	/
	Sum rule fusion (ours)		*75.9*	*75.7*
	Product rule fusion (ours)		*77.5*	*77.2*

Table 2. Comparison of emotion recognition results (in %)

The differences between the weighted sum-rule and weighted product-rule are minor, with the highest UW recall of 77.5% achieved by the weighted product-rule fusion procedure.

Comparison with the state-of-the-art

Last but not least, we compared the performance of both developed sub-systems as well as the multi-modal emotion recognition system as a whole to results published in the literature. The results of this comparison are shown in Table 2.

Since there is no strictly defined protocol for the eNTERFACE'05 corpus, different experimental setups were used with the cited results, thus, a strict comparison is not possible. Note, however that our experimental protocol was as least as challenging as any from the cited sources. It is evident that our results at least match the highest reported results from the literature. Furthermore, our results are obtained without incorporating any prosodic or voice quality features, which could further improve the results.

4. Conclusion

In the paper we presented a multi-modal emotion recognition system. Both, audio and video sub-systems were implemented using novel approaches. For the audio sub-system we have shown that the standard UBM-MAP procedure can be further improved by incorporating gender-specific information. For the video sub-system we presented an approach to emotion recognition based on image-set matching. Both sub-systems were evaluated individually, resulting in competitive performance, when compared to the state-of-the-art results from the literature. The fusion of both sub-systems resulted in an additional increase in the emotion recognition performance when compared to the results obtained with the uni-modal systems. Moreover, the achieved average unweighted recall of 77.5% on the eNTERFACE'05 corpus also compares favorably with other techniques from the literature.

For our future work with respect to multi-modal emotion recognition we plan to evaluate other possibilities to exclude non-emotion related information from the audio signals. For the video sub-system we plan to assess different, possibly non-linear, options for image-set matching, such as kernel canonical correlation analysis [25, 26] or related techniques.

5. Acknowledgments

The work presented in this paper was supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the postdoctoral project BAMBI (ARRS ID Z2-4214),and by funding from the European Union's Seventh Framework Programme (FP7-SEC-2010-1) under grant agreement number 261727. The authors also acknowledge the support of the COST Action IC1106 Qualinet.

6. References

[1] R. Gajšek, V. Štruc and F. Mihelič, "Multi-modal emotion recognition using canonical correlations and acoustic features", *Proceeding of ICPR 2010*, Istanbul, Turky, pp. 4133-4136, 2010.

[2] R. Gajšek, V. Štruc and F. Mihelič, "Multimodal emotion recognition based on the decoupling of emotion and speaker recognition", *Proceeding of TSD 2010*, Springer, LNCS vol. 6231, pp. 275-282, 2010.

[3] N. Sebe, I. Cohen, T. Gevers and T.S. Huang, " Multimodal approaches for emotion recognition : A survey", *Proc. of SPIE*, pp. 56-67, 2005.

[4] T.K. Kim, J. Kittler and R. Cipolla, "Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations". *IEEE Transactions on PAMI*, vol. 29, no. 6, pp. 1005-1018, 2007.

[5] O. Yamaguchi, K. Fukui and K.I. Maeda, "Face Recognition using Temporal Image Sequence", *Proc of AFGR*, pp. 318-323, 1998.

[6] P. Viola and M. Jones, "Robust Real-Time Face Detection", *Int. J. Comp. Vis.*, vol. 57, no. 2, pp. 137-154, 2004.

[7] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances", *Proc. of ASRU*, 2009.

[8] C. Busso, S. Lee and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 4, pp. 582-596, 2009.

[9] B. Schuller, S. Steidl and A. Batliner, "The Interspeech 2009 Emotion Challenge", *Proc. of Interspeech 2009*, Brighton, UK, pp. 312-315, 2009.

[10] M. Kockmann, L. Burget and J. Cernocky, "Brno University of Technology System for Interspeech 2009 Emotion Challenge", *Proc. of Interspeech 2009*, Brighton, UK, pp. 348--351, 2009.

[11] O. Martin, I. Kotsia, B. Macq and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database", *22nd Int. Conf. Data Engineering Workshops*, pages 8, 2006.

[12] M. Paleari and B. Huet, "Toward emotion indexing of multimedia excerpts", *CBMI*, London, UK, June 2008.

[13] M. Mansoorizadeh and N.M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech", *Multimedia Tools Appl.*, vol. 49, no. 2, pp. 277-297, 2010.

[14] D. Datcu and L. Rothkrantz, "Multimodal recognition of emotions in car environments", *DCI&I 2009*, Prag, Czech Republic, 2009.

[15] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neurosicence*, vol. 3, no. 1, pp. 71-86, 1991.

[16] F. Eyben, M. Wöllmer and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit", *Proc. ACII 2009*, Amsterdam, Netherlands, pp. 576-581, 2009.

[17] B. Vesnicer, J. Žganec-Gros, N. Pavešić and V. Štruc, "Face Recognition using Simplified Probabilistic Linear Discriminant Analysis", *International Journal of Advanced Robotic Systems*, vol. 9, pp. 1-10, 2012.

[18] J. Križaj, V. Štruc and S. Dobrišek, "Towards Robust 3D Face Verification using Gaussian Mixture Models", *International Journal of Advanced Robotic Systems*, vol. 9, pp. 1-10, 2012.

[19] D.A. Reynolds, T. F. Quatieri and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-40, 2000.

[20] T. Bocklet, A. Maier, J.G. Bauer, F. Burkhardt and E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines", *Proc. of ICASSP*, pp. 1605-1608, 2008.

[21] J. Kittler, "Combining classifiers: A theoretical framework ", *Pattern Analysis and Applications*, vol. 1, pp. 18-27, 1998.

[22] A. Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems", *Pattern Recognition*, vol. 38, pp. 2270-2285, 2005.

[23] P. Eckman, "Facial expression and Emotion", *American Psychologist*, vol. 48, pp. 384-392, 1993.

[24] M. Paleari, R. Benmokhtar and B. Huet, "Evidence Theory-Based Multimodal Emotion Recognition", *MMM '09: Proc. of the 15th Inter. Multimedia Modeling Conf. on Advances in Multimedia Modeling*, Springer, Berlin Heidelberg, pp. 435-446, 2008.

[25] S.Y. Huang, H.H. Lee and C.K. Hsiao, "Kernel Canonical Correlation Analysis and its Application to Nonlinear Measure of Association and Test of Independence", 2006.

[26] T. Melzer, M. Reiter, H. Bischof, "Appearance models based on kernel canonical correlation analysis", *Pattern Recognition*, vol. 36, no. 9, pp. 1961-1971, 2003.