# EMOTIONS RECOGNITION BY SPEECH AND FACIAL EXPRESSIONS ANALYSIS

*Simina Emerich[1], Eugen Lupu[1], Anca Apatean[1]*

[1] Communication Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania
Simina.Emerich@com.utcluj.ro

**ABSTRACT**

*In this paper, we propose a bimodal emotion recognition system using the combination of facial expressions and speech signals. The models obtained from a bimodal corpus with six acted emotions and ten subjects were trained and tested with different classifiers, such as Support Vector Machine, Naive Bayes and K-Nearest Neighbor. In order to fuse visual and acoustic information, two different approaches were implemented: feature level fusion and match score level fusion. Comparative studies reveal that the performance and the robustness of emotion recognition systems can be improved by the use of fusion-based techniques. Further, the fusion performed at the feature level showed better results than the one performed at the score level.*

## 1. INTRODUCTION

In human-computer interaction, emotional processes are inseparably connected to rational decisions; hence affective interaction has gained great attention. Therefore, it has become an important issue to identify the user emotional state. Based on psychological theory, it is widely accepted that six archetypal emotions can be identified: surprise, fear, disgust, anger, happiness and sadness. Facial motion and tone of the speech play a major role in expressing these emotions.

Emotions can significantly change the message sense: sometimes it is not what was said that is the most important, but how it was said. The face tends to be most visible form of emotion communication, but it is also most easily controlled in response to different social situations when compared to the voice and other ways of expression. Further a brief review of existing emotion recognition systems is presented.

**Facial Expression Recognition Studies:** Since 1970, Paul Ekman has performed extensive studies on human facial expressions. He found that facial expressions of emotion are not culturally determined, but universal to human culture and thus biological in origin. He developed the Facial Action Coding System (FACS) where movements on the face are described by a set of action units (AUs). [7].

The studies in computer-assisted recognition of facial expressions started in 1990s. The features used are typically based on local spatial position of specific points and regions of the face (edge of the mouth, eyes, eyebrows).

Mase (1991) was one of the first researchers who used image processing techniques to recognize facial expressions. With 11 windows manually located in the face, the muscle movements were extracted by the use of optical flow. K-nearest neighbour rule was employed for the classification task of four emotions, with an accuracy of 80%. Rosenblum (1996) and Otsuka (1997) also developed an optical flow region-based approach, by applying a Radial Basis Function Network and a Hidden Markov Model, respectively. Tian et al. (2000) explored AUs recognition by using permanent and transient facial features (lips, wrinkles). Geometrical models were used to locate their shapes and appearances. They achieved 96% accuracy with a Neural-Network-Based classifier. Cohen (2003, 2004) introduced the Bayesian Network classifiers in the static settings and a multi-level HMM classifier to automatically segment an arbitrary long sequence to the corresponding facial expressions [3], [4].

**Vocal Emotion Recognition Studies:** Following the long tradition of speech analysis, many efforts were taken to recognize affective states from vocal information. Starting in the 1930s, some important voice feature vectors have been chosen for research: fundamental frequency, time-energy distribution vector, MFCC, LPCC coefficients, etc. Williams and Stevens (1972) studied the spectrograms of real emotional speech and compared them with acted speech. They found similarities which suggest the use of acted data. A qualitative correlation between emotion and speech feature was presented by Murray and Arnott (1993). Petrushin (1998) compared human and machine recognition of emotions in speech and achieved similar rates for both. To exploit the dynamic variation along an utterance, Mel-Frequency Cepstral Coefficients were employed. Nwe (2001) achieved an average accuracy of 70% for six emotions acted by two speakers using 12 MFCC features as input to a discrete Hidden Markov Model. Busso (2004) also argued that statistics relating to MFCCs carry emotional information. Yu et. al. (2002) used Support Vector Machines as binary classifiers. On four distinct emotions, they achieved an accuracy of 73%. Lee (2002) tried to distinguish between negative and positive emotions, in call center environment, using linear discrimination, k-NN and SVM classifiers achieving a maximum accuracy rate of 75%. Batliner (2003) studied a 4-class problem with elicited emotions in spontaneous speech [3].

