Recognition of Emotional States in Natural Human-Computer Interaction

R. Cowie[1], E. Douglas-Cowie[1], K. Karpouzis[2], G. Caridakis[2],
M. Wallace[3] and S. Kollias[2]

[1]School of Psychology
Queen's University
University Road, Belfast, BT7 1NN, Northern Ireland, UK
[2]Image, Video and Multimedia Systems Laboratory
National Technical University of Athens
15780, Zographou, Athens, Greece
[3]Department of Computer Science, University of Indianapolis, Athens Campus
9 Ipitou St., GR-105 57 Athens, Greece
{r.cowie, e.douglas-cowie}@qub.ac.uk, wallace@uindy.gr,
{kkarpou, gcari, skollias}@image.ntua.gr

Abstract. Affective and human-centered computing have attracted a lot of attention during the past years, mainly due to the abundance of environments and applications able to exploit and adapt to multimodal input from the users. The combination of facial expressions with prosody information allows us to capture the users' emotional state in an unintrusive manner, relying on the best performing modality in cases where one modality suffers from noise or bad sensing conditions. In this paper, we describe a multi-cue, dynamic approach to detect emotion in naturalistic video sequences, where input is taken from nearly real world situations, contrary to controlled recording conditions of audiovisual material. Recognition is performed via a recurrent neural network, whose short term memory and approximation capabilities cater for modeling dynamic events in facial and prosodic expressivity. This approach also differs from existing work in that it models user expressivity using a dimensional representation, instead of detecting discrete 'universal emotions', which are scarce in everyday human-machine interaction. The algorithm is deployed on an audiovisual database which was recorded simulating human-human discourse and, therefore, contains less extreme expressivity and subtle variations of a number of emotion labels. Results show that in turns lasting more than a few frames, recognition rates rise to 98%.

1 Introduction

The introduction of the term 'affective computing' by R. Picard [46] epitomizes the fact that computing is no longer considered a 'number crunching' discipline, but should be thought of as an interfacing means between humans and machines and sometimes even between humans alone. To achieve this, application design must take into account the ability of humans to provide multimodal input to computers, thus

moving away from the monolithic window-mouse-pointer interface paradigm and utilizing more intuitive concepts, closer to human niches ([47], [48]). A large part of this naturalistic interaction concept is expressivity [49], both in terms of interpreting the reaction of the user to a particular event or taking into account their emotional state and adapting presentation to it, since it alleviates the learning curve for conventional interfaces and makes less technology-savvy users feel more comfortable. In this framework, both speech and facial expressions are of great importance, since they usually provide a comprehensible view of users' reactions; actually, Cohen commented on the emergence and significance of multimodality, albeit in a slightly different human-computer interaction (HCI) domain, in [55] and [56], while Oviatt [50] indicated that an interaction pattern constrained to mere 'speak-and-point' only makes up for a very small fraction of all spontaneous multimodal utterances in everyday HCI [51]. In the context of HCI, [54] defines a multimodal system as one that 'responds to inputs in more than one modality or communication channel' abundance, while Mehrabian [52] suggests that facial expressions and vocal intonations are the main means for someone to estimate a person's affective state [53], with the face being more accurately judged, or correlating better with judgments based on full audiovisual input than on voice input ([54], [57]). This fact led to a number of approaches using video and audio to tackle emotion recognition in a multimodal manner ([3], [58] - [63], [67]), while recently the visual modality has been extended to include facial, head or body gesturing ([64] and [65], extended in [66]).

Additional factors that contribute to the complexity of estimating expressivity in everyday HCI are the fusion of the information extracted from modalities ([50]), the interpretation of the data through time and the noise and uncertainty alleviation from the natural setting ([4], [71]). In the case of fusing multimodal information [72], systems can either integrate signals at the feature level ([73]) or, after coming up with a class decision at the feature level of each modality, by merging decisions at a semantic level (late identification, [73] and [74]), possibly taking into account any confidence measures provided by each modality or, generally, a mixture of experts mechanism [6].

Regarding the dynamic nature of expressivity, Littlewort [76] states that while muscle-based techniques can describe the morphology of a facial expression, it is very difficult for them to illustrate in a measurable (and, therefore detectable) manner the dynamics, i.e. the temporal pattern of muscle activation and observable feature movement or deformation. She also makes a case of natural expressivity being inherently different in temporal terms than posed, presenting arguments from psychologists ([77] and [78]), proving the dissimilarity of posed and natural data, in addition to the need to tackle expressivity using mechanisms that capture dynamic attributes. As a general rule, the naturalistic data chosen as input in this work, is closer to human reality since intercourse is not acted and expressivity is not guided by directives (e.g. Neutral expression **à** one of the six universal emotions **à** neutral). This amplifies the difficulty in discerning facial expressions and speech patterns [70]. Nevertheless it provides the perfect test-bed for the combination of the conclusions drawn from each modality in one time unit and use as input in the following sequence of audio and visual events analyzed.

The current work aims to interpret sequences of events by modeling the user's behavior in a natural HCI setting through time. With the use of a recurrent neural network, the short term memory provided through its feedback connection, works as a memory buffer and the information remembered is taken under consideration in every next time cycle. Theory on this kind of network backs up the claim that it is suitable for learning to recognize and generate temporal patterns as well as spatial ones [1]. In addition to this, results show that this approach can capture the varying patterns of expressivity with a relatively low-scale network, which is not the case with other works operating on acted data.

The paper is structured as follows: in Section 2 we provide the fundamental notions upon which the remaining presentation is based. This includes the overall architecture of our approach as well as the running example which we will use throughout the paper in order to facilitate the presentation of our approach. In Section 3 we present our feature extraction methodologies, for both the visual and auditory modalities. In Section 4 we explain how the features extracted, although fundamentally different in nature, can be used to drive a recursive neural network in order to acquire an estimation of the human's state. In Section 5 we present results from the application of our methodology on naturalistic data and in Section 6 we list our concluding remarks.

2 Fundamentals

2.1 Emotion representation

When it comes to recognizing emotions by computer, one of the key issues is the selection of appropriate ways to represent the user's emotional states. The most familiar and commonly used way of describing emotions is by using categorical labels, many of which are either drawn directly from everyday language, or adapted from it. This trend may be due to the great influence of the works of Ekman and Friesen who proposed that the archetypal emotions correspond to distinct facial expressions which are supposed to be universally recognizable across cultures [34][35].

On the contrary psychological researchers have extensively investigated a broader variety of emotions. An extensive survey on emotion analysis can be found in [20]. The main problem with this approach is deciding which words qualify as genuinely emotional. There is, however, general agreement as to the large scale of the emotional lexicon, with most lists of descriptive terms numbering into the hundreds; the Semantic Atlas of Emotional Concepts lists 558 words with 'emotional connotations'. Of course, it is difficult to imagine an artificial systems being able to match the level of discrimination that is implied by the length of this list.

Although the labeling approach to emotion representation fits perfectly in some contexts and has thus been studied and used extensively in the literature, there are other cases in which a continuous, rather than discrete, approach to emotion representation is more suitable. At the opposite extreme from the list of categories are

dimensional descriptions, which identify emotional states by associating them with points in a multidimensional space. The approach has a long history, dating from Wundt's [21] original proposal to Schlossberg's reintroduction of the idea in the modern era [22]. For example, activation-emotion space as a representation has great appeal as it is both simple, while at the same time makes it possible to capture a wide range of significant issues in emotion [23]. The concept is based on a simplified treatment of two key themes:

§ Valence: The clearest common element of emotional states is that the person is materially influenced by feelings that are valenced, i.e., they are centrally concerned with positive or negative evaluations of people or things or events.

§ Activation level: Research from Darwin forward has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e., the strength of the person's disposition to take some action rather than none.

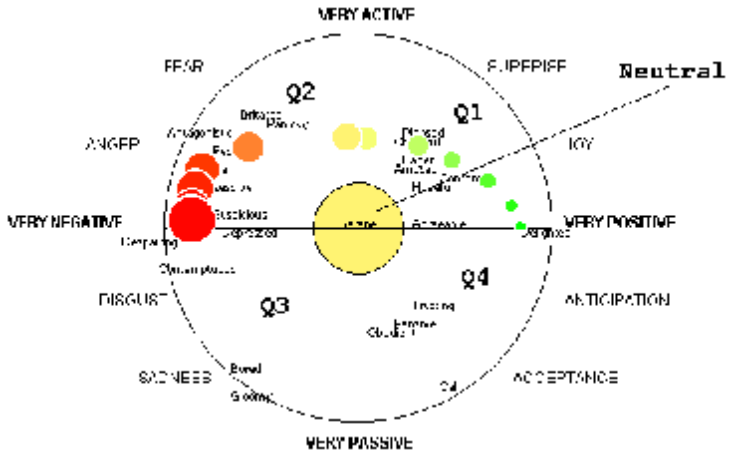

Figure 1. The activation/valence dimensional representation [30]

There is general agreement on these two main dimensions. Still, in addition to these two, there are a number of other possible dimensions, such as power-control, or approach-avoidance. Dimensional representations are attractive mainly because they provide a way of describing emotional states that is more tractable than using words. This is of particular importance when dealing with naturalistic data, where a wide range of emotional states occur. Similarly, they are much more able to deal with non discrete emotions and variations in emotional state over time.

In this work we have focused on the general area in which the human emotion lies, rather than on the specific point on the diagram presented in Figure 1. One of the reasons that has lead us to this decision is that it is not reasonable to expect human annotators to be able to discriminate between an extra pixel to the left or to the right as being an indication of a shift in observed emotional state, and therefore it does not make sense to construct a system that attempts to do so either. Thus, as is also

displayed in Figure 1, we have segmented the emotion representation space in broader areas.

As we can see in the figure, labels are typically given for emotions falling in areas where at least one of the two axes has a value considerably different than zero. On the other hand, the beginning of the axes (the center of the diagram) is typically considered as the neutral emotion. For the same reasons as mentioned above, we find it is not meaningful to define the neutral state so strictly. Therefore, we have added to the more conventional areas corresponding to the four quadrants a fifth one, corresponding to the neutral area of the diagram, as is depicted in Figure 1.

Label	Location in FeelTrace [75] diagram
Q1	positive activation, positive evaluation (+/+)
Q2	positive activation, negative evaluation (+/-)
Q3	negative activation, negative evaluation (-/-)
Q4	negative activation, positive evaluation (-/+)
Neutral	close to the center

Table 1: Emotion classes

2.2 Methodology outline

As we have already mentioned, the overall approach is based on a multimodal processing of the input sequence. 'Multimodal processing' is a general term referring to the combination of multiple input queues in order to enhance the operation of a system. In fact, there are two different methodologies that fall under this general label; decision-level and feature-level.

In the first one, independent systems are developed, each one considering one of the available information queues. The results of the different systems are then considered as independent sources of evidence concerning the optimal result, and the overall output of the system is computed through some averaging approach. This approach has the benefit of being very easy to implement when the independent systems are already available in the literature.

