# Multimodal recognition of emotions in car environments

Dragoş Datcu[A] and Léon J.M. Rothkrantz[B]

***Abstract* — Within the last couple of years, automatic multimodal recognition of human emotions has gained a considerable interest from the research community. By taking into account more sources of information, the multimodal approaches allow for more reliable estimation of the human emotions. They increase the confidence of the results and decrease the level of ambiguity with respect to the emotions among the separate communication channels. This paper provides a thorough description of a bimodal emotion recognition system that uses face and speech analysis. Basically, we use hidden Markov models - HMMs to learn and to describe the temporal dynamics of the emotion clues in the visual and acoustic channels.**

**Key Words — emotion recognition, facial expression analysis, speech analysis**

## 1. Introduction

The complexity of the emotion recognition using multiple modalities is higher than the complexity of the unimodal methods. Some causes for that relate

to the asynchronous character of the emotion patterns and the ambiguity and the correlation which possibly occur in the different informational channels.

For instance, speaking while expressing emotions implies that the mouth shape corresponds to a mix of the influence of both the pronounced phoneme and the internal emotional state. In this case, the use of the regular algorithms we have used so far for facial expression recognition [3] show limited performance and reliability. In order to apply fusion, the model differentiates the silence video segments and the segments that show the subject speaking. Basically, we use hidden Markov models - HMMs to learn and to describe the temporal dynamics of the emotion clues in the visual and acoustic channels. This approach provides a powerful method enabling to fuse the data we extract from separate modalities.

The models we build in this paper run emotion analysis on the data segments which embody activity in both visual and audio channels. In the following section, we present the algorithms and the results achieved in some recent and relevant research works in the field of multimodal emotion recognition. Then, we describe our new system that may be used in car environments (figure 1).

We present the details of all steps involved in the analysis, from the preparation of the multimodal database and the feature extraction to the classification of six prototypic emotions. Apart from working with unimodal recognizers, we conduct experiments on both early fusion and decision level fusion of visual and audio features.

The novelty of our approach consists of the dynamic modelling of emotions using hidden Markov models (HMMs) in combination with Local Binary Patterns (LBPs) [17] as visual features and mel-frequency cepstral coefficients (MFCCs) as audio features. In the same time, we propose a new method for visual feature selection based on the multi-class Adaboost.M2 classifier. A cross database method is employed to identify the set of most relevant features from a unimodal database and to proceed with applying it in the context of the multimodal setup.

We report on the results we have achieved so far for the discussed models. The last part of the paper relates to conclusions and discussions on the possible ways to continue the research on the topic of multimodal emotion recognition.

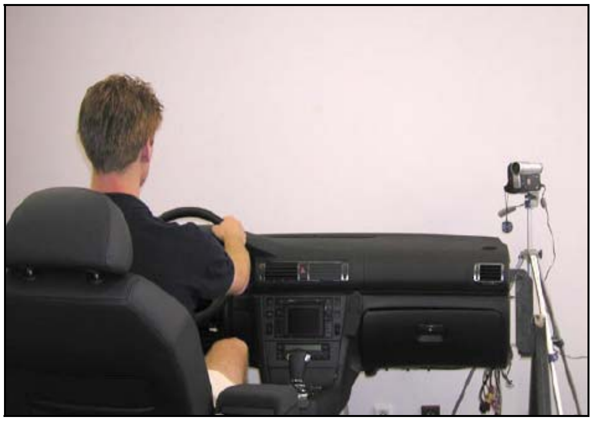


**Fig. 1: Experimental setup for multimodal emotion recognition in car environment**

## 2. Related work

A noticeable approach for approaching the recognition of emotion represents the multimodal analysis. The multimodal integration of speech and face analysis can be done by taking into account features at different levels of abstraction.

Depending on that, the integration takes the form of fusion at the low, intermediate or high levels. The low-level fusion is also called early fusion or fusion at the signal level and the high-level fusion is also called semantic, late

**A –** Dr. ir. Dragoş Datcu, e-mail: d.datcu@tudelft.nl
**B –** Prof. drs. dr. Léon Rothkrantz, e-mail: l.j.m.rothkrantz@tudelft.nl
SeWaCo, Netherlands Defence Academy/Department of
Mediamatica, Delft University of Technology, The Netherlands

fusion or fusion at the decision level. Several researchers have recently tackled these types of integration.

Han et al. [10] propose a method for bimodal recognition of four emotion categories plus the neutral state, based on hierarchical SVM classifiers. Binary SVM classifiers make use of fusion of low-level features to determine the dominant modality that, in turn, leads to the estimation of emotion labels. The video processing implies the use of skin colour segmentation for face detection and optical density and edge detection for face feature localization. The algorithm extracts twelve geometrical features based on the location of specific key points on the face area. In case of speech analysis, twelve feature values are computed using the contours of pitch and the energy from the audio signal.

On a database of 140 video instances, the authors report an improvement of 5% compared to the performance of the facial expression recognition and an improvement of 13%, compared to the result of the emotion recognition from speech.

Wimmer et al. [24] study early feature fusion models based on statistically analysing multivariate time-series for combining the processing of video based and audio based low-level descriptors (LLDs). Paleari and Huet [18] research the multimodal recognition of emotions on Enterface 2005 database. They use mel-frequency cepstral coefficients - MFCC and linear predictive coding – LPC for emotion recognition and optical flow for facial expression recognition together with support vector machines and neural network classifiers. The recognition rate of emotion classification is less than 35% for speech-oriented analysis and less than 30% in case of face-oriented analysis. Though, combining the two modalities leads to an improvement of 5% in case of fusion at the decision level and to almost 40% recognition rate in case of early fusion.

