Features for Multimodal Emotion Recognition: An Extensive Study

Marco Paleari & Ryad Chellali
TEleRobotics and Applications
Italian Institute of Technology
Genoa, Italy
{marco.paleari; ryad.chellali} @ iit.it

Benoit Huet
Multimedia Department
EURECOM
Sophia Antipolis, France
benoit . huet @ eurecom.fr

***Abstract*—The ability to recognize emotions in natural human communications is known to be very important for mankind. In recent years, a considerable number of researchers have investigated techniques allowing computer to replicate this capability by analyzing both prosodic (voice) and facial expressions. The applications of the resulting systems are manifold and range from gaming to indexing and retrieval, through chat and health care. No study has, to the best of our knowledge, ever reported results comparing the effectiveness of several features for automatic emotion recognition. In this work, we present an extensive study conducted on feature selection for automatic, audio-visual, real-time, and person independent emotion recognition. More than 300,000 different neural networks have been trained in order to compare the performances of 64 features and 11 different sets of features with 450 different analysis settings. Results show that: 1) to build an optimal emotion recognition system, different emotions should be classified via different features and 2) different features, in general, require different processing.**

***Index Terms*—Emotion recognition; facial expressions; vocal expressions; prosody; affective computing.**

I. Introduction

The ability to recognize emotions is intrinsic in human beings and is known to be very important for natural interactions, decision making, memory, and other cognitive functions [1], [2]. As an example, during face to face meeting, it has been suggested that as much as 93% of what we communicate may be transferred through paralanguage (e.g. voice tone and volume, body language, facial expressions, etc.) [3].

In an attempt to render human–computer and human–robot interaction more similar to human-human communication and enhance its naturalness researchers have, in the last decade, approached the topic of automatic, computer based, emotion recognition [4], [5]. Indeed, the information about the emotion felt by a user interacting with a computer can be used in many different ways for human–machine interaction and computer–mediated human communications [2]. Few examples regards tele–applications such as tele–medicine and tele–learning, indexing and retrieval of media, and generally all domains of human–machine interactions from gaming to advanced domotics, security, or e–learning.

Many different techniques have been tested by literature to perform automatic emotion recognition using different modalities (auditory, visual, haptic, etc.) mainly focusing on the visual and auditory modalities. State of the art on video processing usually analyzes the facial expression by following keypoints on the face [6]–[10]. Head posture independency may be obtained with the use of elastic graph matching or active appearance models. The second most used modality for emotion recognition is probably audio. State of the art on audio processing usually takes advantage of characteristics of the voice as pitch, energy, harmonicity, speech rate, and mel-frequency cepstral coefficients [5], [7], [9]–[11]. A third modality which is often employed is physiology. State of the art process signals from the autonomous nervous system (ANS) as heartbeat, galvanic skin response, or body temperature [12], [13]. Few other modalities, such as gestures, postures, speech semantics, and others, are thought to carry affective information but are still only partially exploited and published.

Notwithstanding the fact that a considerable number of publications and surveys have been published on the topic of automatic emotion recognition, few authors, if any, discuss the matter of feature and modality selection, parameter adjustment, or classifier selection. In previous works [9], [14], [15] we have discussed how different settings, classifiers, modalities, fusion techniques, etc. perform with respect to each other. In this paper, we aim at analyzing extensively the matter of significant audio and video feature selection. For doing so we will analyze and compare the behavior of 75 different sets of one or more features extracted from audio and video of the shots presented in the eNTERFACE’05 [16] database.

II. Multimodal Approach

In our approach, emotion recognition is performed by fusing information coming from both visual and auditory modalities. We are targeting the identification of the six “universal” emotions listed by Ekman and Friesen [17] (i.e. anger, disgust, fear, happiness, sadness, and fear).

The idea of using more than one modality arises from two main observations: 1) when one, or the other, modality is not available (e.g. the subject is silent or hidden from the camera) the system will still be able to return an emotional estimation thanks to the other one and 2) when both modalities are available, the diversity and complementarity of the information, should couple with an improvement on the general performances of the system.