In the second approach, a single system considers all input queues at the same time in order to reach a single conclusion. This approach has the drawback of being often difficult to implement, as different information queues are often different in nature and are thus difficult to incorporate in one uniform processing scheme. On the other hand, when successfully realized, the feature level approach produces systems that are able to achieve considerably better performances [36].

Our approach is of the latter type; the general architecture of our approach is depicted in Figure 2.

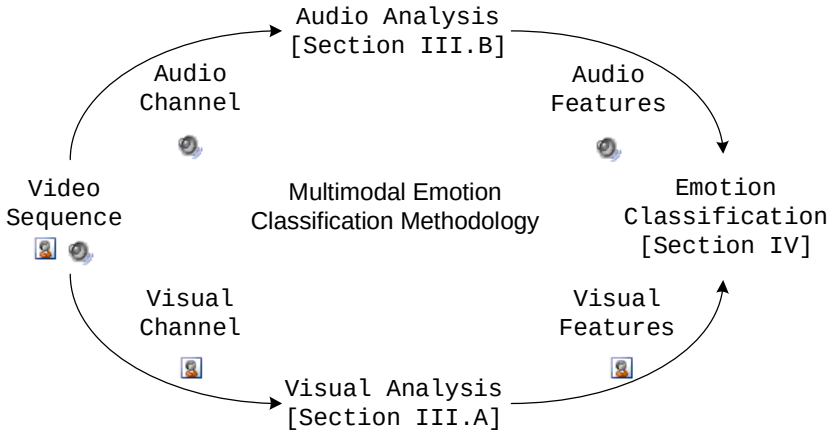

Figure 2. Graphical outline of the proposed approach

The considered input sequence is split into the audio and visual sequences. The visual sequence is analyzed frame by frame using the methodology presented in Section 3 and the audio sequence is analyzed as outlined in Section 3.2 and further explained in [3]. Visual features of all corresponding frames are fed to a recurrent network as explained in Section 4, where the dynamics in the visual channel are picked up and utilized in classifying the sequence to one of the five considered emotional classes mentioned in Table 1. Due to the fact that the features extracted from the audio channel are fundamentally different in nature than those extracted from the visual channel, the recurrent network structure is altered accordingly in order to allow both inputs to be fed to the network at the same time, thus allowing for a truly multimodal classification scheme.

The evaluation of the performance of our methodology includes statistical analysis of application results, quantitative comparisons with other approaches focusing on naturalistic data and qualitative comparisons with other known approaches to emotion recognition, all listed in Section 5.

2.3 Running Example

In developing a multimodal system one needs to integrate diverse components which are meant to deal with the different modalities. As a result, the overall architecture comprises a wealth of methodologies and technologies and can often be difficult to grasp in full detail. In order to facilitate the presentation of the multimodal approach proposed herein for the estimation of human emotional state we will use the concept of a running example.

Our example is a sample from the dataset on which we will apply our overall methodology in section 5. In Figure 3 we present some frames from the sequence of the running example.

Frame: 00815 Frame: 00820 Frame: 00825

Frame: 00830 Frame: 00835

Figure 3. Frames from the running example

3 Feature Extraction

3.1 Visual Modality

3.1.1 State of the art

Automatic estimation of facial model parameters is a difficult problem and although a lot of work has been done on selection and tracking of features [37], relatively little wp1p1ork has been reported [38] on the necessary initialization step of tracking algorithms, which is required in the context of facial feature extraction and expression recognition. Most facial expression recognition systems use the Facial Action Coding System (FACS) model introduced by Ekman and Friesen [34] for describing facial expressions. FACS describes expressions using 44 Action Units (AU) which relate to the contractions of specific facial muscles.

Additionally to FACS, MPEG-4 metrics [26] are commonly used to model facial expressions and underlying emotions. They define an alternative way of modeling facial expressions and the underlying emotions, which is strongly influenced by neurophysiologic and psychological studies. MPEG-4, mainly focusing on facial expression synthesis and animation, defines the Facial Animation Parameters (FAPs) that are strongly related to the Action Units (AUs), the core of the FACS. A comparison and mapping between FAPs and AUs can be found in [39].

Most existing approaches in facial feature extraction are either designed to cope with limited diversity of video characteristics or require manual initialization or intervention. Specifically [38] depends on optical flow, [40] depends on high resolution or noise-free input video, [41] depends on color information, [42] requires two head-mounted cameras and [43] requires manual selection of feature points on the first frame. Additionally very few approaches can perform in near-real time. In this work we combine a variety of feature detection methodologies in order to produce a robust FAP estimator, as outlined in the following.

3.1.2 Face localization

The first step in the process of detecting facial feature is that of face detection. In this step the goal is to determine whether or not there are faces in the image and, if yes, to return the image location and extent of each face [7]. Face detection can be performed with a variety of methods [8][9][10]. In this paper we have chosen to use nonparametric discriminant analysis with a Support Vector Machine (SVM) which classifies face and non-face areas, thus reducing the training problem dimension to a fraction of the original with negligible loss of classification performance [11].

In order to train the SVM and fine-tune the procedure we used 800 face examples from the NIST Special Database 18. All these examples were first aligned with respect to the coordinates of the eyes and mouth and rescaled to the required size and then the set was extended by applying small scale, translation and rotation perturbations to all samples, resulting in a final training set consisting of 16695 examples.

The accuracy of the feature extraction step that will follow greatly depends on head pose, and thus rotations of the face need to be removed before the image is further processed. In this work we choose to focus on roll rotation, since it is the most frequent rotation encountered in real life video sequences. So, we need to first estimate the head pose and then eliminate it by rotating the facial image accordingly. In order to estimate the head pose we start by locating the two eyes in the detected head location.

For this tack we utilize a multi-layer perceptron (MLP). As activation function we choose a sigmoidal function and for learning we employ the Marquardt-Levenberg learning algorithm [12]. In order to train the network we have used approximately 100 random images of diverse quality, resolution and lighting conditions from the ERMIS database [13], in which eye masks were manually specified. The network has 13 input neurons; the 13 inputs are the luminance Y, the Cr & Cb chrominance values and the 10 most important DCT coefficients (with zigzag selection) of the neighboring 8x8 pixel area. The outputs are 2, one for eye and one for non eye regions. Through pruning, the remaining architecture of the MLP has been trimmed and optimized to comprise two hidden layers of 20 neurons each.

The locations of the two eyes on the face are initially estimated roughly using the approximate anthropometric rules presented in Table 2 and then the MLP is applied separately for each pixel in the two selected regions of interest. For rotations up to 30 degrees, this methodology is successfully at a rate close to 100% in locating the eye pupils accurately.

Figure 4. Eye location using the MLP

Figure 5. Detail from Figure 4

In Figure 4 and Figure 5 we see the result of applying the MLP in the first frame of the running example. Once we have located the eye pupils, we can estimate the head roll rotation by calculating the angle between the horizontal plane and the line defined by the eye centers. We can then rotate the input frame in order to bring the head in the upright position. Finally, we can then roughly segment the rotated frame into three overlapping rectangle regions of interest which include both facial features and facial background; these three feature-candidate areas are the left eye/eyebrow, the right eye/eyebrow and the mouth. The segmentation is once more based on the approximate anthropometric rules presented in Table 2.

Figure 6. Frame rotation based on eye locations

Segment	Location	Width	Height
Left eye and eyebrow	Top left	0.6 x (width of face)	0.5 x (height of face)
Left eye and eyebrow	Top right	0.6 x (width of face)	0.5 x (height of face)
Mouth and nose	Bottom center	width of face	0.5 x (height of face)

Table 2: Anthropometric rules for feature-candidate facial areas [3]

Figure 7. Regions of interest for facial feature extraction

3.1.3 Nose localization

The nose is not used for expression estimation by itself, but is a fixed point that facilitates distance measurements for FAP estimation (see Figure 21). Thus, it is

sufficient to locate the tip of the nose and it is not required to precisely locate its boundaries. The most common approach to nose localization is starting from nostril localization; nostrils are easily detected based on their low intensity. In order to identify candidate nostril locations we apply the threshold t_n on the luminance channel of the area above the mouth region

$$t_n = \frac{\overline{L^n} + 2\min\left(L^n\right)}{3}$$

where L^n is the luminance matrix for the examined area and $\overline{L^n}$ is the average luminance in the area. The result of this thresholding is presented in Figure 8. Connected objects in this binary map are labeled and considered as nostril candidates. In poor lighting conditions, long shadows may exist along either side of the nose, resulting in more than two nostril candidates appearing in the mask. Using statistical anthropometric data about the distance of left and right eyes (bipupil breadth, D_{bp}) we can remove these invalid candidate objects and identify the true nostrils. The nose centre is defined as the midpoint of the nostrils.

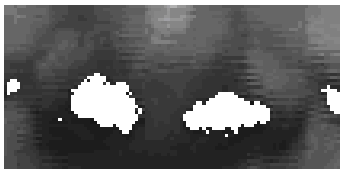

Figure 8. Candidate nostril locations

3.1.4 Eyebrow localization

Eyebrows are extracted based on the fact that they have a simple directional shape and that they are located on the forehead, which due to its protrusion, has a mostly uniform illumination.

The first step in eyebrow detection is the construction of an edge map of the grayscale eye and eyebrow region of interest. This map is constructed by subtracting the dilation and erosion of the grayscale image using a line structuring element. The selected edge detection mechanism is appropriate for eyebrows because it can be directional, preserves the feature's original size and can be combined with a threshold to remove smaller skin anomalies such as wrinkles. This procedure can be considered as a special case of a non-linear high-pass filter.

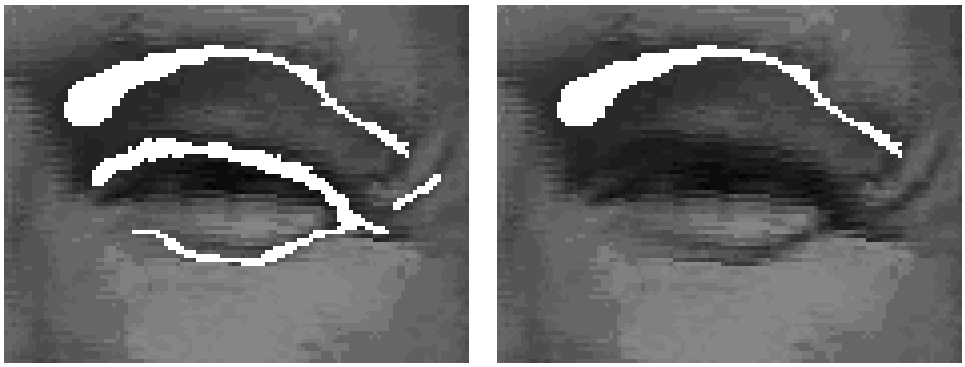

Figure 9. Eyebrow detection steps

Each connected component on the edge map is then tested against a set of filtering criteria that have been formed through statistical analysis of the eyebrow lengths and positions on 20 persons of the ERMIS database [25]. The results of this procedure for the left eyebrow are presented in Figure 9. The same procedure is also applied for the right eyebrow.