Another study on the Enterface 2005 database is presented by Mansoorizadeh and Charkari [14]. They apply principal components analysis - PCA to reduce the size of the audio and visual feature vectors and binary support vector machines - SVM for the bimodal person-dependent classification of basic emotions.

Depending on the type of fusion, the inputs of the SVM models contain either separate or combined audio-visual feature vectors. For speech analysis, the features relate to the energy, the pitch contour, the first 4 formants, their bandwidth, and 12 MFCC components of the audio signal. For face analysis, the features represent geometric features that are computed based on a set of specific key points on the face area. The likelihood results of the binary SVM classifiers are used in a rule based system to determine the emotion labels of the video instances. The authors report the 53% the classification rate for emotion recognition from speech, 36.00% for facial expression recognition, 52.00% for feature level fusion and 57.00% for decision level fusion.

The work of Hoch et al. (Hoch et al. 2005) presents an algorithm for bimodal emotion recognition in automotive environment. The fusion of results from unimodal acoustic and visual emotion recognizers is realized at abstract decision level. For the analysis, the authors used a database of 840 audiovisual samples that contain recordings from seven different speakers showing three emotions. By using a fusion model based on a weighted linear combination, the performance gain becomes nearly 4% compared to the results in the case of unimodal emotion recognition.

Song et al. [21] present emotion recognition based on Active Appearance Models AAM for facial feature tracking. The Facial Animation Parameters – FAPs are extracted from video data and are used together with low level audio features as input for a HMM to classify the human emotions. Paleari and Lisetti [19] present a multimodal fusion framework for emotion recognition that relies on MAUI - Multimodal Affective User Interface paradigm. The approach is based on the Scherers theory Component Process Theory (CPT) for the definition of the user model and to simulate the agent emotion generation.

Sebe et al. [20] propose a Bayesian network topology for recognizing emotions from audio and facial expressions. The database they used includes recordings of 38 subjects who show 11 classes of affects. According to the authors, the achieved performance results pointed to around 90% for bimodal classification of emotions from speech and facial expressions compared to 56% for the faceonly classifier and about 45% for the prosody-only classifier. Zeng et al. [25] conducted a series of experiments related to the multimodal recognition of spontaneous emotions in a realistic setup for Adult Attachment Interview. They use Facial Action Coding System - FACS [6] to label the emotion samples. Their bimodal fusion model combines facial texture and prosody in a framework of Adaboost multi-stream hidden Markov model (AdaMHMM). Joo et al. [12] investigate the use of S-type membership functions for creating bimodal fusion models for the recognition of five emotions from speech signal and facial expressions. The achieved recognition rate of the fusion model was 70.4% whereas the performance of the audio-based analysis was 63% and the performance of the face-based analysis was 53.4%. Caridakis et al. [2] describe a multi-cue, dynamic approach in naturalistic video sequences using recurrent neural networks. The approach differs from the existing works at the time, in the way that the expression of the user is modeled using a dimensional representation of activation and valence instead of the prototypic emotions. The facial expressions are modelled in terms of geometric features from MPEG-4 facial animation parameters - FAPs, and are computed using the location of 19 key points on the face image. Combining FAPs and audio features related to pitch and rhythm leads to the multimodal recognition rate of 79%, as opposed to facial expression recognition rate of 67% and emotion from speech detection rate of 73%.

The work of Meng et al. [16] presents a speech-emotion recognizer that works in combination with an automatic speech recognition system. The algorithm uses Hidden Markov Model HMM as a classifier. The features considered for the experiments consisted of 39 MFCCs plus pitch, intensity and 3 formants, including some of their statistical derivatives. A emotion recognition study on a language independent database has been done in [23]. The authors extract MFCC and formant frequency features from the speech signal and Gabor wavelet features from the face images. The classification of six emotions uses neural networks and Fisher's linear discriminant analysis - FLDA. The results indicate the higher efficiency of using the audio signal with 66.43% recognition rate over the visual processing with 49.29% recognition rate. The audio-visual fusion has classification rate of 70%. Busso et al. [1] explore the properties of both unimodal and multimodal systems for emotion recognition in case of four emotion classes. In this study, the multimodal fusion is realized separately at the semantic level and at the feature level. The overall performance of the classifier based on feature level fusion is 89.1% which is close to the performance of the semantic fusion based classifier when the product-combining criterion is used. Go et al. [9] uses Z-type membership functions to compute the membership degree of each of the six emotions based on the facial expression and the speech data. The facial expression recognition algorithm uses multi-resolution analysis based on discrete wavelets. An initial gender classification is done by the pitch of the speech signal criterion.

The authors report final emotion recognition results of 95% in case of male and 98.3% for female subjects. Fellenz et al. [7] uses a hybrid classification procedure organized in a two-stages architecture to select and fuse the features extracted from face and speech to perform the recognition of emotions. In the first stage, a multi-layered perceptron (MLP) is trained with the back propagation of error procedure. The second symbolic stage involves the use of PAC learning paradigm for Boolean functions.