For our experiments the eNTERFACE’05 database [16] (see

978–1–4244–1674–5/08/$25.00 © 2010 IEEE

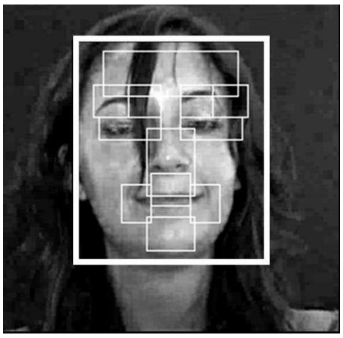


Fig. 1. Anthropometric 2D model

figure 1) has been selected. This database is composed of over 1300 emotionally tagged videos portraying non-native English speaker displaying a single emotion while verbalizing a semantically relevant English sentence. The 6 universal emotions from Ekman and Friesen [17] are portrayed, namely anger, disgust, fear, happiness, sadness, and surprise. Videos have a duration ranging from 1.2 to 6.7 seconds (2.8 ± 0.8 sec). This database is publicly available on the Internet but carries few drawbacks mainly due to the low quality of the video compression and actor performances. Please refer to [9] for an analysis of the database qualities and drawbacks.

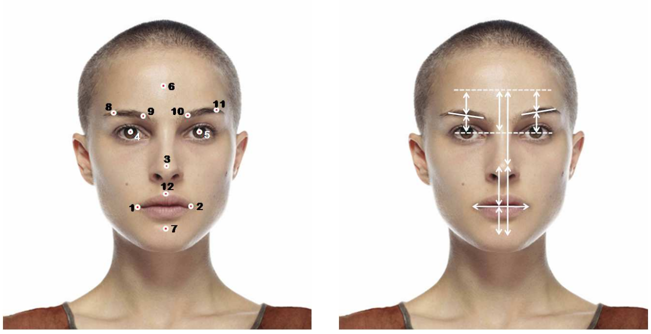
A. Facial Expression Recognition

We have developed a system performing real time, user independent, emotional facial expression recognition from still pictures and video sequences which demonstrated to work for emotion recognition [9], [14], [15]. In order to satisfy the computational time constraints required for real-time operation, we employ Tomasi Lucas-Kanade's algorithm [18] to track characteristic face points as opposed to more complex active appearance models [6], [8], [10].

As a first step we analyze the video and detect the position of the face. To the face we apply a two dimensional anthropometric model of the human face to define 12 different region of interest (see figure 1) similarly to what it was done by Sohail and Bhattacharya in [19].

1) Coordinates Feature Set: Using an approach based on the Lucas–Kanade [18] algorithm we extract 24 features per frame, corresponding to 12 pairs of the feature points (FP) $x(i)$ and $y(i)$ coordinates (see figure 2(a)) representing the average movement of points belonging to the regions of interest defined above.

2) Distances Feature Set: This set of 24 coordinate signals represents a first feature set. We have attempted to extract some more meaningful features, from these 24, in a similar way to the one adopted by MPEG-4 Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs) and to the work of Valenti et al. [8]. This process resulted in 14 features $distance(j)$ defined as mouth corner distance, chin distance to mouth, nose distance to mouth, nose distance to chin, left eye to eyebrow distance, right eye to eyebrow distance, left eyebrow alignment, right eyebrow alignment, left eyebrow to forehead distance, right eyebrow to forehead distance, forehead to eye line distance, head x displacement, head y displacement, and normalization factor proportional to head z displacement (see figure 2(b)).



(a) Feature Points

(b) Distances

Fig. 2. Video Features

B. Prosodic Expression Recognition

Our system for speech emotion recognition, takes deep inspiration from the work of Noble [11]. We use PRAAT [20] to collect 26 features from the audio part of the videos. These features are: the fundamental frequency or pitch (f_0), the energy of the signal (E), the first three formants (f_1, f_2, f_3), the harmonicity of the signal (HNR), the first nine linear predictive coding coefficients (LPC_1 to LPC_9), and the first ten mel-frequency cepstral coefficients ($MFCC_1$ to $MFCC_{10}$).

C. Sets of Features

In sections II-A and II-B we have shown how to extract $24+14+26=64$ features from video and audio which are related to the emotional expressions. To these 64 individual features we add some sets of features by grouping and concatenating them.

For the coordinates we have defined 4 sets: 1) mouth region coordinates, 2) eyes region coordinates, 3) nose coordinates, and 4) nose and forehead coordinates.

For the distances we have defined 3 sets: 1) mouth region distances, 2) eyes region distances, and 3) head displacements.

