Emotion Modelling and Facial Affect Recognition in Human-Computer and Human-Robot Interaction

Lori Malatesta[1], John Murray[2], Amaryllis Raouzaiou[1], Antoine Hiolle[2],
Lola Cañamero[2] and Kostas Karpouzis[1]
[1]*Image, Video and Multimedia Systems Lab, National Technical University of Athens,*
[2]*Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire,*
[1]*Greece*
[2]*UK*

1. Introduction

As research has revealed the deep role that emotion and emotion expression play in human social interaction, researchers in human-computer interaction have proposed that more effective human-computer interfaces can be realized if the interface models the user's emotion as well as expresses emotions. Affective computing was defined by Rosalind Picard (1997) as computing that relates to, arises from, or deliberately influences emotion or other affective phenomena. According to Picard's pioneering book, if we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, and even to have and express emotions. These positions have become the foundations of research in the area and have been investigated in great depth after their first postulation.

Emotion is fundamental to human experience, influencing cognition, perception, and everyday tasks such as learning, communication, and even rational decision making. Affective computing aspires to bridge the gap that typical human-computer interaction largely ignored thus creating an often frustrating experience for people, in part because affect had been overlooked or hard to measure.

In order to take these ideas a step further, towards the objectives of practical applications, we need to adapt methods of modelling affect to the requirements of particularshowcases. To do so, it is fundamental to review prevalent psychology theories on emotion, to disambiguate their terminology and identify the fitting computational models that can allow for affective interactions in the desired environments.

1.1 Applications of affective computing

Affective computing deals with the design of systems and devices that can recognize, interpret, generate, and process emotions. We are going to fledge out the potentials this research domain can provide in the field of new media applications and identify the

Source: State of the Art in Face Recognition, Book edited by: Dr. Mario I. Chacon M.,
ISBN -3-902613-42-4, pp. 250, January 2009, I-Tech, Vienna, Austria

matching theoretical background that will act as a tool for effectively modelling emotional interaction in such environments.

1.1.1 Detecting and recognizing emotional information

Detecting emotional information usually involves passive sensors, which capture data about the user's physical state or behaviour. The data gathered is often analogous to the cues humans use to perceive emotions in others. For example, a video camera might capture facial expressions, body posture and gestures, while a microphone might capture speech. Other sensors detect emotional cues by directly measuring physiological data such as skin temperature and galvanic resistance.

Recognizing emotional information requires the extraction of meaningful patterns from the gathered data. This is done by parsing the data through various processes such as facial expression detection, gesture recognition, speech recognition, or natural language processing.

1.1.2 Emotion in machines

By the term “emotion in machines” we refer to the simulation of emotions using computers or robots. The goal of such simulation is to enrich and facilitate interactivity between human and machine. The most common and probably most complex application of this simulation lies in the field of conversational agents. A related area is that of affective robot companions, although their expressive and communicative capabilities are usually simpler that those of conversational agents, due to a large extent to other constraints and research challenges related to their embodiment. Such a simulation is closely coupled with emotional understanding and modelling as explained below. This being said, it is important to mention that less sophisticated simulation approaches often produce surprisingly engaging experiences in the area of new media. It is often the case that our aim is not to fully simulate human behaviour and emotional responses but to simply depict emotion in a pseudo intelligent way that makes sense in the specific context of interaction.

1.1.3 Emotional understanding

Emotional understanding refers to the ability of a device not only to detect emotional or affective information, but also to store, process, build and maintain an emotional model of the user. This is not to be mistaken for the related term “emotion understanding”, which is used to refer to the use of robotic and computer simulations as tools and models to investigate research hypotheses contributing to the understanding of human and animal emotions (Cañamero, 2005; Cañamero & Gaussier, 2005). The goal of emotional understanding is to understand contextual information about the user and her environment, and formulate an appropriate response. This is difficult because human emotions arise from complex external and internal contexts.