## 3. Dataset preparation

The models we are going to build for multimodal emotion recognition are based on the use of hidden Markov model - HMM classifier. In the current research context, HMM is used as a supervised machine learning technique. Based on that, the HMM training and testing processes rely on the use of fully labeled samples of audio-visual data instances. At the moment of starting this research, finding a fully annotated database turned to be difficult to fulfil. This was first because of the lack of multimodal databases. Some databases had no emotion labels and were not proper for audio-visual processing. We specifically avoided using multimodal data sets that have recordings with noise and utterance overlapping in the audio signal, or with occlusion and too much rotation of the subjects' face.

The database we have eventually decided to use for our research is Enterface 2005 [15]. This database contains audio-visual recordings of 42 subjects who represent 14 different nationalities. A percentage of 81% are men, while the remaining 19% are women. At the recording time, 31% of the subjects wore glasses and 17% had beard. The recording procedure first consisted of listening to six successive short stories, each of them eliciting a particular emotion. The emotions relate to the prototypic emotions which are: happiness, sadness, surprise, anger, disgust and fear, as identified by Ekman [6]. Then, the subjects had to read, memorize and finally utter five different reactions to each story, all by using English language. For each story, the subjects were asked to produce messages that contain only the emotion to be elicited and to show as much expressiveness as possible. The recording setup implied the use of a monochromatic dark grey panel for the image background and constant illumination. The audio-visual data was encoded using Microsoft AVI format. The image frames were stored using the image resolution of 720x576 pixels, at the frame rate of 25 frames per second. The audio samples were stored using uncompressed stereo 16-bit format at the sample rate of 48000 Hz. We have started the data pre-processing step from the set of 1293 samples from Enterface 2005 database. In the context of multimodal processing, we had to first verify the appropriateness of each video sample. As a result, we have removed a subset of 463 instances.

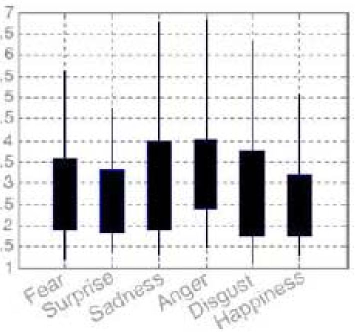


**Fig. 2: Utterance duration (in seconds) for each emotion class**

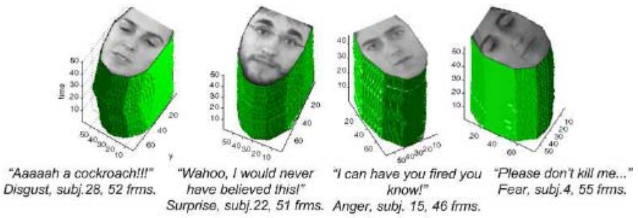

time
"Aaaaah a cockroach!!!"
Disgust, subj.28, 52 frms.
"Wahoo, I would never have believed this!"
Surprise, subj.22, 51 frms.
"I can have you fired you know!"
Anger, subj. 15, 46 frms.
"Please don't kill me..."
Fear, subj.4, 55 frms.

**Fig. 3: 60×80 video samples containing the face area only**

From the set of 830 remaining samples, 135 accounted for emotion class fear, 143 for surprise, 137 for sadness, 145 for anger, 141 for disgust and 129 for happiness. This subset represents a well-balanced multimodal database of simulated emotion recordings from 30 subjects. Figure 2 illustrates the duration in seconds, of the utterances from the final multimodal database. Like in the case of unimodal vision oriented methods for extracting and normalizing the actual face images from each video frame. At first, we used Viola&Jones face detection algorithm [22] and Active Appearance Models [5] to obtain the location and the shape of the faces. Then, we have removed the unnecessary image patches and scaled down the face images to 60 pixels width by 80 pixels height. Here, unnecessary image patches relate to the visible parts of background, subject's hair and cloth. For aligning the faces, we used the reference key point located at the middle of the line segment delimited by the inner corners of the eyes. Figure 3 illustrates the result of applying the previously described methods on four video samples containing faces.

## 3.1 Emotion estimation from speech

The assessment of the emotion levels from speech can be naturally done by identifying patterns in the audio data and by using them in a classification setup. The features we extract are the energy component and 12 mel-frequency cepstral coefficients together with their delta and the acceleration terms from 25 ms audio frames, with 10 ms frame periodicity from a filter bank of 26 channels.

A Hamming window is used on each audio frame during the application of Fourier transform. The feature extraction procedure determines the conversion of the original audio sampling rate of 48kHz to the MFCC frame rate of 100Hz. Each MFCC frame contains 39 terms, as indicated previously. The recognition of emotions is realized using the HMM algorithm. Each emotion has associated one distinct HMM and the set of HMMs forms a multi-class classifier. For evaluation, we use 3-fold cross validation. The samples from the same subject are part of either the training set or the test set. This restriction assumes that the testing is done on instances of subjects other than those of the subjects included in the training data set.