Finally, for the audio variables we have defined 4 sets: 1) pitch and energy, 2) audio formants, 3) LPC coefficients, and 4) MFCC coefficients.

This has been done with the purpose of gathering the information from different features belonging to the same set together. We expect sets of features to perform better for emotion recognition than each one of the single features. Furthermore, we want to compare different groups (e.g. regions of the face) to each other in order to better understand which ones are more interesting for automatic emotion recognition and which one need further development or finer precision.

D. Feature Vectors

As a result of the presented operation 75 sets of one or more features are created. For each one of these sets of feature f, we need to extract a feature vector which best represent the affective information. It is expected that affective information is transferred via the dynamics of the facial and prosodic expressions [21]. In order to incorporate dynamics to our framework, we have taken overlapped sliding windows $w(f)$ of the signals changing the size of the window from 1 to 50 frames with a step of one frame; longer time windows carry more information about the dynamics of the signal, shorter better represent the current state of the expression.

In addition to the original signal we investigate its dynamic properties. In particular we consider the following:

- the feature's values in time t
- the feature's first derivative Δ
- the feature's second derivative $\Delta\Delta$

We have anticipated that some statistical characteristic of the signal inside a time window may be interesting as well. For this reason we have considered, beside the signal in time, its mean and standard deviation; therefore, for each one of the three time analysis mode we consider:

- the raw feature values $raw(w(f))$
- the windows mean values $mean(w(f))$
- the windows standard deviation values $stdev(w(f))$

This section presented the extraction of relevant emotional features. Next sections will discuss our experimental settings and the results of our study.

III. Analysis Procedure

In this section the setup for the experimental analysis is presented. We randomly split the subjects in the eNTERFACE'05 database into two parts for test and training and tested the system under the completely user–independent condition, i.e. test subjects were never fed to the system during training.

Of the total 44 subjects in the database, forty were used for training and four for testing. The idea of keeping four subjects for testing instead of one originates from the need of having reliable results without performing a complete leave one out test. Four subjects provide a reasonable amount of testing samples and represent meaningful infra-subject differences without impacting too much the size of the training base. This step was repeated 3 times to validate the results over different subjects (we analyzed a total of 12 subjects for test).

All tests were carried using feed–forward neural networks with one hidden layer of 20 neurons. The output layer consists of 6 neurons (one per emotion).

For each possible combination of $feature\_set$, $mode$ and $window\_size$, we have trained a minimum of 3 different Neural Networks (NN) and averaged the different scores to reduce the "randomness" intrinsic in NN training. This results in more than 300,000 different neural networks[1].

The size of the hidden layer has been chosen arbitrarily.It may be interesting to investigate whether this parameter shall be adapted to the size of the input feature vector through some kind of heuristic. We believe that, given our objective, it is more suitable to present results obtained with a common set of parameters for every training than arbitrarily choosing heuristics to set these parameters.

Variable	ANG	DIS	FEA	HAP	SAD	SUR
R mouth corner x	0.45	0.65	**0.81**	0.57	**0.80**	0.52
R mouth corner y	0.59	0.54	0.84	0.86	0.89	0.41
L mouth corner x	0.58	**0.75**	0.64	0.54	**0.67**	0.50
L mouth corner y	0.56	0.57	**0.77**	0.60	**0.78**	0.52
nose x	0.55	**0.80**	0.85	0.46	0.61	**0.70**
nose y	0.57	0.45	0.84	0.63	0.86	0.54
right eye x	0.59	**0.81**	0.90	**0.67**	0.64	**0.78**
right eye y	0.55	0.52	**0.80**	0.52	0.52	0.52
left eye x	0.83	**0.77**	0.85	0.44	0.61	**0.69**
left eye y	0.54	0.43	0.87	**0.66**	**0.82**	0.43
forehead x	0.60	**0.74**	0.85	0.61	0.46	**0.77**
forehead y	0.60	**0.68**	**0.66**	0.49	0.84	0.47
chin x	0.43	0.60	**0.78**	0.54	**0.70**	0.56
chin y	0.54	0.52	0.63	**0.71**	0.85	0.50
ext. R eyebrow x	0.64	0.60	0.84	0.58	0.47	0.58
ext. R eyebrow y	0.59	0.46	**0.77**	0.56	0.84	0.35
int. R eyebrow x	0.60	**0.74**	0.87	0.51	0.52	0.60
int. R eyebrow y	**0.73**	0.47	0.84	0.57	0.95	0.42
int. L eyebrow x	0.52	0.55	0.88	0.50	**0.70**	0.57
int. L eyebrow y	0.64	0.51	**0.80**	0.62	0.96	0.40
ext. L eyebrow x	0.58	0.40	0.93	0.52	0.93	0.58
ext. L eyebrow y	0.56	0.47	**0.69**	0.57	0.93	0.45
upper lip x	0.96	0.96	0.94	0.66	0.92	**0.82**
upper lip y	**0.81**	**0.73**	0.99	0.60	0.99	**0.78**
mouth region	0.54	0.49	0.64	**0.74**	0.89	0.40
eyes region	**0.72**	**0.79**	0.89	0.52	0.85	0.45
nose	0.58	0.49	0.85	0.59	**0.77**	0.40
nose and forehead	0.57	0.51	0.88	0.59	0.83	0.44