**Multimodal Emotion Recognition Studies:** In order to improve the unimodal systems' recognition accuracy several studies attempted to exploit the advantage of using multimodal systems, especially by fusing audio-visual information. De Silva et al (2000) proposed a rule-based singular classification of audio-visual data recorded from two subjects into six emotions. From the audio data, they selected prosodic features, and from the video data, they chose the maximum distances and velocities between six specific facial points. Using decision-level fusion, a recognition rate of 72% was

reported. A set of singular classification methods was proposed by Chen and Huang (2000), in which audio-visual data collected from five subjects was classified into the Ekman's six basic emotions[5]. In both studies, the performance of the system increased when both modalities were fused. A large-scale audio-visual dataset was collected by Zeng et al. (2004), it containing five affective responses (confusion, interest, boredom and frustration) in addition to the six basic ones. They used the Naive Bayes classifier as the update rule, achieving an accuracy of almost 90%. There are only a few attempts to combine information from body movement and gestures. Kaliouby's (2005) model infers acted mental states from head movements and facial expressions. Gunes and Piccardi (2006) fused at different levels facial expressions and body gestures. Another multimodal system based on facial expression, body gestures and speech was implemented by Castellano in 2007[1].

The remainder of the paper is organized as follows: Section 2 is an overview of proposed methods and tools. Section 3 explains the steps of the feature extraction, presents the databases used to perform the experiments and how information was fused. At the end, evaluation results are summarized in Section 4 and final conclusions are presented in Section 5.

## 2. METHODS AND TOOLS

One of the most important tasks is to find a proper choice of feature vectors. The first challenge was to select several indicators attributable to the emotional behavior. In order to fulfill this, many features have been explored and further we present the used ones: Mel Frequency Cepstral Coefficients and the statistical moments for speech, respectively wavelet coefficients and the seven moments of Hu for images.

### 2. 1. The Mel-Frequency Cepstral Coefficients

There is a variety of temporal and spectral features that can be extracted from human speech. Prosodic features have been known to be an important indicator of emotional states, and thus they have been used in the design of many vocal emotion recognition systems. Several recent studies have also shown that there are variations across emotional states in the spectral features at the phoneme level (especially for vowel sounds) [8]. Our purpose is to explore the spectral features by using the Mel-frequency cepstral coefficients (MFCCs). They have been widely employed in speech recognition also, because of superior performance when compared to other features. The mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

### 2.2. Wavelet Transform

Wavelet Transform is a relatively new analysis technique and replaces the Fourier transform sinusoidal waves by a family generated by translations and dilations of a window called mother wavelet. The suitability of wavelet transforms (WT) for image analysis is well established: a representation in terms of the frequency content of local regions over a range of scales provides an ideal framework for the analysis of image features. We used two-dimensional wavelets and corresponding scaling functions obtained from one-dimensional wavelets by tensorial product. This kind of 2DWT leads to a decomposition of approximation coefficients at level $j$ in four components: the approximation at level $j+1$ and the details in three orientations (horizontal, vertical and diagonal)[9]. The result of wavelet transform relates to the mother wavelet. In our study, we mainly worked with Daubechies wavelets.

### 2.3. Moments

The mathematical concept of moments has been used for many years in many fields, such as pattern recognition and image understanding. The mean is the first statistic moment from the computed set of 7 statistical features used by us for voice processing. The next six statistical features are median, mode, variance, standard deviation, skewness and kurtosis. The skewness denotes the third order moment, which characterizes the degree of asymmetry of a distribution around its mean, or the degree of deviation from symmetry about the mean. The kurtosis is the fourth-order moment of a distribution and is a measure of the degree of the histogram sharpness; is a measure of the flatness or peakedness of a curve.

When moments are used to describe an image, the global properties of respective image are exploited. A significant work considering moments for pattern recognition was performed by Hu [10]. He derived a set of seven moment invariants, using non-linear combinations of geometric moments. These invariants remain the same under image translation, rotation and scaling.

### 2.4. Support Vector Machine

SVMs are well known in the pattern recognition community and are highly popular due to their generalization capabilities achieved by structural risk minimization oriented training. In many cases, its performance is significantly better than that of competing methods. Non-linear problems are solved by a transformation of the input feature vectors into a generally higher dimensional feature space by a mapping function, the *Kernel*. Maximum discrimination is obtained by an optimal placement of the maximum separation between the borders of two classes. The plane is spanned by the *Support Vectors*. SVM can handle two-class problems but a variety of strategies exist for multiclass discrimination. To construct an optimal hyperplane, SVM employees an iterative training algorithm, used to minimize the error function at the training. A large number of kernels can be used in SVM models, including linear, polynomial, radial basis function (RBF) and sigmoid. We concentrated on an RBF and Polynomial kernel, because both give promising results [6].