Possible features of a system that displays emotional understanding might be adaptive behaviour, for example, avoiding interaction with a user it perceives to be angry. In the case of affect-aware applications, emotional understanding makes sense in tracking the user’s emotional state and adapting environment variables according to the state recognised. Questions regarding the level of detail of the tracking performed, the theoretical grounds for the analysis of the data collected, and the types of potential output that would make sense for such an interactive process are paramount.

2. Affect-related concepts

A lot of confusion exists regarding emotion research terminology, and not without a reason. Different definitions of the role and nature of emotions arise from different scientific approaches since emotion research is typically multidisciplinary. Different disciplines (i.e. psychology, cognitive neuroscience, etc) provide theories and corresponding models that are based on diverse underlying assumptions, are based on different levels of abstraction and may even have different research goals altogether.

The actual definition of *emotions* largely remains an open question: some define it as the physiological changes caused in our body, while the others treat it as purely intellectual thought processes. In psychology research, the term 'affect' is very broad (Rusting, 1998), and has been used to cover a wide variety of experiences such as emotions, moods, and preferences. In contrast, the term 'emotion' tends to be used to refer to fairly brief but intense experiences, although it is also used in a broader sense. Finally, moods or states describe low-intensity but more prolonged experiences.

From a cognitive neuroscience point of view, Damasio (2003) makes a distinction between emotions, which are publicly observable body states, and feelings, which are mental events observable only to the person having them. Based on neuroscience research he and others have done, Damasio argues that an episode of emotion begins with an emotionally "competent" stimulus (such as an attractive person or a scary house) that the organism automatically appraises as conducive to survival or well-being (a good thing) or not conducive (bad). This appraisal takes the form of a complex array of physiological reactions (e.g., quickening heartbeat, tensing facial muscles), which is mapped in the brain. From that map, a feeling arises as "an idea of the body when it is perturbed by the emoting process".

It is apparent that there is no right or wrong approach, and an attempt on a full terminology disambiguation would not be possible without biasing our choices towards one theory over the other. This is to make the point that the context of each approach has to be carefully defined. Next we are going to enumerate core elements of emotion and ways to distinguish them from other affective phenomena. This will lead us to a short description of the directions of affective computing. Subsequently we will put forward the most prevalent psychological theories of emotion along with corresponding computational modelling approaches and couple them to the affective computing goals and more specifically to the goals of practical applications.

2.1 Defining 'emotion' and 'feeling'

Emotion, according to Scherer (1987, 2001), can be defined as an episode of interrelated, synchronized changes in the states of all or most of five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism. The components of an emotion episode are the particular states of the subsystems mentioned. The process consists of the coordinated changes over time.

Most current psychological theories postulate that

- Subjective experience
- Peripheral physiological response patterns, and
- Motor expression

are major components of emotion. These three components have often been called the *emotional response triad*. Some theorists include the cognitive and motivational domains as components of the emotion process. The elicitation of action tendencies and the preparation of action have also been implicitly associated with emotional arousal. However, only after explicit inclusion of motivational consequences in theories (and Frijda's forceful claim for the emotion-differentiating function of action tendencies, see (Frijda, 1986)), have these important features of emotion acquired the status of a major component. The inclusion of a cognitive information-processing component has met with less consensus. Many theorists still prefer to see emotion and cognition as two independent but interacting systems. However, one can argue that all subsystems underlying emotion components function independently much of the time, and that the special nature of emotion as a hypothetical construct consists of the coordination and synchronization of all these systems during an emotion episode (Scherer, 2004).

How can emotions, as defined above, be distinguished from other affective phenomena such as feelings, moods, or attitudes? Let us take the term 'feeling' first. Scherer aligns feeling with the "subjective emotional experience" component of emotion, thus reflecting the total pattern of cognitive appraisal as well as motivational and somatic response patterning that underlie the subjective experience of an emotion. If we use the term 'feeling', a single component denoting subjective experience process, as a synonym for 'emotion' (the total multi-modal component process), this is likely to produce serious confusion and hamper our understanding of the phenomenon.