TABLE I
CR^+ FOR THE COORDINATE FEATURES

IV. Results

Previous sections presented the extraction of the features as well as the analysis procedure. Because the quantity of experiments that have been carried does not allow us to present them, here, extensively we try to resume, in this section, the main outcomes.

We decided to employ as metric the recognition rate of the positive samples CR^+.

$$CR^+_{emo} = \frac{samples(emotion_{emo})\_correctly\_classified}{samples(emotion_{emo})}$$

It is important to notice each CR^+ we report represent the best result that could be found varying the triple composed by the $window\_size$ ([1,50]), the $feature\_set$ (t, Δ, or $\Delta\Delta$), and the $mode$ (raw, $mean$, or $stdev$) and shall therefore be seen as an upper–bound.

It is also very important to keep in mind that for each triple the CR^+ is obtained by computing the mean confusion matrix among a minimum of three neural networks and does not rely on a single neural network training. Doing otherwise, may deeply influence the results.

A. Video features

1) Coordinate Features: In this section, we report the results obtained from the coordinates of the 12 feature points and from the set of features of the same kind.

Table I reports, for each one of the 6 emotions, the CR^+ score for the best mode. We notice is that fear (FEA) and sadness (SAD) are generally more easily recognized than the others. Anger (ANG) is better recognized using the coordinates of the eyes and eyebrows than for the coordinates of the mouth region; this result is confirmed for the coordinate's sets with the

[1] 75 set of features by 50 $window_size$s ([1-50]) by 3 $feature_set$s (t, Δ, or $\Delta\Delta$) by 3 $mode$s (raw, $mean$, or $stdev$) multiplied by 3 different trainings for each setting and 3 different train and test databases.

	t	Δ	$\Delta\Delta$
raw	14	14	9
mean	21	16	21
stdev	17	21	35

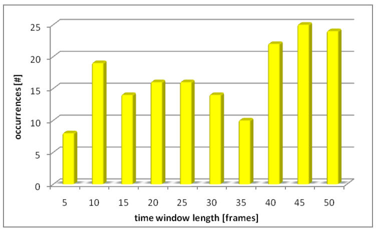


Fig. 3. Modes and window lengths for the coordinate features

set of points belonging to the eyes reaching 72% recognition rate and the second–best set performing only 58%. The same behavior is found for disgust (DIS) and fear. Happiness (HAP) is better recognized using the points belonging to the mouth than for the coordinates of the eye region; this result is confirmed from the sets of features. Finally, for both sadness and surprise we are not able to say that a particular region performs better than the others. Every region works well in recognizing sadness and the few coordinates which works better with surprise are more or less equally spread.

This result is also confirmed from the scores obtained with the coordinate's sets. The best set of features is the one grouping the coordinates of the eye regions with roughly 70%.

In figure (and table) 3, we report the distribution of the modes which have been selected to get these results. We can notice that the $\Delta\Delta$ variable is slightly preferred to both Δ and time analysis and that standard deviation ($stdev$) is being preferred to both $mean$ and raw signal. Please note that, because of the increased complexity of the relative feature vectors, the study if the raw signal is slightly disadvantaged to the other two modes.

We can also observe the histogram of the window lengths which have been selected, as the one returning the best results. From this graph we observe that, in the case of the coordinate features, windows longer than 35 frames are preferred concentrating around 50% of the best trainings. In average, the coordinate features perform 65.5%.