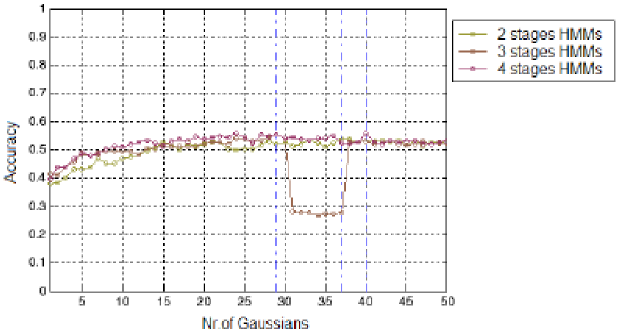


**Fig. 4: The accuracy of HMM-based classifiers of emotions in speech signal. The number of states is 2, 3, 4 and the number of Gaussians varies from 1 to 50. Three fold cross validation method is used for performance estimation**

The method is supposed to give a better estimation of the performance of the classifiers. For finding the best HMM model, we conduct experiments in which we investigates the optimal values for the HMM parameters. In this way, we build and test models which use 2,3 and 4 HMM states (figure 4). The 2 state HMMs encode the emotion onset and offset. The 3 state HMMs encode the emotion onset, apex and offset. The models with 4 states encode the neutral state and emotion onset, apex and offset states. For each state configuration, we build distinct models of HMMs with Gaussian mixtures with different number of components (1..50 components).

The results of testing all the models, are illustrated in figure 4. Following the evaluation, it results that the most efficient configuration is to use 4 states and 40 Gaussians per mixture and that the accuracy of this classifier is 55.90%. Table 1 presents the confusion matrix of this classification model.

**Tab. 1: The confusion matrix of the HMM that has 4 states and 40 Gaussian components; the accuracy of the emotion recognition from speech model is 55.90% for six basic emotion categories**

| | Fear | Surprise | Sadness | Anger | Disgust | Happy |
|---|---|---|---|---|---|---|
| Anger | **91.72** | 2.07 | 0.69 | 1.38 | 2.76 | 1.38 |
| Disgust | 24.11 | **44.68** | 9.22 | 11.35 | 4.26 | 6.38 |
| Fear | 25.19 | 14.81 | **41.48** | 6.67 | 5.19 | 6.67 |
| Happy | 19.38 | 18.60 | 3.88 | **48.06** | 6.98 | 3.10 |
| Surprise | 23.78 | 9.09 | 8.39 | 8.39 | **38.46** | 11.89 |
| Sadness | 5.84 | 5.84 | 10.22 | 2.19 | 6.57 | **69.34** |

## 3.2 Video analysis

The goal of the video analysis is to build models enabling these to dynamically process the video data and to generate labels according to the six basic emotion classes. The input data is represented by video sequences that can have different number of frames, as determined by the utterance-based segmentation method.

The limits of each video data segment are identified by using the information obtained during the analysis of the audio signal. Based on the set of frames, a feature extraction is applied for preparing the input to the actual classifier.

HMM models are then employed to classify the input sequence in terms of the emotion classes.

One problem that has to be taken into account while developing the facial expression recognizers is that both the input set of features and the classifier models should be chosen in such a way so as to be able to handle the time dependent variability of the face appearance.

More specifically, some of the inner dynamics of the face are generated due to the effect of the speech process that is present in the data. Taking into account the aforementioned issues, the focus of the research is to study the selection of most relevant visual features and to use the values of these features as data observations for the HMM-based classifiers.

Adaptive boosting - Adaboost is a binary boosting method proposed by Freund and Schapire [8], which have very high generalization performance. The algorithm works by assigning and iteratively updating the weights on training data instances in a dynamic manner, according to the errors at the previous learning step. Misclassified data get higher weights, leading the learning process to focus more on the hardest examples. The algorithm is a type of large margin classifiers which minimizes an exponential function of the margin over the training set. An interesting aspect of Adaboost is the capacity to identify outliers which are defined as mislabelled, ambiguous or hard-to-classify training instances.

For the feature selection, we conducted a separate research using a second data set namely the Cohn-Kanade database [13]. In this context, the multi-class classification method Adaboost.M2 is used as a feature selection algorithm. The procedure is based on the primary property of the Adaboost.M2 to identify the most important features while running the training phase of the classification process. We use the same set of prototypic emotions as for the main study on the Enterface05 dataset.

The first problem is to make a proper data set of representative face image samples. The basic set of non-ambiguous facial expression samples from the Cohn-Kanade database include 251 instances. Each instance corresponds to the last frame of the video sequence and represents the face at the apex of one facial expression. Subsequently, we changed the structure of the database so as to reflect balanced classes of emotions (table 2).

**Tab. 2: The structure of the balanced set of 303 samples selected from the Cohn-Kanade database**

| | Nr. Non-ambiguous samples | | | | | | | Nr. Mixed emotion samples | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Distance** | **0** | **1** | **2** | **3** | **4** | **5** | total | **0** | **1** | **2** | **3** | **4** | **5** | total |
| Fear | 2 | 10 | 17 | 11 | 7 | 3 | **50** | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| Surprise | 0 | 24 | 4 | 0 | 0 | 0 | **28** | 0 | 10 | 12 | 0 | 0 | 0 | **22** |
| Sadness | 0 | 5 | 24 | 14 | 7 | 0 | **50** | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| Anger | 0 | 4 | 12 | 12 | 18 | 7 | **53** | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| Disgust | 0 | 0 | 24 | 4 | 0 | 0 | **28** | 0 | 4 | 18 | 0 | 0 | 0 | **22** |
| Happiness | 17 | 21 | 10 | 2 | 0 | 0 | **50** | 0 | 0 | 0 | 0 | 0 | 0 | **0** |

Because the sets of the visual features we derive from the face samples, are too large to be used directly as observations in the HMM classification setup, we have to decrease their size. This can be done by transforming the original visual features to other set of more representative features.