If we accept feeling as one of emotions' components, then the next step is to differentiate emotion from other types of affective phenomena. Instances of these phenomena, which can vary in degree of affectivity, are often called "emotions" in the literature. There are five such types of affective phenomena that should be distinguished from emotion: *preferences*, *attitudes*, *moods*, *affective dispositions* and *interpersonal stances*.

2.2 Emotions in applied intelligence

Having distinguished emotions against other types of affective phenomena, it is now of particular interest, in regard to the new media domain, to present a suggested distinction on a different level. Scherer (2004) questioned the need to distinguish between two different types of emotion: (1) *aesthetic* emotions (2) *utilitarian* emotions. The latter correspond to the "garden variety" of emotions usually studied in emotion research, such as anger, fear, joy, disgust, sadness, shame, guilt. These types of emotions can be considered utilitarian in the sense of facilitating our adaptation to events that have important consequences for our well-being. Such adaptive functions are the preparation of action tendencies (fight, flight), recovery and reorientation (grief, work), motivational enhancement (joy, pride), or the creation of social obligations (reparation). Because of their importance for survival and well-being, many utilitarian emotions are high-intensity emergency reactions, involving the synchronization of many subsystems, as described earlier. In the case of aesthetic emotions, adaptation to an event that requires the appraisal of goal relevance and coping potential is absent, or much less pronounced. Kant defined aesthetic experience as "disinterested pleasure" (Kant, 1790), highlighting the complete absence of utilitarian considerations. Thus, *my* aesthetic experience of a work of art or a piece of music is not shaped by the appraisal of the work's ability to satisfy *my* bodily needs, further *my* current goals or plans, or correspond to *my* social values. Rather, aesthetic emotions are produced by the appreciation of the intrinsic qualities of a work of art or an artistic performance, or the beauty of nature. Examples of such aesthetic emotions are: being moved or awed, full of wonder, admiration, bliss, ecstasy, fascination, harmony, rapture, solemnity.

This differentiation of emotions has an impact on the way an appraisal-based modelling approach would be implemented. It would not make sense to try and model all the proposed components of an appraisal process in cases where only aesthetic emotions are expected. On the other hand it would make sense to provide a deeper model in cases where anger or frustration are common emotional states such as in the example of an interactive Television environment.

2.3 Emotion representation

When it comes to machine-based recognition of emotions, 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 (Ekman, 1978, 1993).

On the contrary, psychology researchers have extensively investigated a broader variety of emotions. An extensive survey on emotion analysis can be found in (Cowie, 2001). 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 (Averill, 1975) 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 (1903) original proposal to Schlossberg's reintroduction of the idea in the modern era (Schlosberg, 1954). 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 (Cowie, 2001). 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 onwards 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.

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.

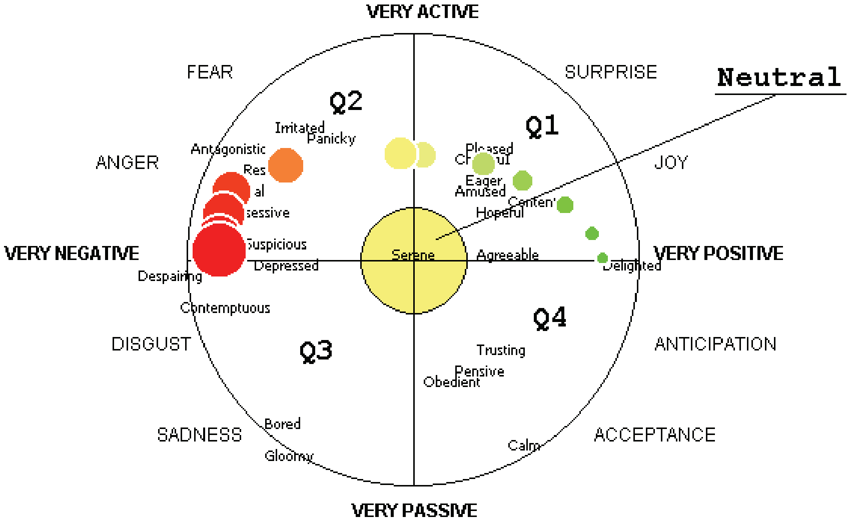

Fig. 1. The activation/valence dimensional representation, after (Whissel, 1989)