3.1.5 Eye localization

A wide variety of methodologies have been proposed in the literature for the extraction of different facial characteristics and especially for the eyes, in both controlled and uncontrolled environments. What is common among them is that, regardless of the overall success rate that they have, they all fail in some set of cases, due to the inherent difficulties and external problems that are associated with the task. As a result, it is not reasonable to select a single methodology and expect it to work optimally in all cases. In order to overcome this, in this work we choose to utilize multiple different techniques in order to locate the most difficult facial features, i.e. the eyes and the mouth.

§ MLP based mask

This approach refines eye locations extracted by the MLP network that was used in order to identify the eye pupils in the eye detection phase. It builds on the fact that eyelids usually appear darker than skin due to eyelashes and are almost always adjacent to the iris. Thus, by including dark objects near the eye centre, we add the eyelashes and the iris in the eye mask. The result is depicted in Figure 10

§ Edge based mask

This is a mask describing the area between the upper and lower eyelids. Since the eye-center is almost always detected correctly from the MLP, the horizontal edges of the eyelids in the eye area around it are used to limit the eye mask in the vertical direction. For the detection of horizontal edges we utilize the Canny edge operator due to its property of providing good localization. Out of all edges detected in the image we choose the ones right above and below the detected eye center and fill the area between them in order to get the final eye mask. The result is depicted in Figure 11

Figure 10 The MLP based eye mask

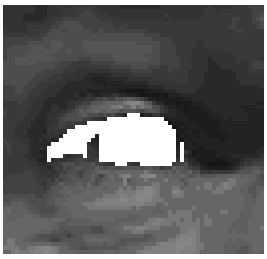

Figure 11. The edge based eye mask

§ Region growing based mask

This mask is created using a region growing technique; the latter usually gives very good segmentation results corresponding well to the observed edges. The construction of this mask relies on the fact that facial texture is more complex and darker inside the eye area and especially in the eyelid-sclera-iris borders, than in the areas around them. Instead of using an edge density criterion, we utilize a simple yet effective new method to estimate both the eye centre and eye mask.

For each pixel in the area of the center of the eye we calculate the standard deviation of the luminance channel in its 3x3 neighborhood and then threshold the result by the luminance of the pixel itself. This process actually results in the area of the center of the eye being extended in order to include some of its adjacent facial characteristics. The same procedure is also repeated for 5x5 neighborhoods; by using different block sizes we enhance the procedure's robustness against variations of image resolution and eye detail information. The two results are then merged in order to produce the final mask depicted in Figure 12. The process is found to fail more often than the other approaches we utilize, but it is found to perform very well for images of very-low resolution and low color quality. The overall procedure is quite similar to that of a morphological bottom hat operation, with the difference that the latter is rather sensitive to the structuring element size.

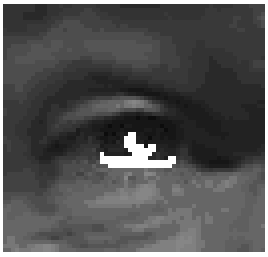

Figure 12. The standard deviation based eye mask

Figure 13. The luminance based eye mask

§ Luminance based mask

Finally, a second luminance based mask is constructed for eye and eyelid border extraction, using the normal probability of luminance using a simple adaptive threshold on the eye area. The result is usually a blob depicting the boundaries of the eye. In some cases, though, the luminance values around the eye are very low due to

shadows from the eyebrows and the upper part of the nose. To improve the outcome in such cases, the detected blob is cut vertically at its thinnest points on either side of the eye centre; the resulting mask's convex hull is depicted in Figure 13.

§ Mask fusion

The reason we have chosen to utilize four different masks is that there is no standard way in the literature based on which to select the ideal eye localization methodology for a given facial image. Consequently, having the four detected masks it is not easy to judge which one is the most correct and select it as the output of the overall eye localization module. Instead, we choose to combine the different masks using a committee machine.

Given the fact that each one of the different methodologies that we have utilized has some known strong and weak points, the committee machine that is most suitable for the task of mask fusion is the mixture of experts dynamic structure, properly modified to match our application requirements [6]. The general structure of this methodology is presented in Figure 14. It consists of k supervised modules called the experts and a gating network that performs the function of a mediator among the experts. The main assumption is that each one of the experts operates best in different regions of the input space in accordance with a probabilistic model that is known a priori, hence the need of the gating network.

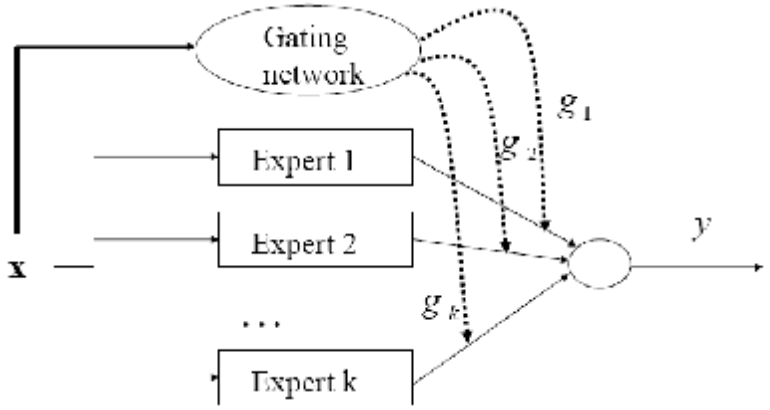

Figure 14. Mixture of experts architecture

The role of the gating network is to estimate, based on the input, the probability g_i that each individual expert i operates correctly, and to provide these estimations to the output combiner module. The gating network consists of a single layer of *softmax* neurons; the choice of *softmax* as the activation function for the neurons has the important properties of

$$0 \le g_i \le 1, \forall i \in 1..k$$

$$\sum_{i=1}^{k} g_i = 1$$

i.e. it allows for the estimations to be interpreted as probabilities. In our work we have $k = 4$ experts; the implementations of the eye detection methodologies

presented earlier in the section. The gating network favors the color based feature extraction methods in images of high color and resolution, thus incorporating the a priori known probabilities of success for our experts in the fusion process.

Additionally, the output combiner module which normally operates as $y = \bar{g} \cdot \bar{e}$, where $\bar{e}$ is the vector of expert estimations, is modified in our work to operate as

$$y = \frac{\bar{g} \cdot \bar{f} \cdot \bar{e}}{\left|\bar{f}\right|}$$

where $\bar{f}$ is the vector of confidence values associated with the output of each expert, thus further enhancing the quality of the mask fusion procedure. Confidence values are computed by comparing the location, shape and size of the detected masks to those acquired from anthropometric statistical studies.

The modified combiner module fuses the four masks together by making pixel by pixel decisions. The result of the procedure for the left eye in the frame of the running example is depicted in Figure 15.

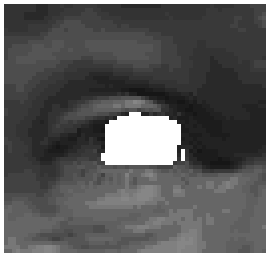

Figure 15. The final mask for the left eye

3.1.6 Mouth localization

Similarly to the eyes, the mouth is a facial feature that is not always detected and localized successfully, mainly due to the wide range of deformations that are observed in it in sequences where the human is talking, which is the typical case in our chosen field of application. In this work we utilize the following methodologies in order to estimate the location and boundaries of the mouth:

§ MLP based mask

An MLP neural network is trained to identify the mouth region using the neutral image. The network has similar architecture as the one used for the eyes. The train data are acquired from the neutral image. Since the mouth is closed in the neutral image, a long region of low luminance region exists between the lips. Thus, the mouth-candidate region of interest is first filtered with Alternating Sequential Filtering by Reconstruction (ASFR) to simplify and create connected areas of similar luminance. Luminance thresholding is then used to find the area between the lips.

This area is dilated vertically and the data depicted by this area are used to train the network.

The MLP network that has been trained on the neutral expression frame is the one used to produce an estimate of the mouth area in all other frames. The output of the neural network on the mouth region of interest is thresholded in order to form a binary map containing several small sub-areas. The convex hull of these areas is calculated

to generate the final mask for the mouth. The result of this procedure is depicted in Figure 16.

Figure 16. The MLP based mouth mask

Figure 17. Edge based mouth mask

§ Edge based mask

In this second approach, the mouth luminance channel is again filtered using ASFR for image simplification. The horizontal morphological gradient of the mouth region of interest is then calculated. Since the position of the nose has already been detected, and, as we have already explained, the procedure of nose detection rarely fails, we can use the position of the nose to drive the procedure of mouth detection. Thus, the connected elements that are too close to the nose center to be a part of the mouth are removed. From the rest of the mask, very small objects are also removed. A morphological closing is then performed and the longest of the remaining objects in the horizontal sense is selected as the final mouth mask. The result of this procedure is depicted in

§ Lip corner based mask

The main problem of most intensity based methods for the detection of the mouth is the existence of the upper teeth, which tend to alter the saturation and intensity uniformity of the region. Our final approach to mouth detection takes advantage of the relative low luminance of the lip corners and contributes to the correct identification of horizontal mouth extent which is not always detected by the previous methods.

The image is first thresholded providing an estimate of the mouth interior area, or the area between the lips in case of a closed mouth. Then, we discriminate between two different cases:

1. there are no apparent teeth and the mouth area is denoted by a cohesive dark area
2. there are teeth and thus two dark areas appear at both sides of the teeth

In the first case mouth extend is straightforward to detect; in the latter mouth centre proximity of each object is assessed and the appropriate objects are selected. The convex hull of the result is then merged through morphological reconstruction with an horizontal edge map to include the upper and bottom lips.

In order to classify the mouth region in one of the two cases and apply the corresponding mouth detection methodology we start by selecting the largest connected object in the thresholded input image and finding its centroid. If the horizontal position of the centroid is close to the horizontal position of the tip of the nose, then we assume that the object is actually the interior of the mouth and we have the first case where there are no apparent teeth. If the centroid is not close to the horizontal position of the nose then we assume that we have the second case where there are apparent teeth and the object examined is the dark area on one of the two sides of the teeth.

The result from the application of this methodology on the running example frame is depicted in Figure 19.

§ Mask fusion

The fusion of the masks is performed using a modified mixture of experts model similar to the one used for the fusion of the different masks for the eyes. The main difference here is that we cannot assess the probability of success of either of the methods using information readily available in the input, such as resolution or color depth, and therefore the gating network has the trivial role of weighting the three experts equally.

This does not mean that the output combiner module is also trivial. Quite the contrary, we still utilize the modified version of the module, where anthropometric statistics are used to validate the three masks and the degree of validation is utilized in the process of mask fusion. The resulting mask for the mouth for the examined frame is depicted in Figure 18.