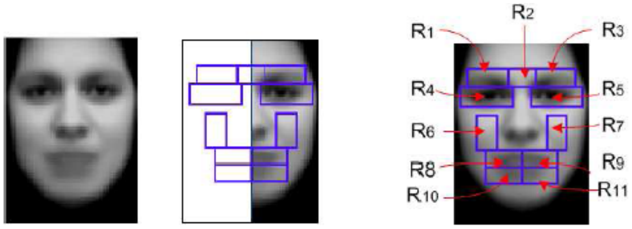
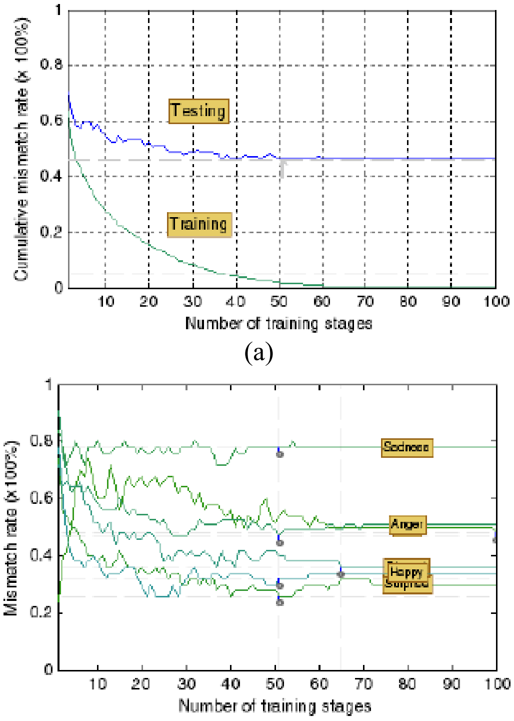


**Fig. 5: Average face sample from the balanced Cohn-Kanade database. A symmetric facial feature model is used to delimit rectangular face regions from which specific visual features are extracted**

The boosting methods represent a specific class of algorithms that can be successfully used to select representative features. We do the feature selection by following the same steps we have made for unimodal facial expression recognition. We use local binary patterns - LBPs and Adaboost.M2 classifier. The result of this part of research will be later applied for the facial expression recognition in video data of speaking subjects.

Previous studies on facial expression recognition in single images, like the one of Dubuisson et al. [4], showed that different face regions produce features with different informative power for classification. In our dynamic recognition setup, we want to also investigate the contribution of speaking mouth region and other face regions to the classification. In addition to using the whole face image, we define two symmetric models of face regions around the face features.



(a)

(b)

**Fig. 6: Train and test mismatch rate of Adaboost.M2 using LBPs from 7 face regions**

Figure 5 illustrates the face regions taken into account. Regions R8, R9, R10 and R11 are located on the mouth area and therefore are considered to be essentially influenced during the production of speech and during expressing emotions. The first face region model consists of using regions R1 ... R7 and the second model consists of using regions R1 ... R11. We generated 27.226 LBP features located on the whole face image, 22.276 LBP features based on the second face region model and 14.176 LBP features based on the first face region model. The Adaboost.M2 classifier was then used to identify the features that provided the best facial expression recognition results. For evaluation we used 20-folds cross validation method. Figure 6 illustrates the train and test mismatch rates of the Adaboost.M2 classifier using 7 face regions, for the six prototypic facial expressions. Table 3 shows the confusion matrix of this classifier. The set of LBP features selected by this classifier, are projected on the average face in figure 7.

**Tab. 3: Confusion matrix of the Adaboost.M2 facial expression classifier using LBP features extracted from 7 face regions**

| | Fear | Surprise | Sadness | Anger | Disgust | Happy |
|---|---|---|---|---|---|---|
| Fear | **46.00** | 4.00 | 26.00 | 8.00 | 2.00 | 14.00 |
| Surprise | 12.00 | **74.00** | 6.00 | 2.00 | 6.00 | 0.00 |
| Sadness | 28.00 | 12.00 | **22.00** | 16.00 | 16.00 | 6.00 |
| Anger | 18.86 | 1.88 | 16.98 | **52.83** | 5.66 | 3.77 |
| Disgust | 8.00 | 8.00 | 10.00 | 14.00 | **60.00** | 4.00 |
| Happy | 6.00 | 0.00 | 8.00 | 4.00 | 14.00 | **68.00** |

## HMM-based facial expression recognition

Making facial expression recognizers with hidden Markov models implies the identification of the optimal model parameters. Finding the best number of states, the best number of Gaussian mixture components and the best set of Local Binary Features - LBPs, represent a non-trivial task. We start from the results of the Adaboost.M2 classifiers. In the case of facial expression recognition using LBP features extracted from 7 face regions, we have found that the optimal number of training stages is 51. At each training stage, Adaboost.M2 selects a subset of six LBP features, one for each facial expression category.