For SVM classification a vector is constructed by concatenation of length normalized emotion patterns. To transform a feature sequence with dynamic length to a static one, length normalization and sub-sampling techniques were applied. It is not known beforehand which $C$ and $g$ or $d$ are the best solution for a problem; consequently a model selection (parameter search) must be done.

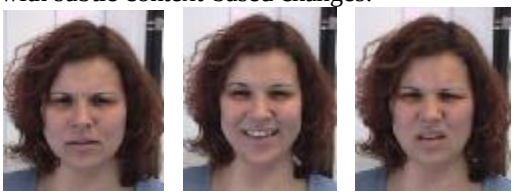
## 3. EXERIMENTS BACKGROUND

### 3.1 The Employed Databases

Emotion recognition has been investigated with three main types of databases: acted emotions, natural spontaneous emotions and elicited emotions. The best results are generally obtained with acted emotion databases because they contain strong emotional expressions.

In this paper two acted emotional databases were employed. For vocal emotions studies, the data comes from the Berlin Database of Emotional Speech. It contains about 500 utterances spoken by actors in a happy, angry, fear, sad and disgusted way as well as in a neutral version. The classification items can be chosen from 10 different actors (males and females) uttering ten different sentences [11].

For facial expressions the Feedtum Emotion Database was used. It does not contain natural facial expressions, but volunteers were asked to act. The image sequences are taken in a laboratory environment and frontal face views [12]. Each record starts and finishes with a neutral state. 18 subjects participated in the sessions, an example being given in figure 1. Authentic facial expressions are difficult to collect because they are relatively rare, short lived and filled with subtle context-based changes.



a. angry b. happy c. disgust

d. sad e. fear f. normal

Figure 1. Facial expressions in the Feedtum database[12]

### 3.2. Feature Extraction

Feature extraction involves simplifying the amount of resources needed to describe a large set of data. Analysis with a large number of variables generally requires a large amount of memory and computation power. Feature extraction can be seen as a method of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Therefore, the main purpose in the feature extraction process is to generate features that exhibit high information packing properties. In the first step we explored the use of either facial expressions or speech to detect human affective states and later our efforts focused on emotion recognition using both modalities.

For each speech frame of 30 ms, a set of mel-frequency cepstrum coefficients was computed. The number of MFCC was chosen as 20. A series with the mean of MFCCs was detected for every utterance. Figure 2 shows the variation in 2nd MFCCs for a speaker uttering *Der Lappen liegt auf dem Eisschrank* (The tablecloth is laying on the fridge) in emotional states of happiness and anger.

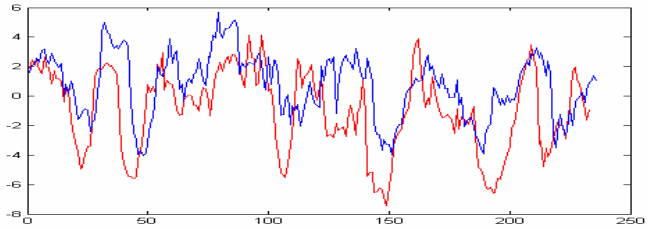



Figure 2. 2nd MFCC trace for happiness and anger utterances

In addition 7 statistical features were computed for every utterance, obtaining a 27 length feature vector. Given a digital image, or a region within an image, the feature extraction task implies the taking out of a quantity of information capable to characterize the original image. Working with two-dimensional signals, the number of samples is much bigger than in the case of a one-dimensional signal, thus the necessity of information quantity reduction is obvious.

On our images of 256x256 pixels, jpg format, the wavelet transform was computed by the iterative method of Mallat, yielding wavelet coefficients at each level in a resolution pyramid, where at each successive level the image resolution is decreased by factor 2. Five iteration steps were applied, and the approximation coefficients together with the seven moments of Hu, subjected to classification analyses. For every image the computed vector dimension is 8x8+7.

### 3.3. Information Fusion

Humans recognize one another emotional state based on the evidence presented by multiple characteristics (behavioral or physical) in addition to several contextual details associated with the environment. Each modality on its own cannot always be reliably used to perform classification. Bimodal systems can be expected to offer substantial improvement in the matching accuracy, depending upon the information being combined and the fusion methodology adopted. Bimodal systems also address the problem of noisy data. For example, in the presence of ambient noise, when an individual's voice characteristics cannot be accurately measured, the facial characteristics may be used to perform classification. Fusion can be accomplished at various levels. In our approaches we fuse information at the feature level and at the match score level.