Figure 18. The final mask for the mouth

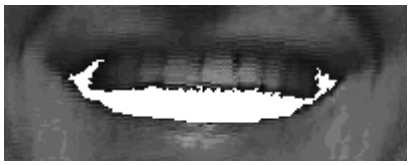

Figure 19.The lip corner based mouth mask

3.1.7 Feature points and FAPs

The facial feature masks detected in the previous section are not used directly for the emotion recognition procedure. They are merely the basis from which other, more refined, information elements will be drawn. Specifically, we utilize the masks in order to detect the marginal points of the studied elements on the face. **Table 3** presents the complete list of points detected on the human face; these are a subset of the complete list of facial feature points defined in the MPEG-4 standard [26]. For example, Figure 20 depicts the feature points detected on the frame of the running example.

Feature Point	MPEG-4	Description
1	4.5	Outer point of Left eyebrow
2	4.3	Middle point of Left eyebrow
3	4.1	Inner point of Left eyebrow
4	4.6	Outer point of Right eyebrow
5	4.4	Middle point of Right eyebrow
6	4.2	Inner point of Right eyebrow
7	3.7	Outer point of Left eye
8	3.11	Inner point of Left eye
9	3.13	Upper point of Left eyelid
10	3.9	Lower point of Left eyelid
11	3.12	Outer point of Right eye
12	3.8	Inner point of Right eye
13	3.14	Upper point of Right eyelid
14	3.10	Lower point of Right eyelid

15	9.15	Nose point
16	8.3	Left corner of mouth
17	8.4	Right corner of mouth
18	8.1	Upper point of mouth
19	8.2	Lower point of mouth

Table 3: Considered feature points

Figure 20. Feature points detected on the input frame

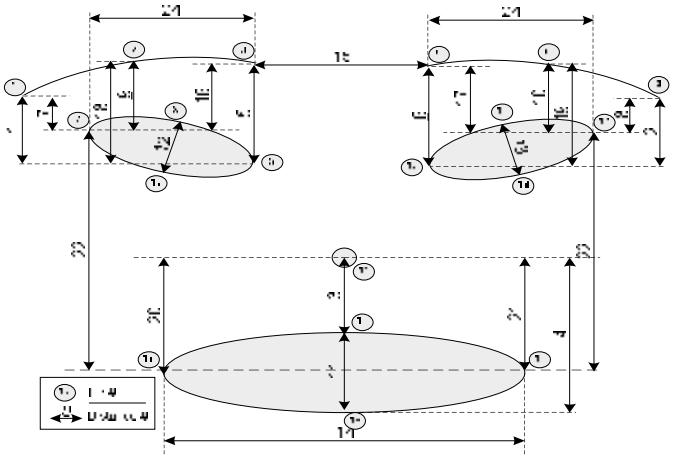

Figure 21: Feature Point Distances

As people change their facial expression their face is altered and the position of some of these points is changed (see Figure 22). Therefore, the main information unit we will consider during the emotion classification stage will be the set of FAPs that describe a frame.

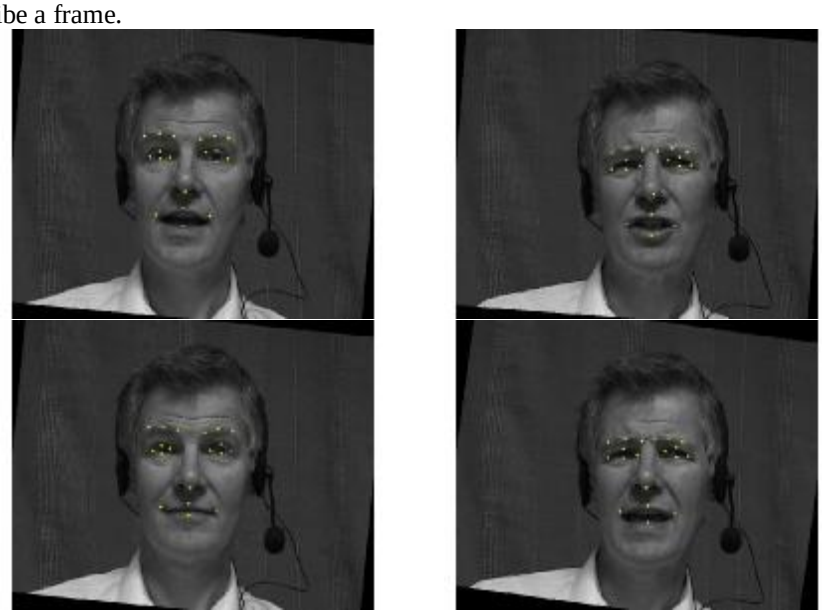

Figure 22. Feature points detected from frames belonging to different sequences

In order to produce this set we start by computing a 25-dimensional distance vector $\overline{d}$ containing vertical and horizontal distances between the 19 extracted FPs, as shown in Figure 21. Distances are not measured in pixels, but in normalized scale-invariant MPEG-4 units, i.e. ENS, MNS, MW, IRISD and ES [26]; unit bases are measured directly from FP distances on the neutral image, for example ES is calculated as the distance between FP_9 and FP_{13} (distance between eye pupils). The first step is to create the reference distance vector $\overline{d}_n$ by processing the neutral frame and calculating the distances described in Figure 21 and then a similar distance vector $\overline{d}_i$ is created for each examined frame i. FAPs are calculated by comparing $\overline{d}_n$ and $\overline{d}_i$.

3.2 Auditory Modality

3.2.1 State of the art

Figure 23 presents classification results from a selection of recent studies that it seems fair to take as representing the state of the art. They show a complex picture, key parts of which have not been formulated clearly.

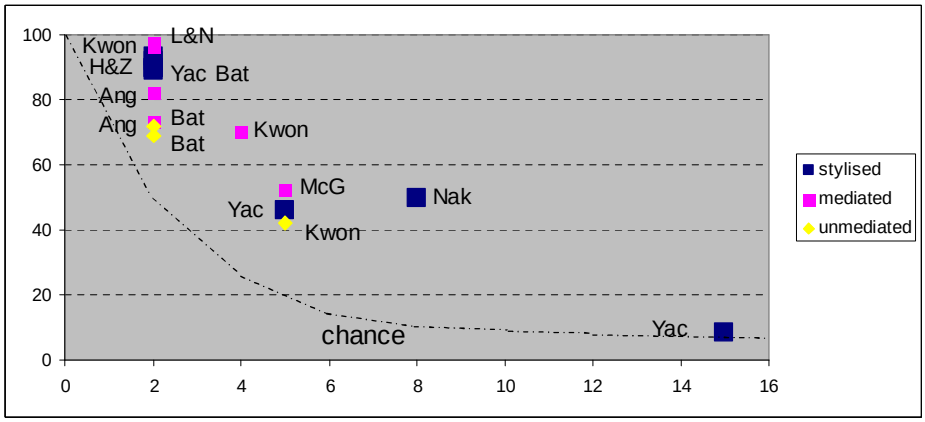

Figure 23: Plot of discrimination results from key recent studies of emotion recognition – Lee and Narayan 2003 ([87]); Kwon et al 2003 ([86]), Zhou & Hansen 1999 ([91]), Ang 2002 ([82]), Yacoub et al 2003 ([90]), Batliner et al 2003 ([83]), McGilloway et al 2000 ([88]), and Nakatsu et al 1999 ([89]); against number of categories to be discriminated (horizontal axis)

The horizontal axis shows one variable that clearly affects classification rate, that is, the number of categories considered. Under some circumstances, performance approaches 100% with two-choice classification. However, when techniques that achieved 90% in pairwise discriminations were used to assign samples to a set of 15 possible emotions [90], recognition rate falls to 8.7% (though note that this is still above chance). Intermediate points suggest an almost linear change between the two extremes.

The implication of the fall is that the techniques used in contemporary systems have limitations are concealed by using pairwise discrimination as a test. Yacoub et al clarified the issue in elegant follow-up studies, showing that their system effectively discriminated two emotion clusters, happiness/hot anger and sadness/boredom. These strongly suggest that that the information available may be more strongly related to dimensions than to categories: they appear to be a high- and a low-activation group. Of course, when analyzers that function as activation detectors are applied to stimuli which are either neutral or irritated, they will appear to detect irritation with high reliability: but the appearance is quite misleading.

The other major dimension, represented by the different types of symbol on the graph, is the extent to which the material has been stylized to simplify the task. The graph uses a simplified classification into three broad levels. Fully stylized speech is produced by competent actors, often in a carefully structured format. The second level, mediated speech, includes two main types: emotion simulated by people without particular acting skill or direction; and samples selected from a naturalistic database as clear examples of the category being considered. The third level includes speech that arises spontaneously from the speaker's emotional state, and which includes naturally occurring shades, not only well-defined examples.

3.2.2 Feature extraction

An important difference between the visual and audio modalities is related to the duration of the sequence that we need to observe in order to be able to gain an understanding of the sequence's content. In case of video, a single frame is often enough in order for us to understand what the video displays and always enough for us to be able to process it and extract information. On the other hand, an audio signal needs to have a minimum duration for any kind of processing to be able to be made.

Therefore, instead of processing different moments of the audio signal, as we did with the visual modality, we need to process sound recordings in groups. Obviously, the meaningful definition of these groups will have a major role in the overall performance of the resulting system. In this work we consider sound samples grouped as tunes, i.e. as sequences demarcated by pauses. The basis behind this is that that although expressions may change within a single tune, the underlying human emotion does not change dramatically enough to shift from one quadrant to another. For this reason, the tune is not only the audio unit upon which we apply our audio feature detection techniques but also the unit considered during the operation of the overall emotion classification system.

From Section 3.2.1 it is quite obvious that selecting the right set of audio features to consider for classification is far from a trivial procedure. Batliner et al. [92] classify features in categories and present a systematic comparison of different sets of features and combination strategies in the presence of natural, spontaneous speech. In order to overcome this in our work, we start by extracting an extensive set of 377 audio features. This comprises features based on intensity, pitch, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment characteristics and pause length.

We analyzed each tune with a method employing prosodic representation based on perception called Prosogram [44]. Prosogram is based on a stylization of the fundamental frequency data (contour) for vocalic (or syllabic) nuclei. It gives globally for each voiced nucleus a pitch and a length. According to a 'glissando treshold' in some cases we don't get a fixed pitch but one or more lines to define the evolution of pitch for this nucleus. This representation is in a way similar to the 'piano roll' representation used in music sequencers. This method, based on the Praat environment, offers the possibility of automatic segmentation based both on voiced part and energy maxima. From this model - representation stylization we extracted several types of features: pitch interval based features, nucleus length features and distances between nuclei.

Given that the classification model used in this work, as we will see in Section 4, is based on a neural network, using such a wide range of features as input to the classifies means that the size of the annotated data set as well as the time required for training will be huge. In order to overcome this we need to statistically process the acoustic feature, so as to discriminate the more prominent ones, thus performing feature reduction. In our work we achieve this by combining two well known techniques: analysis of variance (ANOVA) and Pearson product-moment correlation coefficient (PMCC). ANOVA is used first to test the discriminative ability of each feature. This resulting in a reduced feature set, containing about half of the features tested. To further reduce the feature space we continued by calculating the PMCC for all of the remaining feature pairs; PMCC is a measure of the tendency of two

variables measured on the same object to increase or decrease together. Groups of highly correlated (>90%) features were formed, and a single feature from each group was selected.