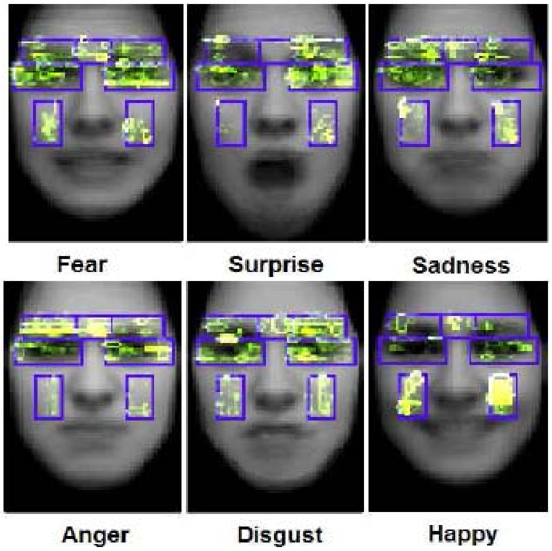
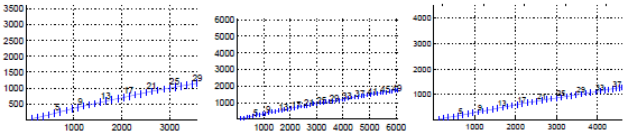


**Fig. 7: Projection of the set of 51 LBPs of the model of 7 face regions, on average face images showing the six basic emotions**

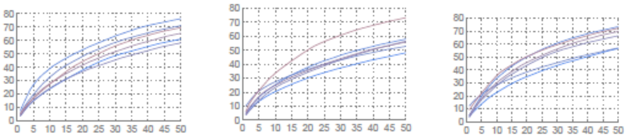
As a consequence, there would be 306 features which account for the six facial expressions of the final optimal classifier. Taking into account the fact that for evaluation we have used cross validation with 20 folders, it results that the final LBP feature set contains 6120 LBP features. However, the set obtained by concatenating all the subsets of the 20 folders of the cross validation method, does not include only distinct features. In fact, an important part of the subset relates to features that are commonly selected by Adaboost.M2 during training multiple folders. In addition, Adaboost.M2 may select the same feature multiple times during the training during the training at the same the cross validation folder. This is depicted graphically in figure 8(b). For example, the set of features collected by taking the first 45 most important LBPs from 7 face regions, for all emotion categories, includes 5400 features, though the same set contains only 1599 distinct LBP features.

We define importance of LBP features based on the number of times an LBP is selected by the optimal Adaboost.M2 classifier during training. Figure 9 illustrates the accumulated percentage of importance for the set of LBPs for the whole face and for the two face region models. In the figure, the LBPs are presented in the descending order. Using the feature importance measure, we make separate data sets by gradually choosing the first most important features for all emotion classes, from the feature sets of the 20 cross validation folders.



(a) Whole face. (b) 7reg. (c) 11reg.

**Fig. 8: The concatenated set of LBP features extracted from the whole face and two types of face regions. The x axis represents the size of the feature set; the y axis represents the number of distinct LBP features selected by Adaboost.M2**



(a) Whole face. (b) 7reg. (c) 11reg.

**Fig. 9: Importance of LBP features extracted from the three face region models. The features are sorted in the descending order of the selections(%) by the Adaboost.M2 classifier, for each basic emotion category**

For evaluation of facial expression recognition, we have generated HMM models for each emotion category. The training data sets have been created by taking into account the emotion label of each video sample. At the testing stage, each video instance is analysed using six HMM models, one for each emotion. Figure 10 shows the performance of different HMM classifiers on the data set of LBPs extracted from the whole face region and on the data sets of LBPs extracted from 7 and 11 face regions. The best facial expression recognition model uses 268 distinct features that corresponds to the selection of 45 features from each facial expression category. The accuracy of this classifier is 37.71%. Table 4 shows the confusion matrix of the HMM classifier.

## 3.3 Fusion model

Considering the previous results of unimodal emotion estimation, it turns out that the use of audio data leads to better recognition rate (55.9%), when compared to the use of facial expressions-oriented models (37.71%). The next step in the attempt to get higher performance for the emotion recognition, is to combine the information from the two unimodal approaches. Depending on the type of information taken into consideration, we can define separate categories of integration. Using combined sets of audio and video features as input for the classification models is considered to fall in the category of low-level data fusion. This approach is also called early fusion or signal level fusion.

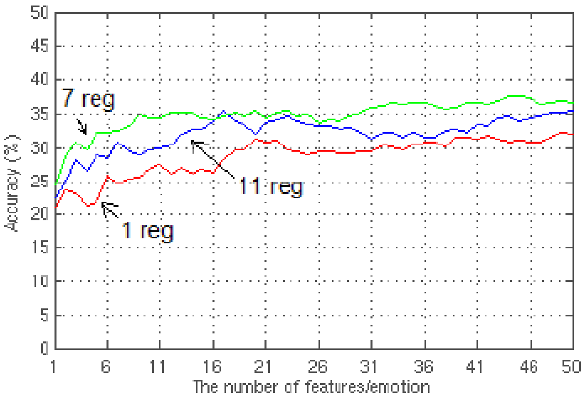


**Fig. 10: Facial expression recognition recognition results by using HMM models with different number of LBP features. The best HMM uses 45 LBP features for each facial expression category, from 7 different face regions**

**Tab. 4: Confusion matrix of the best HMM facial expression classifier using LBP features**

| (%) | Anger | Disgust | Happy | Surprise | Sadness | Fear |
|---|---|---|---|---|---|---|
| Anger | **18.62** | 15.86 | 11.03 | 24.13 | 17.24 | 13.10 |
| Disgust | 9.92 | **60.28** | 10.63 | 63.82 | 6.38 | 6.38 |
| Happy | 6.20 | 17.05 | **48.06** | 18.60 | 5.42 | 4.65 |
| Surprise | 10.48 | 2.79 | 9.09 | **53.14** | 16.08 | 8.39 |
| Sadness | 17.51 | 10.94 | 10.94 | 25.54 | **19.70** | 15.32 |
| Fear | 9.62 | 16.29 | 6.66 | 26.66 | 14.07 | **26.66** |

Conversely, the use of final emotion estimates from unimodal face and speech analysis is defined as high level fusion. This alternative is also called late fusion or fusion at the decision level.