As we can see in Figure 1, 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 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 (Cowie, 2000) 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

2.4 Other emotion representation models

Having reviewed the areas of affective computing, it is time to start focusing on the available theories, descriptions and models that can support these goals. We start with reviewing the main groups of emotion descriptions as identified by the members of the Humaine Network of Excellence (Humaine, 2008). It is important to stress the difference that exists between emotion models and emotion descriptions. By 'emotion descriptions' we refer to different ways of representing emotions and their underlying psychological theories, whereas with the term 'emotional models' we talk about the computational modelling of these theories in specific context.

2.4.1 Categorical representations

Categorical representations — using a word to describe an emotional state — are the simplest and most widespread. Such category sets have been proposed on different grounds, including evolutionarily basic emotion categories; most frequent everyday emotions; application-specific emotion sets; or categories describing other affective states, such as moods or interpersonal stances (Feeltrace core vocabulary in Cowie, 1999; Ortony, Clore and Collins list of emotion words in Ortony, 1988; Ekman's list of six basic emotions in Ekman, 1993).

2.4.2 Other dimensional descriptions

Dimensional descriptions capture essential properties of emotional states, such as arousal (active/passive) and valence (negative/positive). Emotion dimensions can be used to describe general emotional tendencies, including low-intensity emotions. In addition to these two, there are a number of other possible dimensions, such as power, control, or approach/avoidance, which add some refinement. The most obvious is the ability to distinguish between fear and anger, both of which involve negative valence and high activation. In anger, the subject of the emotion feels that he or she is in control; in fear, control is felt to lie elsewhere.

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. A further attraction is the fact that dimensional descriptions can be translated into and out of verbal descriptions. This is possible because emotion words can, to an extent, be understood as referring to positions in activation-evaluation space.

2.4.3 Appraisal theories and representations

Appraisal theories focus on the emotion elicitation process in contrast with the previously mentioned approaches that emphasize on the consequences/symptoms of an emotional episode. Appraisal representations characterise emotional states in terms of the detailed evaluations of eliciting conditions, such as their familiarity, intrinsic pleasantness, or relevance to one's goals. Such detail can be used to characterise the cause or object of an emotion as it arises from the context, or to predict emotions in AI systems (Lazarus, 1991, Scherer, 1987, Frijda, 1986).

Appraisal theories are very common in emotion modelling since their structure caters for simulating their postulations in computational models. Moreover, it is often the case that an appraisal theory was formulated explicitly in order to be implemented in a computer. Such an example is the OCC theory (Ortony, 1988). This is sometimes a source of confusion, since the underlying emotion theory is unavoidably very closely linked with the actual modelling approach.

According to cognitive theories of emotion (Lazarus, 1987), emotions are closely related to the situation that is being experienced (or, indeed, imagined) by the agent.

3. Facial expression recognition

3.1 Feature representation

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 (Tomasi, 1991), relatively little work has been reported (Tian, 2001) 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 (Ekman, 1978) for describing facial expressions. FACS describes expressions using 66 Action Units (AU) which relate to the contractions of specific facial muscles.

Additionally to FACS, MPEG-4 metrics (Tekalp, 2000) are commonly used to model facial expressions and underlying emotions. They define an alternative way of modelling 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.

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 (Tian, 2001) depends on optical flow, (Leung, 2004) depends on high resolution or noise-free input video, (Sebe, 2004) depends on colour information, (Cootes, 2001) requires two head-mounted cameras and (Pantic, 2000) 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.