**Feature level fusion** involves consolidating the evidence presented by two feature sets of the same individual. The feature sets (from voice and facial images) are non-homogeneous, so we concatenated them to form a single feature vector. The individual feature values of used vectors exhibit significant differences in their range as well as their distribution. The goal is to modify the location (mean) and scale (variance) via a transformation function in order to map them into a common domain. From the variety of normalization schemes, we selected the z-score normalization that uses the arithmetic mean and standard deviation of the training data.

**Fusion at the match score level** The match score is a measure of similarity between the input and template bio-

metric feature vectors. When match scores output by different biometric matchers are consolidated in order to arrive at a final recognition decision, fusion is said to be done at the match score level. In this study, the match scores have been combined using the weighted sum of scores fusion rule after the normalization of each matcher's output. We considered the minimum and maximum values for the given set of training match scores and then we applied the min-max normalization.

### 3.4. Classification

First, we inspired from Weka [13], where an important number of classifiers are developed for data mining task. It was observed that SMO classifier from Weka, which is a classical SVM, gives the higher rates compared to other classifiers: Naïve Bayes, *K-Nearest Neighbor*. Therefore, we chose to employ an SVM classifier for the emotional recognition systems. Different kernels for this type of classifier have been also tested: normalized polynomial, polynomial, RBF and the kernel based on Pearson function. From all these types of kernels, the polynomial and RBF ones give the best results. So, we concentrated on these types of SVMs. Because the interest was to obtain the best SVM classifier, a search on the SVM's parameters space was made. In that way, we focused on the Matlab toolbox from [14], which is an efficient implementation of the SVM algorithm. The algorithm was run with the POLY and RBF kernel, with the following parameters: C∈ {0.1, 1, 10, 50, 100, 150, 200, 250, 300, 350, 400, 450}, the gamma values for the RBF kernel $g \in \{0.01, 0.5, 0.1, 1, 5\}$ and the exponent for the POLY kernel d∈ {1, 2, …,12}. The best results are detailed further. The most suitable parameters combination was: C=200 and $g = 1$ for an RBF kernel.

80% of the scores were used for training and the remaining 20% for testing, in order to evaluate the systems performances. A 10-fold cross validation technique was also employed. The training data was randomly split into ten sets, 9 of which were used in training and the 10th for validations. Then iteratively another nine were picked and so forth.

## 4. EXPERIMENTS AND RESULTS

By the use of three different systems based on speech signal, facial expression and bimodal information, six emotions (sadness, happiness, anger, disgust, fear and neutral state) are recognized. The first purpose was to quantify the performance of the unimodal systems, to recognize the strengths and weaknesses of these approaches and after that to fuse these modalities in order to increase the overall recognition rate of the system. The results, presented in the first three tables are obtained by using the SVM classifier (RBF Kernel) and 10-fold cross validation technique.

The confusion matrix of the emotion classification system based only on the acoustic information (6 classes, 45 utterances for each class) is given in Table 1. The overall performance in this case was 87.7%. The diagonal components reveal that all the emotions can be recognized with more than 80% accuracy, by using a 27 coefficients feature vector. It can be seen that some pairs of emotions are usually confused more. Happiness is misclassified as disgust (8.8%) and vice versa (6.6%). Another important confusion appears between sadness and neutral state (6.6% and 4.4%, respectively). Happiness is also confused with anger. These results can be explained by similarity patterns observed in the spectral features of these emotions (Figure 2). Results showed that spectral features play a significant role in emotion recognition. The spectral characteristics of speech differ for various emotions even for the same sentence. Both prosodic and spectral features play significant roles in emotion recognition and our intension is to combine those information sources to further improve the performance.

Table 2 shows the confusion matrix of the facial expression classifier and allows us to analyze in detail the limitation of this emotion recognition system (6 classes, 45 images for each class, 71 coefficients for each image). The overall performance of this classifier was 90.7%. Happiness is recognized with high accuracy. In the facial expressions domain, happiness is confused often with fear. In the acoustic domain, these states can be separated with a good accuracy, so it is expected that the bimodal classifier will give better performance for them. Considering the performances of the unimodal systems the one based on facial expressions appears to be the most successful.