Prior to building models which integrate audio and video data, the first problem that regards the video segmentation must be solved. We identify the beginning and end points of audio-video data chunks based on the turn-based segmentation.

The long pauses in conversation are used as indicators for identifying the edges of a segment. Once the audio-video segments are obtained, we then proceed by removing the sub-segments that denote the lack of speech. Based on the resulting data segments, the distinct sets of audio and video features are further extracted following the same procedures as in the case of unimodal emotion recognition. In case of the audio signal, we extract sequences of MFCC frames at the rate of 100 frames/second, each frame being sized to 39 acoustic features. As result to video processing, we extract visual feature sets at the rate of 25 feature sets/second. Extracting LBP features from the whole face image leads to sets of 331 features/set. Similarly, using LBP features from 7 face regions generates sets of 307 features/set and using LBP features from 11 face regions generates sets of 335 features/set.

Because of the difference between the 100Hz rate of MFCC frames and the 25Hz rate of video frames, a special feature formatting procedure has to be done to first synchronize the unimodal sets of features. This additional step can be done by up-scaling the observation rate of the visual feature sets to the observation rate of the audio feature sets. The recognition of emotions based on low-level fusions of audio-visual data is done by using the synchronized bimodal observation vectors with HMMs. For each emotion category, we create a separate HMM and combine all the models to obtain a multi-class emotion classifier. For evaluation, we have used 3 fold cross validation method with the additional restriction that the train and test data sets do not contain samples on the same subject, for all subjects.

The simplest model consists of HMMs with one Gaussian for each state. Combining the acoustic features and LBP features extracted from 7 face regions leads to the final model accuracy of 38.55%. Using acoustic features and LBP features extracted from 11 face regions leads to the accuracy of 39.15%. Both results are superior to the results of the recognition of emotions using visual data. Still, they are worse than the results from the emotion extraction from speech. Setting the number of HMM Gaussian components to 40 and the number of HMM states to 4 like in the case of the best speech-oriented emotion classifier and combining with LBP features from the 7 regions leads to a classifier which shows 22.18% accuracy. The recognition of emotions based on decision level fusion implies the combination of the final classification results obtained by each modality separately. For this, we take into consideration four sets of unimodal classification results namely from the speech-oriented analysis and from the separate LBP-oriented analysis which use visual features from the whole face image, from 7 face regions and from 11 face regions. We use these sets together with a weighting function that allows for setting different importance levels for each set of unimodal results. This weight-based semantic fusion approach models the asynchronous character of the emotion in visual and auditory channels according.

The best model obtained in this way has the accuracy of 56.27%. Although this result reflects an improvement when compared to the emotion recognition from the unimodal approaches considered, it represents only a slight increase of performance.

## 4. Results

Studying the unimodal recognition of emotions on Enterface 2005 shows that the speech-oriented analysis proves to be more reliable than the facial expression analysis. The best classifier we obtained in case of using HMM models with MFCC features has the accuracy of 55.90%. Conversely, the best HMMbased facial expression recognition model we got uses LBP features and has the accuracy of 37.71%. The difference of 18.19% between the classification rates achieved on separate modalities is close to the same difference between the unimodal performances reported by Paleari and Huet [18] and by Mansoorizadeh and Charkari [14]. However, we obtained better results than the results from these two research papers, for the emotion analysis on separate modalities. Moreover, as opposed to the work of Mansoorizadeh and Charkari [14] which attempts the person dependent recognition of emotions, our models are completely independent of the identity of the users. To support this approach, we use n-fold cross validation and separate the samples of each subject in the train set from the test set.

The best facial expression recognition classifier we have obtained is based on the use of local binary patterns - LBPs. The rather low results of the models based on optical flow estimation, can be explained by the limited visual representation of the feature set. Extracting feature observations from consecutive frames of the 25Hz video sequences, does not offer enough information to describe the dynamics of the emotion generation process. A solution is to calculate and to derive features from the face motion flow applied over large integration windows. The fusion of audio and video features leads to results that are at best, close to the best unimodal classification result. In order to improve the fusion results, more investigations are needed.

## 5. Conclusion

The current paper has proposed a method for bimodal emotion recognition using face and speech data. The advantage of such a method is that the resulting models overcome the limited efficiency of single modality emotion analysis. We focus on the person-independent recognition of prototypic emotions from audio-visual sequences.

The novelty of our approach is in the use of hidden Markov models for the classification process. Furthermore, we introduced a new technique to select the most relevant visual features, by running a separate modelling study on a separate database of facial expressions. The HMM and Adaboost.M2 algorithms we have used for the recognition relate to multi-class classification methods. Finally, we show that the fusion at the semantic level provides the best performance for the multimodal emotion analysis.