3.2 Facial feature extraction

Facial feature extraction is a crucial step to numerous applications such us face recognition, human-computer interaction, facial expression recognition, surveillance and gaze/pose detection (Asteriadis, 2007). In their vast majority, the approaches in the bibliography use face detection as a pre-processing step. This is usually necessary in order to tackle with scale problems, as localizing a face in an image is more scale-independent than starting with the localization of special facial features. When only facial features are detected (starting from the whole image and not from the face region of interest), the size and the position of the face in the image have to be pre-determined and, thus, such algorithms are devoted to special cases, such as driver's attention recognition (Smith, 2003) where the user's position with regards to a camera is almost stable. In such techniques, colour (Smith, 2003) predicates, shape of facial features and their geometrical relations (D' Orazio, 2004) are used as criteria for the extraction of facial characteristics.

On the other side, facial features detection is more scale-independent when the face is detected as a pre-processing step. In this case, the face region of interest can be normalized to certain dimensions, thus making the task of facial feature detection more robust. For example, in (Cristinacce, 2004) a multi-stage approach is used to locate features on a face. First, the face is detected using the boosted cascaded classifier algorithm by Viola and Jones (Viola, 2001). The same classifier is trained using facial feature patches to detect facial features. A novel shape constraint, the Pairwise Reinforcement of Feature Responses (PRFR) is used to improve the localization accuracy of the detected features. In (Jerosky, 2001), a three-stage technique is used for eye centre localization. The Hausdorff distance between edges of the image and an edge model of the face is used to detect the face area. At the second stage, the Hausdorff distance between the image edges and a more refined model of the area around the eyes is used for more accurate localization of the upper area of the head.

Finally, a Multi-Layer Perceptron (MLP) is used for finding the exact pupil locations. In (Viola, 2001), an SVM-based approach is used for face detection. Following, eye-areas are located using a feed-forward neural network and the face is brought to a horizontal position based on the eye positions. Starting from these points, edge information and luminance values, are used for eyebrow and nostrils detection. Further masks are created to refine the eye positions, based on edge, luminance and morphological operations. Similar approaches are followed for the detection of mouth points.

In this work, prior to eye and mouth region detection, face detection is applied on the face images. The face is detected using the Boosted Cascade method, described in (Viola, 2001). The output of this method is usually the face region with some background. Furthermore, the position of the face is often not centred in the detected sub-image. Since the detection of the eyes and mouth will be done on blocks of a predefined size, it is very important to have an accurate face detection algorithm. Consequently, a technique to post-process the results of the face detector is used.

More specifically, a technique that compares the shape of a face with that of an ellipse is used (Asteriadis, 2007). According to this, the distance map of the face area found at the first step is extracted. Here, the distance map is calculated from the binary edge map of the area. An ellipsis scans the distance map and a score that is the average of all distance map values on the ellipse contour el, is evaluated.

$$score = \frac{1}{el} \sum_{(x,y)\in el} D(x,y) \tag{1}$$

where D is the distance map of the region found by the Boosted Cascade algorithm. This score is calculated for various scale and shape transformations of the ellipses. The transformation which gives the best score is considered as the one that corresponds to the ellipses that best describes the exact face contour. The lateral boundaries of the ellipses are the new boundaries of the face region.

A template matching technique follows for the facial feature area detection step: The face region found by the face detection step is brought to certain dimensions and the corresponding Canny edge map is extracted. Subsequently, for each pixel on the edge map, a vector pointing to the closest edge is calculated and its *x, y* coordinates are stored. The final result is a vector field encoding the geometry of the face. Prototype eye patches were used for the calculation of their corresponding vector fields and the mean vector field was used as prototype for searching similar vector fields on areas of specified dimensions on the face vector field. The similarity between an image region and the templates is based on the following distance measure:

$$E_{L_2} = \sum_{i\in R_k} \| v_i - m_i \| \tag{2}$$

where $\|\|$denotes the L_2 norm. Essentially for a *NxM* region R_k the previous formula is the sum of the Euclidean distances between vectors v_i of the candidate region and the corresponding mi of the mean vector field (template) of the eye we are searching for (right or left). The candidate region on the face that minimizes E_{L2} is marked as the region of the left or right eye. To make the algorithm faster we utilize the knowledge of the approximate positions of eyes on a face.