2) Distances Features: In the former section we have analyzed the results from the coordinate feature set; in this section we describe how the distances, which we defined in section II-A2, perform with the aid of the same graphs and figures.

Once more we observe that two emotions are better recognized than others: these are, as it was for the coordinate features, fear and sadness (see table II). The emotion which is recognized with the least accuracy is anger, with a maximum recognition of 66% for the feature relative to the alignment of the right eyebrow. We can also notice that the same emotion is better recognized for the distances relative to the eye region and for the head displacements than for the features relative to the mouth. The same behavior is found for the disgust, fear, and surprise. The emotion happiness is better recognized thanks to the movements of the mouth region similarly to what it was found for the coordinate features.

Analyzing the set of features we notice that the eye region works the best for disgust and surprise, the head displacements for anger and fear, and the mouth region performs the best for sadness.

Variable	ANG	DIS	FEA	HAP	SAD	SUR
mouth corner	0.61	0.55	0.64	0.51	0.89	0.41
chin to mouth	0.58	0.52	**0.68**	**0.70**	0.89	0.56
nose to mouth	0.39	0.63	0.44	0.57	0.85	0.40
chin to nose	0.43	0.45	**0.69**	0.59	**0.76**	0.39
left eye-eyebrow	0.57	0.53	**0.74**	0.47	**0.77**	0.40
right eye-eyebrow	0.52	**0.77**	0.83	0.52	0.57	**0.68**
left eyebrow align.	0.57	0.55	**0.79**	0.54	0.52	**0.71**
right eyebrow align.	**0.66**	0.59	0.84	0.51	**0.66**	0.42
L eyebrow-forehead	0.50	0.57	0.83	0.53	0.89	0.50
R eyebrow-forehead	0.56	0.52	**0.79**	0.50	**0.69**	**0.68**
forehead to eyes	0.49	0.53	**0.74**	0.43	0.44	0.53
x displacement	0.52	0.64	0.88	**0.68**	0.46	**0.71**
y displacement	0.49	0.42	0.85	0.37	**0.82**	0.53
z displacement	0.61	0.52	**0.76**	0.46	**0.79**	0.35
mouth region	0.49	0.49	0.61	0.49	0.87	0.34
eyes region	0.51	0.94	**0.77**	0.50	0.50	0.49
head displacements	0.60	**0.66**	0.86	0.50	**0.75**	0.37

TABLE II
CR^+ FOR THE DISTANCE FEATURES

	t	Δ	$\Delta\Delta$
raw	11	8	16
mean	12	10	12
stdev	7	10	16

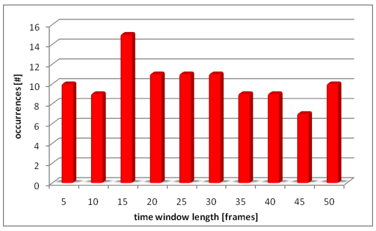


Fig. 4. Modes and window lengths for the distance features

In average the features relative to the eyes and eyebrows seem to perform better than the ones belonging to the mouth region. The same behavior is confirmed from the sets of features: here eye distances and head displacement outperform the mouth region by more than 7%. This result confirms our expectations; speech production influencing negatively the capability of the system to recognize emotions from the video signal.

We can notice from figure (and table) 4 that for the distances features the favorite mode is the second derivative of the signal ($\Delta\Delta$) gathering more than the 43% of the examples. In this case the favorite window lengths are concentrated between 15 and 30 frames. In average these features score 60% and therefore about 5% less than the coordinates features.

We have discussed the results from the video modality. In average, features relative to the eye regions work better than the ones belonging to the mouth region but this is not the case for all emotions. Unexpectedly, we observe that coordinates perform generally better than distances. In our previous tests [9], [14], [15] the two feature sets were compared showing that distances works generally better for emotion recognition. It is now to be assumed that while as a whole the decreased complexity of the distance set helps recognizing emotions, when these variables are taken one at a time (or in small sets) they are less easily exploitable.

B. Audio Features

In this section, we analyze the results of the features relative to the audio modality.