## References

[1] Busso C., Deng Z., Yildirim S., Bulut M., Lee C. M., Kazemzadeh A., Lee S., Neumann U., Narayanan S., *Analysis of emotion recognition using facial expressions, speech and multimodal information*. In ICMI ’04: Proceedings of the 6th international conference on Multimodal interfaces, pages 205–211, New York, NY, USA, 2004. ACM

[2] Caridakis G., Malatesta L., Kessous L., Amir N., Raouzaiou A., Karpouzis K., *Modeling naturalistic affective states via facial and vocal expressions recognition*. In ICMI ’06: Proceedings of the 8th international conference on Multimodal interfaces, pages 146–154, New York, NY, USA, 2006

[3] Datcu D., Rothkrantz L. J. M., *Semantic audio-visual data fusion for automatic emotion recognition*, Euromedia'2008 Porto, ISBN 978-9077381-38-0, pp. 58-65, Eurosis, Ghent, April 2008

[4] Dubuisson S., Davoine F., Masson M., *A solution for facial expression representation and recognition*. SP:IC, 17(9):657–673, October 2002

[5] Edwards G. J., Taylor C. J., Cootes T. F., *Interpreting face images using active appearance models*. In FG ’98: Proceed ings of the 3rd. International Conference on Face & Gesture Recognition, page 300, Washington, DC, USA, 1998. IEEE Computer Society

[6] Ekman P., Friesen W. V., *Facial action coding system: investigator’s guide*. Consulting Psychologists Press, Palo Alto, 1978

[7] Fellenz W. A., Taylor J. G., Cowie R., Cowie E. D. , Piat F., Kollias S. D., Orovas C., Apolloni B., *On emotion recognition of faces and of speech using neuralnetworks, fuzzy logic and the assess system*. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN’00, 2000

[8] Freund Y., Schapire R. E., *A decision-theoretic generalization of online learning and an application to boosting*. Journal of Computer and System Science, 55:119–139, 1997

[9] Go H. J., Kwak K. C., Lee D. J., Chun M. G., *Emotion recognition from the facial image and speech signal*. In SICE 2003 Annual Conference, volume 3, pages 2890–2895, August 2003

[10] Han M.J., Hsu J.H., Song K.T., Chang F.Y., *A new information fusion method for bimodal robotic emotion recognition*. JCP, 3(7):39–47, 2008

[11] Hoch S., Althoff F., McGlaun G., Rigoll G., *Bimodal fusion of emotional data in an automotive environment*. volume 2, pages ii/1085–ii/1088 Vol. 2, March 2005

[12] Joo J. T., Seo S. W., Ko K.E., Sim K.B., *Emotion recognition method based on multimodal sensor fusion algorithm*. In ISIS’07, 2007

[13] Kanade T., Cohn J. F., Tian Y., *Comprehensive database for facial expression analysis*. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG’00), pages 46–53, 2000

[14] Mansoorizadeh M., Charkari M. N., *Bimodal person dependent emotion recognition comparison of feature level and decision level information fusion*. In PETRA '08: Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments, pages 1–4, New York, NY, USA, 2008. ACM

[15] Martin O., Kotsia I., Macq B., Pita I., *The enterface'05 audio-visual emotion database*. In 22nd International Conference on Data Engineering Workshops. ICDEW'06, page 8, 2006

[16] Meng H., Pittermann J., Pittermann A., Minker W., *Combined Speech-Emotion Recognition for Spoken Human-Computer Interfaces*, ICSPC 2007. IEEE International Conference on, 1179–1182, Nov. 2007

[17] Ojala T., Pietikainen M., Maenpaa T., *Multiresolution grayscale and rotation invariant texture classification with local binary patterns*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):971–987, 2002

[18] Paleari M., Huet B., *Toward emotion indexing of multimedia excerpts*. In Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on, pages 425–432, June 2008

[19] Paleari M., Lisetti C. L., *Toward multimodal fusion of affective cues.* In Proceedings of the 1st ACM international workshop on Human-centered multimedia, pages 99–108, 2006

[20] Sebe N., Cohen I., Gevers T., Huang T. S., *Emotion recognition based on joint visual and audio cues*. Pattern Recognition, International Conference on, 1:1136–1139, 2006

[21] Song M., Bu J., Chen C., Li N., *Audio-visual based emotion recognition, a new approach*. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 2:1020–1025, 2004

[22] Viola P., Jones M. J., *Robust real-time face detection*. IJCV:International Journal of Computer Vision, 57(2):137–154, May 2004

[23] Wang Y., Guan L., *Recognizing human emotion from audiovisual information*. Acoustics, Speech, and Signal Processing, 2005. Proceedings. ICASSP'05. IEEE International Conference on, 2:ii/1125–ii/1128 Vol. 2, March 2005

[24] Wimmer M., Schuller B., Arsic D., Radig B., Rigoll G., *Low-level fusion of audio and video feature for multi-modal emotion recognition*. In 3rd International Conference on Computer Vision Theory and Applications. VISAPP, volume 2, pages 145–151, Madeira, Portugal, January 2008

[25] Zeng Z., Hu Y., Roisman G. I., Wen Z., Fu Y., Huang T. S., *Audio-visual spontaneous emotion recognition*. In Artifical Intelligence for Human Computing, volume 4451 of Lecture Notes in Computer Science, pages 72–90. Springer, 2007