For eye centre detection, the normalized area of the eye is brought back to its initial dimensions on the image and a light reflection removal step is employed. The grayscale image of the eye area is converted to a binary image and small white connected components are removed. The areas that correspond to such components on the original image are substituted by the average of their surrounding area. The final result is an eye area having reflections removed. Subsequently, horizontal and vertical derivative maps are extracted from the resulting image and they are projected on the vertical and horizontal axis respectively. The mean of a set of the largest projections is used for an estimate of the eye centre. Following, a small window around the detected point is used for the darkest patch to be detected, and its centre is considered as the refined position of the eye centre.

For the detection of the eye corners (left, right, upper and lower) a technique similar to that described in (Ioannou, 2007) is used: Having found the eye centre, a small area around it is used for the rest of the points to be detected. This is done by using the Generalized Projection Functions (GPFs), which are a combination of the Integral Projection Functions (IPFs) and the Variance Projection Functions (VPFs). The integral projection function's value on row (column) x (y) is the mean of its luminance intensity, while the Variance Projection Function on row x is its mean variance. The GPF's value on a row (column) x (y) is a linear combination of the corresponding values of the derivatives of the IPF and VPF on row x (column y):

$$\begin{aligned} GPF_u(x) &= (1-a) * IPF_u^{'}(x) + a * VPF_u^{'} \\ GPF_v(y) &= (1-a) * IPF_v^{'}(y) + a * VPF_v^{'} \end{aligned} \tag{3}$$

Local maxima of the above functions are used to declare the positions of the eye boundaries.

For the mouth area localization, a similar approach to that of the eye area localization is used: The vector field of the face is used and template images are used for the extraction of a prototype vector field of the mouth area. Subsequently, similar vector fields are searched for on the lower part of the normalized face image. However, as the mouth has, many times,

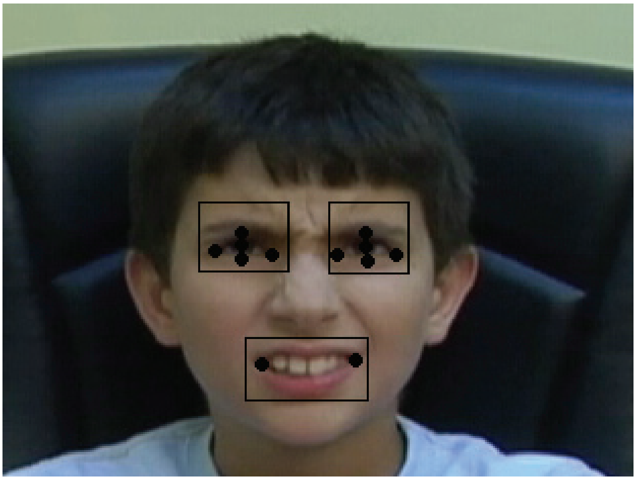

Fig. 2. Detected facial features

similar luminance values with its surrounding skin, an extra factor is also taken into account. That is, at every search area, the mean value of the hue component is calculated and added to the inverse distance from the mean vector fields of the mouth. Minimum values declare mouth existence.

For the extraction of the mouth points of interest (mouth corners), the hue component is also used. Based on the hue values of the mouth, the detected mouth area is binarised and small connected components whose value is close to 0° are discarded similar to the light reflection removal technique employed for the eyes. The remainder is the largest connected component, which is considered as the mouth area. The leftmost and rightmost points of this area are considered as the mouth corners. An example of detected feature points is shown in Figure 2.