Variable	ANG	DIS	FEA	HAP	SAD	SUR
f0 - pitch	0.50	0.38	**0.76**	0.64	**0.76**	0.59
energy	0.86	0.35	0.34	0.47	0.99	0.57
f1	**0.66**	0.37	0.46	0.37	0.90	0.41
f2	0.57	0.46	0.43	0.50	0.93	0.45
f3	0.46	0.51	0.55	**0.67**	0.93	0.61
harmonicity	0.54	0.53	0.55	**0.66**	**0.69**	0.51
LPC_1	**0.82**	0.46	0.37	0.59	0.95	0.47
LPC_2	0.60	0.33	0.59	0.53	1.00	0.64
LPC_3	0.40	0.33	0.44	0.44	1.00	0.53
LPC_4	0.42	0.25	0.40	0.42	1.00	**0.69**
LPC_5	0.47	0.34	0.43	0.41	1.00	0.56
LPC_6	0.49	0.43	0.52	0.39	1.00	**0.73**
LPC_7	0.43	0.46	0.52	0.45	1.00	0.84
LPC_8	0.57	0.48	0.61	0.32	1.00	0.60
LPC_9	0.48	0.42	0.39	0.37	0.97	0.55
$MFCC_1$	0.50	0.62	0.53	0.52	0.91	**0.67**
$MFCC_2$	**0.67**	0.36	0.36	0.58	0.91	0.50
$MFCC_3$	**0.67**	0.47	0.46	0.51	0.50	0.51
$MFCC_4$	**0.71**	0.47	0.38	0.55	0.88	0.44
$MFCC_5$	**0.68**	0.47	0.55	**0.70**	0.54	0.46
$MFCC_6$	0.63	0.38	0.47	0.46	0.59	0.40
$MFCC_7$	**0.69**	0.51	0.44	0.60	0.64	0.46
$MFCC_8$	0.62	0.48	0.35	0.58	**0.78**	0.59
$MFCC_9$	**0.69**	0.54	0.34	0.47	0.58	0.44
$MFCC_{10}$	0.58	0.48	0.38	0.65	**0.69**	0.52
f0 & energy	0.61	0.32	0.63	0.61	0.92	0.48
f1. f2. f3	0.48	0.30	0.43	0.44	0.86	0.62
$LPCs$	0.49	0.39	0.41	0.42	1.00	0.56
$MFCCs$	0.60	0.36	0.49	0.64	**0.80**	0.55

TABLE III
CR^+ FOR THE AUDIO FEATURES

	t	Δ	$\Delta\Delta$
raw	31	31	12
mean	12	10	12
stdev	7	10	16

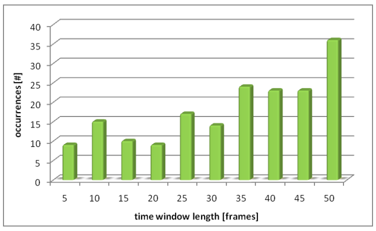


Fig. 5. Modes and window lengths for the audio features

In table III we can observe that sadness is, once more, the best recognized emotion, followed by anger, surprise, and happiness. Anger is well recognized thanks to the energy (high energy in this case), the first LPC, and most of the $MFCCs$; the best set of features is, for this emotion, the one formed by $pitch$ and $energy$. Disgust demonstrates to be an emotion which is particularly hard to recognize from audio only; the best features result to be the 1^{st} and the 9^{th} $MFCC$ and the $harmonicity$ value. Fear is well recognized using the $pitch$ value (high pitch in this case) with 76%; the 2^{nd} and the 8^{th} LPCs works quite well too. For the sets of features, the one composed of $pitch$ and $energy$ is the one returning the best result. Regarding the emotion of happiness we note that the best results are obtained while using the 5^{th} $MFCC$, the 3^{rd} $formant$ and the (higher) $harmonicity$ value. $Pitch$ also gives good results as a single variable. The set of $pitch$ and $energy$ and the one composed by the different $MFCCs$ result in the best scores for the sets of features. Sadness is easily recognized using most features.Finally, surprise is well recognized only with the use of the 7^{th}, 6^{th} , and 4^{th} $LPCs$ and with the 1^{st} $MFCC$. $Formants$ represent the set of features which best contribute to recognize this emotion.

Regarding the sets of features, the set of $pitch$ and $energy$ is the one providing the highest $average(CR^+)$ score followed by the set of the $MFCCs$.