The overall process results in reducing the number of audio features considered during classification from 377 to only 32 [66]; all selected features are numerical and continuous.

4 Multimodal Expression Classification

4.1 The Elman net

In order to consider the dynamics of displayed expressions we need to utilize a classification model that is able to model and learn dynamics, such as a Hidden Markov Model or a recursive neural network In this work we are using a recursive neural network; see Figure 24. This type of network differs from conventional feed-forward networks in that the first layer has a recurrent connection. The delay in this connection stores values from the previous time step which can be used in the current time step, thus providing the element of memory.

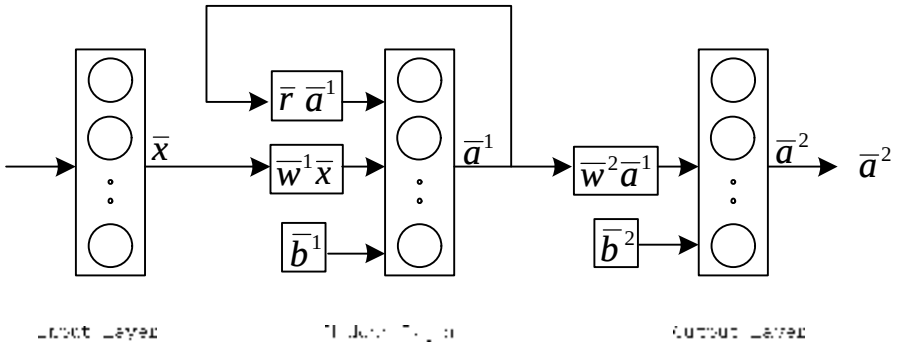

Figure 24. The recursive neural network

Although we are following an approach that only comprises a single layer of recurrent connections, in reality the network has the ability to learn patterns of a greater length as well, as current values are affected by all previous values and not only by the last one.

Out of all possible recurrent implementations we have chosen the Elman net for our work [1][2]. This is a two-layer network with feedback from the first layer output to the first layer input. This recurrent connection allows the Elman network to both detect and generate time-varying patterns.

The transfer functions of the neurons used in the Elman net are tan-sigmoid for the hidden (recurrent) layer and purely linear for the output layer. More formally

$$a_i^1 = \tan sig(k_i^1) = \frac{2}{1+e^{-2k_i^1}} - 1,\ a_j^2 = k_j^2$$

where a_i^1 is the activation of the i-th neuron in the first (hidden) layer, k_i^1 is the induced local field or activation potential of the i-th neuron in the first layer, a_j^2 is the activation of the j-th neuron in the second (output) layer and k_j^2 is the induced local field or activation potential of the j-th neuron in the second layer.

The induced local field in the first layer is computed as:

$$k_i^1 = \overline{w}_i^1 \cdot \overline{x} + \overline{r}_i \cdot \overline{a}^1 + b_i^1$$

where $\overline{x}$ is the input vector, $\overline{w}_i^1$ is the input weight vector for the i-th neuron, $\overline{a}^1$ is the first layer's output vector for the previous time step, $\overline{r}_i$ is the recurrent weight vector and b_i^1 is the bias. The local field in the second layer is computed in the conventional way as:

$$k_j^2 = \overline{w}_j^2 \cdot \overline{a}^1 + b_j^2$$

where $\overline{w}_i^2$ is the input weight and b_j^2 is the bias.

This combination of activation functions is special in that two-layer networks with these transfer functions can approximate any function (with a finite number of discontinuities) with arbitrary accuracy. The only requirement is that the hidden layer must have enough neurons ([32] and [33]).

As far as training is concerned, the truncated back-propagation through time (truncated BPTT) algorithm is used [6].

The input layer of the utilized network has 57 neurons (25 for the FAPs and 32 for the audio features). The hidden layer has 20 neurons and the output layer has 5 neurons, one for each one of five possible classes: Neutral, Q1 (first quadrant of the Feeltrace [75] plane), Q2, Q3 and Q4. The network is trained to produce a level of 1 at the output that corresponds to the quadrant of the examined tune and levels of 0 at the other outputs.

4.1.1 Dynamic and non dynamic inputs

In order for the network to operate we need to provide as inputs the values of the considered features for each frame. As the network moves from one frame to the next it picks up the dynamics described by the way these features are changed and thus manages to provide a correct classification in its output.

One issue that we need to consider, though, is that not all of the considered inputs are dynamic. Specifically, as we have already seen in section 3.2, as far as the auditory modality is concerned the tune is seen and processed as a single unit. Thus, the acquired feature values are referring to the whole tune and cannot be allocated to specific frames. As a result, a recurrent neural network cannot be used directly and unchanged in order to process our data.

In order to overcome this, we modify the simple network structure of Figure 24 as shown in Figure 25. In this modified version input nodes of two different types are utilized:

1. For the visual modality features we maintain the conventional input neurons that are met in all neural networks
2. For the auditory modality features we use static value neurons. These maintain the same value throughout the operation of the neural network.

The auditory feature values that have been computed for a tune are fed to the network as the values that correspond to the first frame. In the next time steps, while visual features corresponding to the next frames are fed to the first input neurons of the network, the static input neurons maintain the original values for the auditory modality features, thus allowing the network to operate normally.

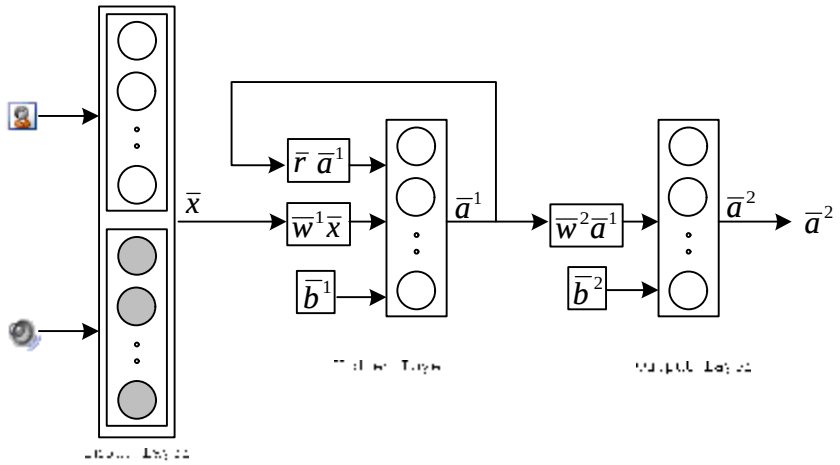

Figure 25. The modified Elman net

One can easily notice that although the network has the ability to pick up the dynamics that exist in its input, it cannot learn how to detect the dynamics in the auditory modality since it is only fed with static values. Still, we should comment that the dynamics of this modality are not ignored. Quite the contrary, the static feature values computed for this modality, as has been explained in section 3.2, are all based on the dynamics of the audio channel of the recording.

4.2 Classification

The most common applications of recurrent neural networks include complex tasks such as modeling, approximating, generating and predicting dynamic sequences of known or unknown statistical characteristics. In contrast to simpler neural network structures, using them for the seemingly easier task of input classification is not equally simple or straight forward.

The reason is that where simple neural networks provide one response in the form of a value or vector of values at their output after considering a given input, recurrent neural networks provide such inputs after each different time step. So, one question to answer is at which time step the network's output should be read for the best classification decision to be reached.

As a general rule of thumb, the very first outputs of a recurrent neural network are not very reliable. The reason is that a recurrent neural network is typically trained to pick up the dynamics that exist in sequential data and therefore needs to see an adequate length of the data in order to be able to detect and classify these dynamics. On the other hand, it is not always safe to utilize the output of the very last time step as the classification result of the network because:

1. the duration of the input data may be a few time steps longer than the duration of the dominating dynamic behavior and thus the operation of the network during the last time steps may be random
2. a temporary error may occur at any time step of the operation of the network

For example, in Figure 26 we present the output levels of the network after each frame when processing the tune of the running example. We can see that during the first frames the output of the network is quite random and changes swiftly. When enough length of the sequence has been seen by the network so that the dynamics can be picked up, the outputs start to converge to their final values. But even then small changes to the output levels can be observed between consecutive frames.

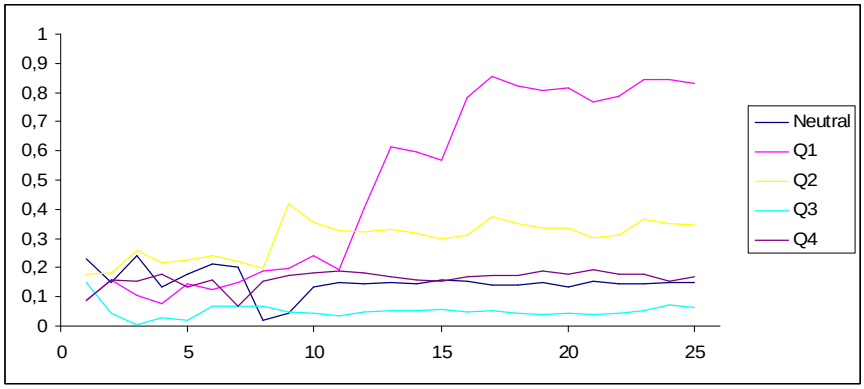

Figure 26. Individual network outputs after each frame

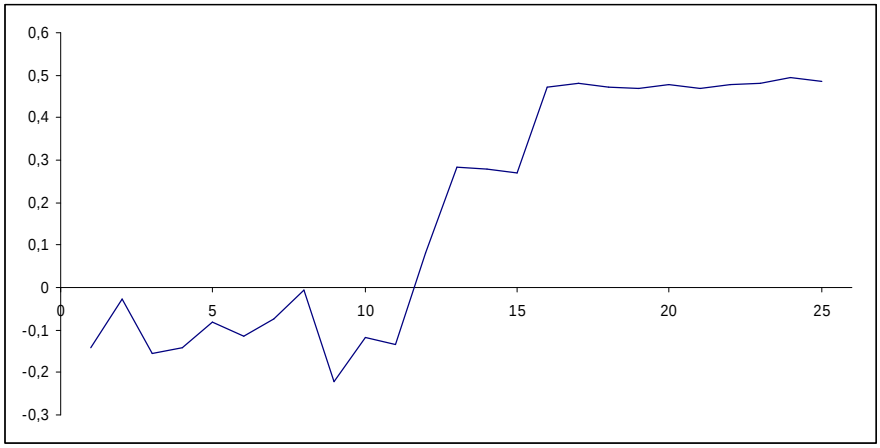

Figure 27. Margin between correct and next best output

Although these are not enough to change the classification decision (see Figure 27) for this example where the classification to Q1 is clear, there are cases in which the classification margin is smaller and these changes also lead to temporary classification decision changes.