TABLE I. CONFUSION MATRIX OF THE EMOTION RECOGNITION SYSTEM BASED ON SPEECH INFORMATION

| **Classified as à** | **a** | **b** | **c** | **d** | **e** | **f** |
|---|---|---|---|---|---|---|
| **a=happy** | **80%** | 6.6% | 4.4% | 6.6% | 0 | 2.2% |
| **b=disgust** | 8.8% | **82.2%** | 2.2% | 2.2% | 0 | 4.4% |
| **c=neutral** | 0 | 2.2% | **91.1%** | 0 | 4.4% | 2.2% |
| **d=anger** | 2.2% | 4.4% | 0 | **93.3%** | 0 | 0 |
| **e=sad** | 0 | 0 | 6.6% | 0 | **93.3%** | 0 |
| **f=fear** | 4.4% | 6.6% | 2.2% | 0 | 0 | **86.7%** |

TABLE II. CONFUSION MATRIX OF THE EMOTION RECOGNITION SYSTEM BASED ON FACIAL EXPRESSIONS

| **Classified as à** | **a** | **b** | **c** | **d** | **e** | **f** |
|---|---|---|---|---|---|---|
| **a=happy** | **91.1%** | 2.2% | 2.2% | 0 | 0 | 4.4% |
| **b=disgust** | 2.2% | **91.1%** | 0 | 2.2% | 2.2% | 2.2% |
| **c=neutral** | 2.2% | 0 | **93.3%** | 0 | 4.4% | 0 |
| **d=anger** | 4.4% | 0 | 0 | **91.1%** | 4.4% | 0 |
| **e=sad** | 2.2% | 2.2% | 0 | 2.2% | **91.1%** | 0 |
| **f=fear** | 11.1% | 2.2% | 0 | 0 | 0 | **86.7%** |

TABLE III. CONFUSION MATRIX OF THE EMOTION RECOGNITION SYSTEM BASED ON FEATURE LEVEL FUSION

| **Classified as à** | **a** | **b** | **c** | **d** | **e** | **f** |
|---|---|---|---|---|---|---|
| **a=happy** | **91.1%** | 2.2% | 2.2% | 0 | 2.2% | 2.2% |
| **b=disgust** | 2.2% | **88.8%** | 0 | 2.2% | 0 | 6.6% |
| **c=neutral** | 2.2% | 2.2% | **91.1%** | 0 | 0 | 4.4% |
| **d=anger** | 2.2% | 0 | 0 | **97.8%** | 0 | 0 |
| **e=sad** | 0 | 0 | 2.2% | 0 | **97.8%** | 0 |
| **f=fear** | 4.4% | 2.2% | 2.2% | 0 | 0 | **91.1%** |

TABLE IV. CLASSIFICATION ACCURACY FOR DIFFERENT CLASSIFIERS

| Technique / Emotion Recognition System | | Speech Information | Facial Expressions | Feature Level Fusion |
|---|---|---|---|---|
| Classifier | 10-fold cross validation | | | |
| | SVM (RBF) | 87.7% | 90.3% | 93% |
| | SVM (POLY) | 85.2% | 88.8% | 90.2% |
| | Naïve Bayes | 67% | 68.14 | 68.7% |
| | K-NN (k=3) | 73.7% | 84.4% | 86.6% |
| | 80% training 20% testing | | | |
| | SVM (RBF) | 83.3% | 86.7% | 91.1% |
| | SVM (POLY) | 83.3% | 85.2% | 88.8% |
| | Naïve Bayes | 66.6% | 70.3% | 70.7% |
| | K-NN (k=3) | 72.2% | 83.3% | 85.2% |

Table 3 displays the confusion matrix of the bimodal system where the facial expressions and acoustic information were fused at the feature-level. The overall performance of this system was 93%. From a total number of 270 instances, 251 were correctly classified. One can observe that anger and sadness are recognized with 97.8% of accuracy. Disgust is the emotion with the lowest recognition rate. This emotion is often confused with fear (6.6%)..

Comparative results, in terms of accuracy rate, obtained from different classifiers are given in Table 4.

Since Support Vector Machine (RBF kernel) seems to be the most suitable classifier, we used it for the match score level fusion. Figure 3 shows that the maximum recognition accuracy is 92% in the case of match score fusion.