As shown in figure (and table) 5, these results are obtained with the use of raw data (52% of the cases) while the $\Delta\Delta$ values are disadvantaged. On top of that, we also notice that 20% of the best trainings are obtained while using the longest 10% of the available window lengths. This may be due to the system need to somehow filter out samples of the audio signal when the subjects are not articulating sounds. In average this modality performs with 56.5% accuracy, roughly 9% less than the coordinate features and 4% less than the distance features. It is, nevertheless, important to notice that, at the moment, the audio emotion appraisal is returned also for frames which do not present audio (silent pieces of the video shots); somehow filtering out these estimations is likely to improve the results.

C. Summary of the Results

If we summarize the results we observe that depending on the particular emotion and feature different modes should be employed. Generally, we noticed that longer time windows provide slightly better results while increasing the number of emotionally relevant features does not seem to always improve the result. With the current settings coordinate features work in average better than distances and audio.

More specifically we can say that:

Anger is best recognized using the x coordinates of the eyes and of the upper lip, the information about the alignment of the eyebrows; for the audio we will use $energy$ and the first LPC.

Disgust is recognized with the x coordinates of the eyes, the nose, and the upper lip and the information of the distances of the eye region while using audio features other than the first $MFCC$ should be avoided.

Fear can mainly be recognized only using video features; from audio, only $pitch$ seem to return good results.

Happiness is characterized by the y coordinates of the mouth corners; the distance chin to mouth may be used too; for the audio features we will mostly rely on the 3^{rd} $formant$, the $harmonicity$, and the 5^{th} $MFCC$.

Sadness is well recognized using most features and in particular audio seem to better discriminate between sadness and all the other emotions.

Surprise is best recognized by the use of the x coordinates of eyes, nose, and upper lip, the mean face x $displacement$, and the right eyebrow alignment. For the audio features we would use the 7^{th}, 6^{th} , and 4^{th} $LPCs$ and the 1^{st} $MFCC$

V. Emotion Recognition System

In the former sections we have presented the modality of extraction of audio and video features as well as a comparative analysis of their interest for emotion recognition. In this section we briefly overview a possible use of these result for a real multimodal emotion recognition system which we have developed [22].

We used the eNTERFACE'05 database and split the subjects into a train (40 subjects) and a test (4 subjects) sets. Thanks to the study presented here we were able to select only the most relevant features for each modality and emotion. The experiments was repeated 4 times with different testing subjects.

out in	Anger	Disgust	Fear	Happin.	Sadness	Surprise
Anger	87%	11%	0%	2%	0%	0%
Disgust	14%	60%	0%	6%	21%	0%
Fear	0%	10%	75%	0%	15%	0%
Happiness	1%	0%	0%	99%	0%	0%
Sadness	15%	20%	1%	0%	64%	0%
Surprise	21%	2%	14%	7%	15%	41%

TABLE IV
CONFUSION MATRIX OF THE RESULTING MUTIMODAL SYSTEM

For this experiment we have computed three different feed–forward neural networks (50 neurons in the hidden layer) per emotion using data respectively from the audio, the coordinate, and the distances feature sets. For each one of the 6 emotions we have employed a Bayesian approach to extract a single multimodal emotion estimate per frame from the three unimodal neural network outputs[2].

The resulting system, simply detecting the most likely emotion by searching from the maximum estimation between the 6 different detectors perform an average recognition rate equal to 45.3% ($wstd = 0.73$)[3].

We have, then, computed the minimum, maximum, average, and standard deviation values for each one of the detector outputs and proceeded to normalize the minimum and average outputs of the 6 different emotions raising the mean recognition rate to 50.3% and decreasing the $wstd$ to 0.19.

Finally we have applied a thresholding strategy to filter out results whose likelihood was too low obtaining a lower recall of 0.125 (i.e. returning in average three estimates per second).

In table IV we report the confusion matrix for this system. As one can see, with the sole exception of surprise which is often confused with anger, fear, and sadness all emotions are recognized in more than 60% of the cases. Happiness is recognized in 99% of the samples in our test bases.

VI. CONCLUDING REMARKS

We have presented an extensive study on three feature sets (i.e. coordinates, distances, and audio) for the task of real-time, person independent, emotion recognition. Many different scenarios for human-computer interaction and human-centered computing will profit from a module performing such a task.