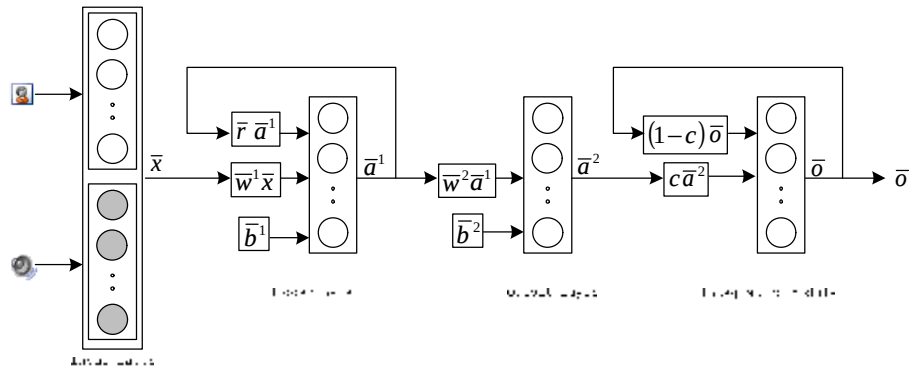

Figure 28. The Elman net with the output integrator

In order to arm our classification model with robustness we have added a weighting integrating module to the output of the neural network which increases its stability. Specifically, the final outputs of the model are computed as:

$$o_j(t) = c \cdot a_j^2 + (1-c) \cdot o_j(t-1)$$

where $o_j(t)$ is the value computed for the j-th output after time step t, $o_j(t-1)$ is the output value computed at the previous time step and c is a parameter taken from the (0,1] range that controls the sensitivity/stability of the classification model. When c is closer to zero the model becomes very stable and a large sequence of changed values of k_j^2 is required to affect the classification results while as c approaches one the model becomes more sensitive to changes in the output of the network. When $c = 1$ the integrating module is disabled and the network output is acquired as overall classification result. In our work, after observing the models performance for different values of c, we have chosen $c = 0.5$.

In Figure 29 we can see the decision margin when using the weighting integration module at the output of the network. When comparing to Figure 27 we can clearly see that the progress of the margin is more smooth, which indicates that we have indeed succeeded in making the classification performance of the network more stable and less dependent on frame that is chosen as the end of a tune.

Of course, in order for this weighted integrator to operate, we need to define output values for the network for time step 0, i.e. before the first frame. It is easy to see that due to the way that the effect of previous outputs wares off as time steps elapse due to c, this initialization is practically indifferent for tunes of adequate length. On the other hand, this value may have an important affect on tunes that are very short. In this work, we have chosen to initialize all initial outputs at

$$\bar{o}(0) = 0$$

Another meaningful alternative would be to initialize $\bar{o}(0)$ based on the percentages of the different output classes in the ground truth data used to train the classifier. We have avoided doing this in order not to add a bias towards any of the

outputs, as we wanted to be sure that the performance acquired during testing is due solely to the dynamic and multimodal approach proposed in this work.

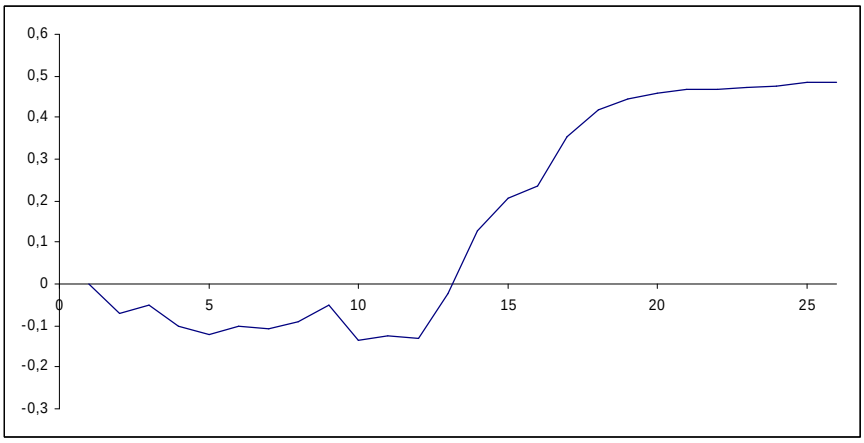

Figure 29. Decision margin when using the integrator

It is worth noting that from a modeling point of view it was feasible to include this integrator in the structure of the network rather than having it as an external module, simply by adding a recurrent loop at the output layer as well. We have decided to avoid doing so, in order not to also affect the training behavior of the network, as an additional recurrent loop would greatly augment the training time and size and average length of training data required.

5 Experimental Results

5.1 The case for naturalistic data

As Douglas-Cowie mentions in [79], the easiest way to collect emotional data is to have actors simulate it; the data produced in such experiments lends itself to extremely high recognition rates possibly in the high 70s [80]. However, this kind of intercourse is very rare in everyday human-human or human-computer interaction contexts and this is attributed to a number of reasons:

- acted experiments usually involve reading a passage or uttering a specific phrase, which produces speech recordings with particular qualities. Older theoretical ([68], [69]), as well as some recent works [81] deal with picking out spontaneous or posed behavior
- turn-based interaction results in associations between expressivity and stimulus: e.g. participants express anger at a particular cue or their computer freezing, thus catering for annotation of context, in addition to expressivity. An initial view of context can be the relation of an expressive utterance to answers to the W-questions (‘who?’, ‘when?’, ‘where?’, ‘why?’)

§ speech and facial expressivity hardly follow a specific pattern; in addition to this, some parameters may be stable, e.g. the *valence* of the observed emotion, while *activation* may change to indicate semantic emphasis. Most 'regular' databases provide samples of participants moving from neutral to highly expressive and back

Since the aim of this work is to emphasize on the ability to classify sequences with naturalistic expressions, we have chosen to utilize the SAL database for training and testing purposes [5]. Recordings were based on the notion of the "Sensitive Artificial Listener", where the SAL simulates what some interviewers and good listeners do, i.e. engages a willing person in emotionally colored interaction on the basis of stock lines keyed in a broad way to the speaker's emotions. Although the final goal is to let the SAL automatically assess the content of the interaction and select the line with which to respond, this had not yet been fully implemented at the time of the creation of the SAL database and thus a "Wizard of Oz" approach was used for the selection of the SAL's answers [45].

A point to consider in natural human interaction is that each individual's character has an important role on the human's emotional state; different individuals may have different emotional responses to similar stimuli. Therefore, the annotation of the recordings should not be based on the intended induced emotion but on the actual result of the interaction with the SAL. Towards this end, FeelTrace was used for the annotation of recordings in SAL [24]. This is a descriptive tool that has been developed at Queen's University Belfast using dimensional representations, which provides time-sensitive dimensional representations. It lets observers track the emotional content of a time-varying stimulus as they perceive it. Figure 1, illustrates the kind of display that FeelTrace users see.

The space is represented by a circle on a computer screen, split into four quadrants by the two main axes. The vertical axis represents activation, running from very active to very passive and the horizontal axis represents evaluation, running from very positive to very negative. It reflects the popular view that emotional space is roughly circular. The centre of the circle marks a sort of neutral default state, and putting the cursor in this area indicates that there is no real emotion being expressed. A user uses the mouse to move the cursor through the emotional space, so that its position signals the levels of activation and evaluation perceived by her/him, and the system automatically records the co-ordinates of the cursor at any time.

For reasons outlined in Section 2.1 the x-y coordinates of the mouse movements on the two-dimensional user interface are mapped to the five emotional categories presented in Table 1. Applying a standard pause detection algorithm on the audio channel of the recordings in examination, the database has been split into 477 tunes, with lengths ranging from 1 frame up to 174 frames. A bias towards Q1 exists in the database, as 42,98% of the tunes are classified to Q1, as shown in Table 4.

	Neutral	**Q1**	**Q2**	**Q3**	**Q4**	**Totals**
Tunes	47	205	90	63	72	477
Percentages	9,85%	42,98%	18,87%	13,21%	15,09%	100,00%

Table 4: Class distribution in the SAL dataset

5.2 Statistical results

From the application of the proposed methodology on the data set annotated as ground truth we acquire a measurement of 81,55% for the system's accuracy. Specifically, 389 tunes were classified correctly, while 88 were misclassified. Clearly, this kind of information, although indicative, is not sufficient to fully comprehend and assess the performance of our methodology.

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	**34**	1	5	3	0	43
Q1	1	**189**	9	12	6	217
Q2	4	3	**65**	2	1	75
Q3	4	6	7	**39**	3	59
Q4	4	6	4	7	**62**	83
Totals	47	205	90	63	72	477

Table 5: Overall confusion matrix

Towards this end, we provide in Table 5 the confusion matrix for the experiment. In the table rows correspond to the ground truth and columns to the system's response. Thus, for example, there were 5 tunes that were labeled as neutral in the ground truth but were misclassified as belonging to Q2 by our system.

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	**79,07%**	2,33%	11,63%	6,98%	0,00%	100,00%
Q1	0,46%	**87,10%**	4,15%	5,53%	2,76%	100,00%
Q2	5,33%	4,00%	**86,67%**	2,67%	1,33%	100,00%
Q3	6,78%	10,17%	11,86%	**66,10%**	5,08%	100,00%
Q4	4,82%	7,23%	4,82%	8,43%	**74,70%**	100,00%
Totals	9,85%	42,98%	18,87%	13,21%	15,09%	100,00%

Table 6: Overall confusion matrix expressed in percentages

Given the fact that our ground truth is biased towards Q1, we also provide in Table 6 the confusion matrix in the form of percentages so that the bias is removed from the numbers. There we can see that the proposed methodology performs reasonably well for most cases, with the exception of Q3, for which the classification rate is very low. What is more alarming is that more than 10% of the tunes of Q3 have been classified as belonging to the exactly opposite quadrant, which is certainly a major mistake.

Still, in our analysis of the experimental results so far we have not taken into consideration a very important factor: that of the length of the tunes. As we have explained in section 5, in order for the Elman net to pick up the expression dynamics of the tune an adequate number of frames needs to be available as input. Still, there is a number of tunes in the ground truth that are too short for the network to reach a point where its output can be read with high confidence.

In order to see how this may have influence our results we present in the following separate confusion matrices for short and normal length tunes. In this context we consider as normal tunes that comprise at least 10 frames and as short tunes with length from 1 up to 9 frames.

First of all, we can see right away that the performance of the system, as was expected is quite different in these two cases. Specifically, there are 83 errors in just 131 short tunes while there are only 5 errors in 346 normal tunes. Moreover, there are no severe errors in the case of long tunes, i.e. there are no cases in which a tune is classified in the exact opposite quadrant than in the ground truth.

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	**29**	0	0	0	0	29
Q1	0	**172**	3	0	0	175
Q2	1	1	**54**	0	0	56
Q3	0	0	0	**30**	0	30
Q4	0	0	0	0	**56**	56
Totals	30	173	57	30	56	346

Table 7: Confusion matrix for normal tunes

Overall, the operation of our system in normal operating conditions (as such we consider the case in which tunes have a length of at least 10 frames) is accompanied by a classification rate of 98,55%, which is certainly very high, even for controlled data, let alone for naturalistic data.