## 5. CONCLUSIONS AND FUTURE WORK

We presented a bimodal framework for emotion analysis and recognition starting from expressive faces and speech. The proposed approach tried to distinguish between six emotions (happiness, angry, fear, sadness, disgusted and neutral state) by using different classifiers. The maximum accuracy was achieved for Support Vector Machine (SVM). The first purpose was to select several suitable features for the task of emotion classification. A secondary aim was to analyze the strengths and the limitations of the unimodal emotion recognition systems based on facial expressions and speech features. The results reveal that the system based on facial expression gave better performance than the one based on speech information only for the considered emotions. It can be observed that even though the system based on audio information had poorer performance than the facial expression emotion classifier, its features have valuable information about emotions, that cannot be extracted from the visual information. Audio and visual data present complementary information. When these two modalities are fused, the performance and the robustness of the emotion recognition system are improved. Further, the fusion performed at the feature level showed better results than the one performed at the score level.

For future research, the integration of physiological states such as heart beat and skin conductivity would be expected to improve the recognition rates and eventually improve the computer's understanding of human emotional states. Gestures are widely believed to play an important role as well.

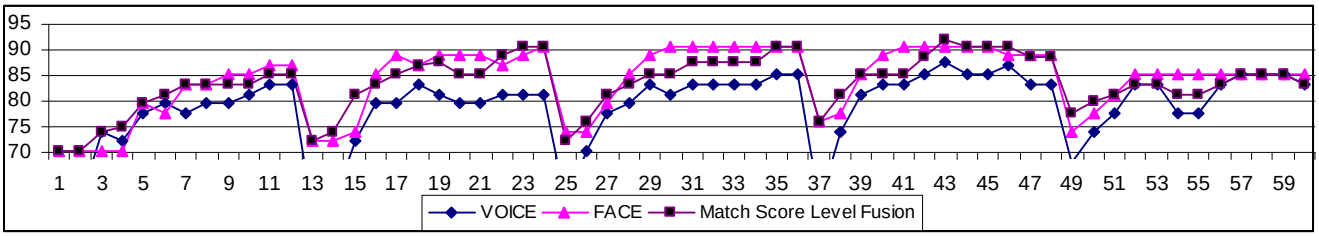



Figure 3. Classification accuracy for Unimodal Systems (Voice and Face) and for the Bimodal System obtained by fusion at the match score level. These rates have been obtained by using different parameter combinations for RBF kernel of SVM classifier

## 6. ACKNOWLEDGMENT

Part of this work has been supported by the research grant project PNCDI No. 339/2007.

## REFERENCES

[1] G. Castellano, L. Kessous, "Multimodal emotion recognition from expressive faces, body gestures and speech", 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, September 2007N.

[2] J. Kim, "Bimodal emotion recognition using speech and physiological changes", Robust Speech Rcognition and Understanding, I-Tech Education and Publishing, Vienna, 2007.

[3] Sebe, I. Cohen, T. Gevers, T.S. Huang, "Multimodal Approaches for Emotion Recognition: A Survey" Internet Imaging VI, SPIE'05, San Jose, USA, January 2005.

[4] C. Busso, Z. Deng, et al, "Analysis of emotion recognition using facial expressions, speech and multimodal information", Proceedings of the 6th international conference on Multimodal interfaces, State College, USA, 2004

[5] Chen, L.S., Huang, T.S. "Emotional expressions in audiovisual human computer interaction". Multimedia and Expo, ICME 2000.

[6] Vapnik, V. "The Nature of Statistical Learning Theory", New York, NY: Springer-Verlag, 1995

[7] Ekman, P., Friesen, W. V. Facial Action Coding System: A Technique for Measurement of Facial Movement. Consulting Psychologists Press Palo Alto, California, 1978.

[8] Lee, Chul Min et al. "Emotion recognition based on phoneme classes", In Interspeech 2004, pg 889-892.

[9] S. Mallat, "A wavelet tour of signal processing", New York, Academic Press, 1999

[10] M.K Hu, "Visual pattern recognition by moment invariants", IRE Trans.Inf. Theory, It-8, 1962, pp.179-187.

[11] http://pascal.kgw.tu-berlin.de/emodb/index-1280.html

[12] http://www.mmk.ei.tum.de/~waf/fgnet/

[13] http://www.cs.waikato.ac.nz/ml/weka/

[14] http://eprints.pascal-network.org/archive/00000348/