The study presented here involves the training of more than 300,000 different NN which are compared to evaluate 64 different features and 11 different sets of features. We have shown that individual emotions are better recognized by different features and/or modalities (audio or video). Similarly, we have demonstrated that different features do, in general, need different processing (i.e. different processing modes or time window lengths) if one wants to effectively extract the emotional information.

In a following section, we have overviewed a working prototype recognizing the correct emotion in more than 75% of the cases. Ongoing work consists in improving the classification rate by taking further advantage of the methodology and results presented here.

Future work, will also inverstigate the idea, developed in [10], of separating the frames of the video shots into two classes of silence/non silence frames and applying different processing to the two classes.

REFERENCES

[1] A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Avon Books, NY, 1994.

[2] R. Picard, *Affective Computing*, MIT Press, Cambridge (MA), 1997.

[3] A. Mehrabian, *Nonverbal Communication*, Aldine-Atherton, 1972.

[4] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[5] Z.Zeng, M. Pantic, G.I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, January 2009.

[6] I. Cohen, N. Sebe, A. Garg, S.W. Lew, and T.S. Huang, "Facial expression recognition from video sequences," in *Proceedings of ICME 2002*, 2002, pp. 121,124.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of ICMI*, 2004, pp. 205–211, State College, PA, USA.

[8] R. Valenti, N. Sebe, and T. Gevers, "Facial expression recognition: A fully integrated approach," in *ICIAPW '07: Proceedings of the 14th International Conference of Image Analysis and Processing - Workshops*, Washington, DC, USA, 2007, pp. 125–130, IEEE Computer Society.

[9] M. Paleari and B. Huet, "Toward Emotion Indexing of Multimedia Excerpts," in *CBMI '08 Sixth International Workshop on Content-Based Multimedia Indexing*, London, June 2008, IEEE.

[10] D. Datcu and L.J.M. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," in *Euromedia' 2008*, Porto, 2008.

[11] J. Noble, "Spoken emotion recognition with support vector machines," *PhD Thesis*, 2003.

[12] F. Nasoz, K. Alvarez, C.L. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cognition, Technology & Work*, vol. 6, no. 1, pp. 4–14, February 2004.

[13] O. Villon, *Modeling affective evaluation of multimedia contents: user models to associate subjective experience, physiological expression and contents description*, Ph.D. thesis, Thesis, Oct 2007.

[14] M. Paleari, B. Huet, and B. Duffy, "SAMMI, Semantic Affect-enhanced MultiMedia Indexing," in *SAMT 2007, 2nd International Conference on Semantic and Digital Media Technologies*, Dec 2007.

[15] M. Paleari, R. Benmokhtar, and B. Huet, "Evidence theory based multimodal emotion recognition," in *MMM '09* 15^{th} *Intl Conference on MultiMedia Modeling*, Sophia Antipolis, France, January 2009.

[16] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE05 Audio-Visual Emotion Database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006.

[17] P. Ekman and W. V. Friesen, "A new pan cultural facial expression of emotion," *Motivation and Emotion*, vol. 10(2), pp. 159–168, 1986.

[18] C. Tomasi and T. Kanade, "Detection and tracking of point features," April 1991, CMU-CS-91-132.

[19] A.S.M. Sohail and P. Bhattacharya, *Signal Processing for Image Enhancement and Multimedia Processing*, vol. 31, chapter Detection of Facial Feature Points Using Anthropometric Face Model, pp. 189–200, Springer US, 2007.

[20] Paul Boersma and David Weenink, "Praat: doing phonetics by computer," January 2008, [http://www.praat.org/].

[21] K. Scherer, *Appraisal processes in emotion: Theory, methods, research*, chapter Appraisal Considered as a Process of Multilevel Sequential Checking, pp. 92–120, New York: Oxford University Press, 2001.

[22] M. Paleari, B. Huet, and R. Chellali, "Towards multimodal emotion recognition: A new approach," in *Proc. of ACM CIVR 2010 Intl. Conf. on Image and Video Retrieval*, Xi'An, China, July 2010.

[2] The Bayesian approach has been preferred to other simple decision level fusion approaches and to the NNET approach [15] as one returning very good results without the need for training.

[3] To evaluate the system as a whole we used a measure of weighted standard deviation $wstd(CR^+) = \frac{std(CR^+)}{m(CR^+)}$. The $wstd$ value will be low if all emotions are recognized with the same likelihood and vice versa if some emotions are much better recognized than others, it will be high.