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	**100,00%**	0,00%	0,00%	0,00%	0,00%	100,00%
Q1	0,00%	**98,29%**	1,71%	0,00%	0,00%	100,00%
Q2	1,79%	1,79%	**96,43%**	0,00%	0,00%	100,00%
Q3	0,00%	0,00%	0,00%	**100,00%**	0,00%	100,00%
Q4	0,00%	0,00%	0,00%	0,00%	**100,00%**	100,00%
Totals	8,67%	50,00%	16,47%	8,67%	16,18%	100,00%

Table 8: Confusion matrix for normal tunes expressed in percentages

5.3 Quantitative comparative study

In a previous work we have proposed a different methodology to process naturalistic data with the goal of estimating the human's emotional state [3]. In that work a very similar approach is followed in the analysis of the visual component of the video with the aim of locating facial features. FAP values are then fed into a rule based system which provides a response concerning the human's emotional state.

In a later version of this work, we evaluate the likelihood of the detected regions being indeed the desired facial features with the help of anthropometric statistics acquired from [29] and produce degrees of confidence which are associated with the FAPs; the rule evaluation model is also altered and equipped with the ability to consider confidence degrees associated with each FAP in order to minimize the propagation of feature extraction errors in the overall result [4].

When compared to our current work, these systems have the extra advantages of

1. considering expert knowledge in the form of rules in the classification process

2. being able to cope with feature detection deficiencies and.

On the other hand, they are lacking in the sense that

3. they do not consider the dynamics of the displayed expression and
4. they do not consider other modalities besides the visual one.

Thus, they make excellent candidates to compare our current work against in order to evaluate the practical gain from the proposed dynamic and multimodal approach. In Table 11 we present the results from the two former and the current approach. Since dynamics are not considered, each frame is treated independently in the preexisting systems. Therefore, statistics are calculated by estimating the number of correctly classified frames; each frame is considered to belong to the same quadrant as the whole tune.

It is worth mentioning that the results are from the parts of the data set that were selected as expressive for each methodology. But, whilst for the current work this refers to 72,54% of the data set and the selection criterion is the length of the tune, in the previous works only about 20% of the frames was selected with a criterion of the clarity with which the expression is observed, since frames close to the beginning or the end of the tune are often too close to neutral to provide meaningful visual input to a system.

Methodology	Classification rate
Rule based	78,4%
Possibilistic rule based	65,1%
Dynamic and multimodal	98,55%

Table 9: Classification rates on parts of the naturalistic data set

Methodology	Classification rate
Rule based	27,8%
Possibilistic rule based	38,5%
Dynamic and multimodal	98,55%

Table 10: Classification rates on the naturalistic data set

5.4 Qualitative comparative study

As we have already mentioned, during the recent years we have seen a very large number of publications in the field of the estimation of human expression and/or emotion. Although the vast majority of these works is focused on the six universal expressions and use sequences where extreme expressions are posed by actors, it would be an omission if not even a qualitative comparison was made to the broader state of the art.

In Table 11 we present the classification rates reported in some of the most promising and well known works in the current state of the art. Certainly, it is not possible or fair to compare numbers directly, since they come from the application on different data sets. Still, it is possible to make qualitative comparisons base on the following information:

1. The Cohen2003 is a database collected of subjects that were instructed to display facial expressions corresponding to the six types of emotions. In the Cohn–Kanade database subjects were instructed by an experimenter to perform a series

of 23 facial displays that included single action units and combinations of action units.

2. In the MMI database subjects were asked to display 79 series of expressions that included either a single AU or a combination of a minimal number of AUs or a prototypic combination of AUs (such as in expressions of emotion). They were instructed by an expert (a FACS coder) on how to display the required facial expressions, and they were asked to include a short neutral state at the beginning and at the end of each expression. The subjects were asked to display the required expressions while minimizing out-of-plane head motions.

3. The original instruction given to the actors has been taken as the actual displayed expression in all abovementioned databases, which means that there is an underlying assumption is that there is no difference between natural and acted expression.

As we can see, what is common among the datasets most commonly used in the literature for the evaluation of facial expression and/or emotion recognition is that expressions are solicited and acted. As a result, they are generally displayed clearly and to their extremes. In the case of natural human interaction, on the other hand, expressions are typically more subtle and often different expressions are mixed. Also, the element of speech adds an important degree of deformation to facial features which is not associated with the displayed expression and can be misleading for an automated expression analysis system.

Consequently, we can argue that the fact that the performance of the proposed methodology when applied to a naturalistic dataset is comparable to the performance of other works in the state of the art when applied to acted sequences is an indication of its success. Additionally, we can observe that when extremely short tunes are removed from the data set the classification performance of the proposed approach exceeds 98%, which, in current standards, is very high for an emotion recognition system.

Methodology	**Classification rate**	**Data set**
TAN	83,31%	Cohen2003
Multi-level HMM	82,46%	Cohen2003
TAN	73,22%	Cohn–Kanade
PanticPatras2006	86,6%	MMI
Proposed methodology	81,55%	SAL Database
Proposed methodology	**98,55%**	Normal section of the SAL database

Table 11: Classification rates reported in the broader state of the art [27],[28]

6 Conclusions

In this work we have focused on the problem of human emotion recognition in the case of naturalistic, rather than acted and extreme, expressions. The main elements of our approach are that i) we use multiple algorithms for the extraction of the “difficult”

facial features in order to make the overall approach more robust to image processing errors, ii) we focus on the dynamics of facial expressions rather that on the exact facial deformations they are associated with, thus being able to handle sequences in which the interaction is natural or naturalistic rather than posed or extreme and iii) we follow a multimodal approach where audio and visual modalities are combined, thus enhancing both performance and stability of the system.

From a more technical point of view, our contributions include: i) A modified input layer that allows the Elman net to process both dynamic and static inputs at the same time. This is used to fuse the fundamentally different visual and audio inputs in order to provide for a truly multimodal classification scheme. ii) A modified output scheme that allows the Elman that integrates previous values, with value significance decreasing exponentially through time. This allows the network to display augmented stability. iii) a modified mixture of experts module that, additionally to characteristics drawn from the experts' input, can also draw information from the experts' output in order to drive the output mediation step. This is used in order to incorporate the results from the statistical anthropometric evaluation of the acquired masks in the operation of the output combiner module.

Practical application of our methodology in a ground truth data set of naturalistic sequences has given a performance of 98,55% for tunes that are long enough for dynamics to be able to be picked up in both the visual and the audio channel.

For our future work, we intend to further extend our work in multimodal naturalistic expression recognition by considering more modalities such as posture and gestures and by incorporating uncertainty measuring and handling modules in order to maximize the system's performance and stability in difficult and uncontrolled environments.

7 Acknowledgements

The authors would like to thank all collaborators within the Humaine Network of Excellence, as well as all those that participated in the SAL data collection and annotation process. This work has been funded by the FP6 Network of Excellence Humaine: Human-Machine Interaction Network on Emotion, http://www.emotion-research.net

8 References

[1] J. L. Elman, Finding structure in time. Cognitive Science, 14, 179-211, 1990.

[2] J. L. Elman, Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning, 7, 195-224, 1991.

[3] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, S. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy network, Special Issue on Emotion: Understanding & Recognition, Neural Networks, Elsevier, Volume 18, Issue 4, May 2005, Pages 423-435.

[4] M. Wallace, S. Ioannou, A. Raouzaiou, K. Karpouzis, S. Kollias, Dealing with Feature Uncertainty in Facial Expression Recognition Using Possibilistic Fuzzy Rule Evaluation, International Journal of Intelligent Systems Technologies and Applications, Volume 1, Number 3-4, 2006.
[5] http://emotion-research.net/toolbox/toolboxdatabase.2006-09-26.5667892524
[6] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall International, 1999.
[7] M. H. Yang, D. Kriegman, N. Ahuja, "Detecting Faces in Images: A Survey", PAMI, Vol.24(1), 2002, pp. 34-58.
[8] C. Papageorgiou, M. Oren and T. Poggio, A general framework for object detection. In international Conference on Computer Vision, 1998.
[9] P. Viola, M. Jones, Rapid Object Detection using a Boosted Cascade of Simple Features. Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on , Volume: 1 , 8-14 Dec. 2001 Pages:I-511 - I-518 vol.1
[10] I. Fasel, B. Fortenberry, and J. R. Movellan, A generative framework for real-time object detection and classification, Computer Vision and Image Understanding, Volume 98 , Issue 1 (April 2005), pp. 182 – 210.
[11] R. Fransens, Jan De Prins, SVM-based Nonparametric Discriminant Analysis, An Application to Face Detection, Ninth IEEE International Conference on Computer Vision Volume 2, October 13 - 16, 2003
[12] M. T. Hagan, and M. Menhaj, Training feedforward networks with the Marquardt algorithm, IEEE Trans. Neural Networks, vol. 5, no. 6, 1994, pp. 989-993.
[13] ERMIS, Emotionally Rich Man-machine Intelligent System IST-2000-29319 (http://www.image.ntua.gr/ermis)
[14] J. Ang, R. Dhilon, A. Krupski, E. Shriberg, and A. Stolcke, Prosody based automatic detection of annoyance and frustration in human computer dialog, in Proc. of ICSLP, 2002, pp. 2037-2040.
[15] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noeth, How to Find Trouble in Communication, Speech Communication, 40, pp. 117-143, 2003.
[16] L. Devillers, L. Vidrascu, Real-life Emotion Recognition Human-Human call center data with acoustic and lexical cues, Ch. Mueller, S. Schoetz (eds.), Speaker characterization, Springer-Verlag, 2007.
[17] A. Batliner, R. Huber, H. Niemann, E. Noeth, J. Spilker, K. Fischer, The Recognition of Emotion, In: Wahlster, W.: Verbmobil: Foundations of Speech-to-Speech Translations. New York, Berlin: Springer, 2000, pp. 122- 130
[18] H. Ai, D. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, A. Purandare, Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs, Proceedings of Interspeech ICSLP, Pittsburgh, PA., 2006.
[19] D. Neiberg, K. Elenius, I. Karlsson, K. Laskowski, Emotion Recognition in Spontaneous Speech, Proceedings of Fonetik 2006, pp. 101-104.
[20] R. Cowie and R. Cornelius, Describing the Emotional States that are Expressed in Speech, Speech Communications, 40(5-32), 2003
[21] W. Wundt, Grundzuge der Physiologischen Psychologie, vol. 2. Engelmann, Leipzig, 1903
[22] H. Schlosberg, A scale for judgment of facial expressions, Journal of Experimental Psychology, 29(497-510), 1954

[23] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, Emotion Recognition in Human-Computer Interaction, IEEE Signal Processing Magazine, January 2001
[24] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M. Schroeder, 'FeelTrace': An Instrument for recording Perceived Emotion in Real Time, Proceedings of ISCA Workshop on Speech and Emotion, pp 19-24, 2000
[25] ERMIS, Emotionally Rich Man-machine Intelligent System IST-2000-29319 http://www.image.ntua.gr/ermis
[26] A. Murat Tekalp, Joern Ostermann, Face and 2-D mesh animation in MPEG-4, Signal Processing: Image Communication 15 (2000) 387-421
[27] I. Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling, Computer Vision and Image Understanding 91 (2003) 160–187
[28] M. Pantic and I. Patras, 'Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences', IEEE Transactions on Systems, Man and Cybernetics - Part B, vol. 36, no. 2, pp. 433-449, April 2006
[29] J. W. Young, Head and Face Anthropometry of Adult U.S. Civilians, FAA Civil Aeromedical Institute, 1963-1993 (final report 1993)
[30] C.M Whissel, (1989) 'The dictionary of affect in language', in Plutchnik, R. and Kellerman, H. (Eds.): Emotion: Theory, Research and Experience: The Measurement of Emotions, Academic Press, New York, Vol. 4, pp.113–131.
[31] P. Oudeyer, The production and recognition of emotions in speech: features and algorithms. International Journal of Human Computer Interaction, 59(1-2):157-183, 2003.
[32] A. M. Schaefer, H. G. Zimmermann, Recurrent Neural Networks Are Universal Approximators, ICANN 2006, pp. 632-640.
[33] B. Hammer, P. Tino, Recurrent neural networks with small weights implement definite memory machines, Neural Computation 15(8), 1897-1929, 2003.
[34] P. Ekman and W. V. Friesen, The facial Action Coding System: A Technique for the Measurement of Facial Movement. San Francisco: Consulting Psychologists Press, 1978
[35] P. Ekman, Facial expression and Emotion. Am. Psychologist, Vol. 48 (1993) 384-392
[36] R. Bertolami,H. Bunke, Early feature stream integration versus decision level combination in a multiple classifier system for text line recognition, 18th International Conference on Pattern Recognition (ICPR'06)
[37] C. Tomasi, T. Kanade, Detection and Tracking of Point Features, Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.
[38] Ying-li Tian, Takeo Kanade, J. F. Cohn, Recognizing Action Units for Facial Expression Analysis, IEEE Transactions on PAMI, Vol.23, No.2, February 2001
[39] I.A. Essa and A.P. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 757-763, July 1997
[40] S. H. Leung, S. L. Wang SL, W. H. Lau, Lip image segmentation using fuzzy clustering incorporating an elliptic shape function, IEEE Trans. on Image Processing, vol.13, No.1, January 2004

[41] N. Sebe, M.S. Lew, I. Cohen, Y. Sun, T. Gevers, T.S. Huang, Authentic Facial Expression Analysis, International Conference on Automatic Face and Gesture Recognition (FG'04), Seoul, Korea, May 2004, pp. 517-522.
[42] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE PAMI 23 (6), 2001, pp. 681-685
[43] M. Pantic, L.J.M Rothkrantz, Expert system for automatic analysis of facial expressions, Image and Vision Computing vol 18, 2000, pp. 881-905.
[44] P. Mertens, The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. in B. Bel & I. Marlien (eds.), Proc. of Speech Prosody, Japan, 2004
[45] J.F. Kelley, Natural Language and computers: Six empirical steps for writing an easy-to-use computer application. Unpublished doctoral dissertation, The Johns Hopkins University, 1983.
[46] R. W. Picard, Affective Computing, MIT Press, 1997.
[47] A. Jaimes, Human-Centered Multimedia: Culture, Deployment, and Access, IEEE Multimedia Magazine, Vol. 13, No.1, 2006.
[48] A. Pentland, Socially Aware Computation and Communication, Computer, vol. 38, no. 3, pp. 33-40, 2005.
[49] R. W. Picard, Towards computers that recognize and respond to user emotion, IBM Syst. Journal, 39 (3–4), 705–719, 2000.
[50] S. Oviatt, Ten myths of multimodal interaction, Communications of the ACM, Volume 42, Number 11 (1999), Pages 74-81.
[51] S. Oviatt, A. DeAngeli, K. Kuhn, Integration and synchronization of input modes during multimodal human-computer interaction, In Proceedings of Conf. Human Factors in Computing Systems CHI'97, ACM Press, NY, 1997, pp. 415 – 422.
[52] A. Mehrabian, Communication without words, Psychology Today, vol. 2, no. 4, pp. 53–56, 1968.
[53] Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, S. Levinson, Audio-Visual Affect Recognition, IEEE Trans. Multimedia, vol. 9, no. 2, Feb. 2007.
[54] A. Jaimes and N. Sebe, Multimodal Human Computer Interaction: A Survey, IEEE International Workshop on Human Computer Interaction, ICCV 2005, Beijing, China.
[55] P. Cohen, Multimodal Interaction: A new focal area for AI. IJCAI 2001, pp. 1467-1473.
[56] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Clow, I. Smith, The efficiency of multimodal interaction: A case study, in Proceedings of International Conference on Spoken Language Processing, ICSLP'98, Australia, 1998.
[57] M. Pantic and L.J.M. Rothkrantz, Towards an affect-sensitive multimodal human-computer interaction, Proc. of the IEEE, vol. 91, no. 9, pp. 1370-1390, 2003.
[58] L.C. De Silva and P.C. Ng, Bimodal emotion recognition, in Proc. Face and Gesture Recognition Conf., 332-335, 2000.
[59] L. Chen and T. S. Huang, Emotional expressions in audiovisual human computer interaction, in Proc. Int. Conf. Multimedia Expo, 2000, pp. 423–426.
[60] L. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, Multimodal human emotion/expression recognition, in Int. Conf. Automatic Face Gesture Recognition, 1998, pp. 396–401.

[61] L. C. De Silva and P. C. Ng, Bimodal emotion recognition, in Proc. Int. Conf. Automatic Face Gesture Recognition, 2000, pp. 332–335.
[62] Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in Proc. ROMAN, 2000, pp. 178–183.
[63] A. Kapoor, R. W. Picard and Y. Ivanov, Probabilistic combination of multiple modalities to detect interest, Proc. of IEEE ICPR, 2004.
[64] H. Gunes and M. Piccardi, Fusing Face and Body Gesture for Machine Recognition of Emotions, 2005 IEEE International Workshop on Robots and Human Interactive Communication, pp. 306 – 311.
[65] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, K. Karpouzis, Modeling naturalistic affective states via facial and vocal expressions recognition, International Conference on Multimodal Interfaces (ICMI'06), Banff, Alberta, Canada, November 2-4, 2006, pp. 146 - 154.
[66] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir,A. Raouzaiou, L. Malatesta, S. Kollias, Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition, in T. Huang, A. Nijholt, M. Pantic, A. Pentland (eds.), AI for Human Computing, LNAI Volume 4451/2007, Springer.
[67] N. Fragopanagos and J. G. Taylor, Emotion recognition in human computer interaction, Neural Networks, vol. 18, pp. 389–405, 2005.
[68] P. Ekman and W. V. Friesen. Felt, false, and miserable smiles. Journal of Nonverbal Behavior, 6:238–252, 1982.
[69] M. G. Frank and P. Ekman. Not all smiles are created equal: The differences between enjoyment and other smiles. Humor: The International Journal for Research in Humor, 6:9–26, 1993.
[70] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image and Vision Computing 24 (2006) 615–625.
[71] M. Pantic, Face for interface, in The Encyclopedia of Multimedia Technology and Networking, M. Pagani, Ed. Hershey, PA: Idea Group Reference, 2005, vol. 1, pp. 308–314.
[72] L. Wu, Sharon L. Oviatt, Philip R. Cohen, Multimodal Integration – A Statistical View, IEEE Transactions on Multimedia, vol. 1, no. 4, December 1999.
[73] P. Teissier, J. Robert-Ribes, and J. L. Schwartz, Comparing models for audiovisual fusion in a noisy-vowel recognition task, IEEE Trans. Speech Audio Processing, vol. 7, pp. 629–642, Nov. 1999.
[74] A. Rogozan, Discriminative learning of visual data for audiovisual speech recognition, Int. J. Artif. Intell. Tools, vol. 8, pp. 43–52, 1999.
[75] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder,'Feeltrace': An instrument for recording perceived emotion in real time, in Proc. ISCAWorkshop on Speech and Emotion, 2000, pp. 19–24.
[76] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image and Vision Computing 24 (2006) 615–625.
[77] P. Ekman and W. V. Friesen. Felt, false, and miserable smiles. Journal of Nonverbal Behavior, 6:238–252, 1982.

[78] M. G. Frank and P. Ekman. Not all smiles are created equal: The differences between enjoyment and other smiles. Humor: The International Journal for Research in Humor, 6:9–26, 1993.
[79] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: Towards a new generation of databases, Speech Communication 40, pp. 33–60, 2003.
[80] R. Banse, K. Scherer, K., Acoustic profiles in vocal emotion expression. J. Pers. Social Psychol. 70 (3), 614–636, 1996.
[81] M.F. Valstar, M. Pantic, Z. Ambadar and J.F. Cohn, 'Spontaneous vs. posed facial behavior: Automatic analysis of brow actions (pdf file)', in Proceedings of ACM Int'l Conf. Multimodal Interfaces (ICMI'06), pp. 162-170, Banff, Canada, November 2006.
[82] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, A. Stolcke, Prosody-based automatic detection of annoyance and frustration in human-computer dialog. Proc. ICSLP 2002, Denver, Colorado, Sept. 2002.
[83] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Noeth, How to find trouble in communication. Speech Communication 40, pp. 117-143, 2003.
[84] J.H.L. Hansen, B.D. Womack, Feature analysis and neural network-based classification of speech under stress. IEEE Transactions on Speech and Audio Processing IV(4), 307- 313, 1996.
[85] J. Hansen, S. Bou-Ghazale, Getting started with SUSAS: A Speech Under Simulated and Actual Stress Database. Proc. Eurospeech 1997, Rhodes, Greece, vol. 5, pp. 2387-2390, 1997.
[86] O. Kwon, K. Chan, J. Hao, Emotion recognition by speech signals. Proc. Eurospeech 2003, Geneva, pp. 125-128, 2003.
[87] C. Lee, S. Narayanan, Emotion recognition using a data-driven fuzzy inference system. Proc. Eurospeech 2003, Geneva.
[88] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, S. Stroeve, Automatic recognition of emotion from voice: a rough benchmark. Proc. ISCA ITRW on Speech and Emotion, 5-7 September 2000, Textflow, Belfast, pp. 207-212, 2000.
[89] R. Nakatsu, N. Tosa, J. Nicholson, Emotion recognition and its application to computer agents with spontaneous interactive capabilities. Proc. IEEE International Workshop on Multimedia Signal Processing, pp. 439-444, 1999..
[90] S. Yacoub, S. Simske, X. Lin, J. Burns, Recognition of emotions in interactive voice response systems, in Proeedings of Eurospeech 2003, Geneva, 2003.
[91] G. Zhou, J. Hansen, J. Kaiser, Methods for stress classification: Nonlinear TEO and linear speech based features. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. IV, pp. 2087-2090, 1999.
[92] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson, Combining efforts for improving automatic classification of emotional user states, In Erjavec, T. and Gros, J. (Ed.), Language Technologies, IS-LTC 2006, pp. 240-245, Ljubljana, Slovenia, 2006.