4. Expression classification

4.1 Recognizing dynamics

In order to consider the dynamics of displayed expressions, one needs to utilize a classification model that is able to model and learn dynamics, such as a Hidden Markov Model or a recurrent neural network (see Figure 3). 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.

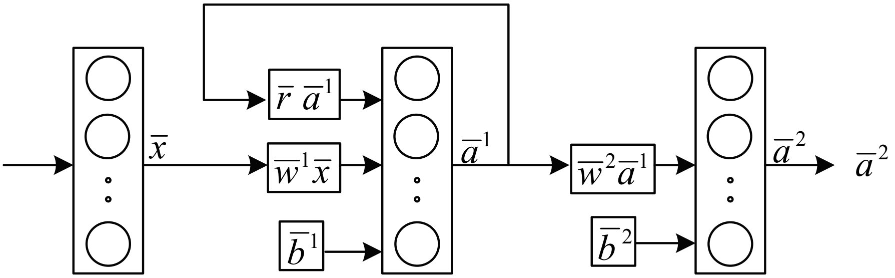

Fig. 3. A recurrent neural network (Elman, 1990, 1991)

Among other implementations of recurrent networks, the Elman net (Elman, 1990, 1991) is the most popular. 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

$$a_i^1 = \tan sig(k_i^1) = \frac{2}{1+\mathrm{e}^{-2k_i^1}} - 1\,,\ a_j^2 = k_j^2 \tag{4}$$

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:

$$k_i^1 = \overline{w}_i^1 \cdot \overline{x} + \overline{r}_i \cdot \overline{a}^1 + b_i^1 \tag{5}$$

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:

$$k_j^2 = \overline{w}_j^2 \cdot \overline{a}^1 + b_j^2 \tag{6}$$

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 (Schaefer, 1996, Hammer, 2003).
As far as training is concerned, the truncated back-propagation through time (truncated BPTT) algorithm is used (Haykin, 1999).
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 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 (Cowie, 2001) and levels of 0 at the other outputs.

4.2 Classification

The most common applications of recurrent neural networks include complex tasks such as modelling, 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 behaviour 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 4 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.

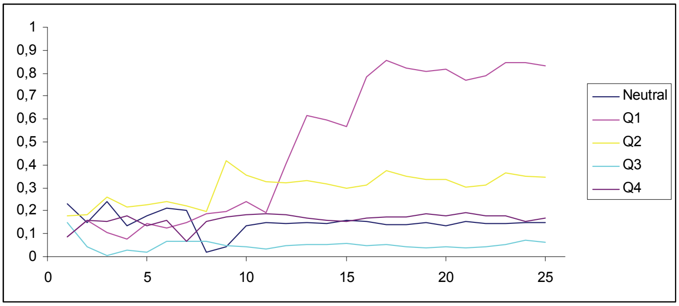

Fig. 4. Individual network outputs after each frame (Caridakis, 2006)

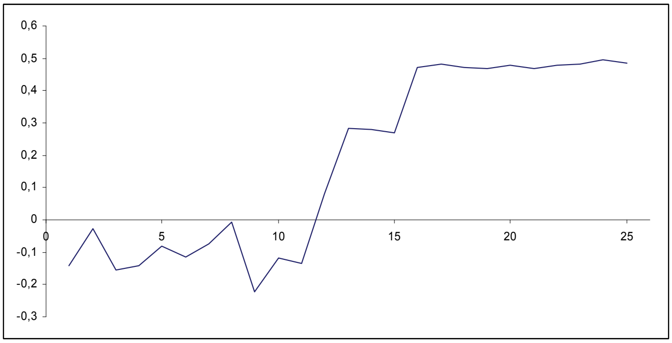

Fig. 5. Margin between correct and next best output

Although these are not enough to change the classification decision (see Figure 5) 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.

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:

$$o_j(t) = c \cdot a_j^2 + (1-c) \cdot o_j(t-1) \tag{7}$$

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 $c = 0.5$.

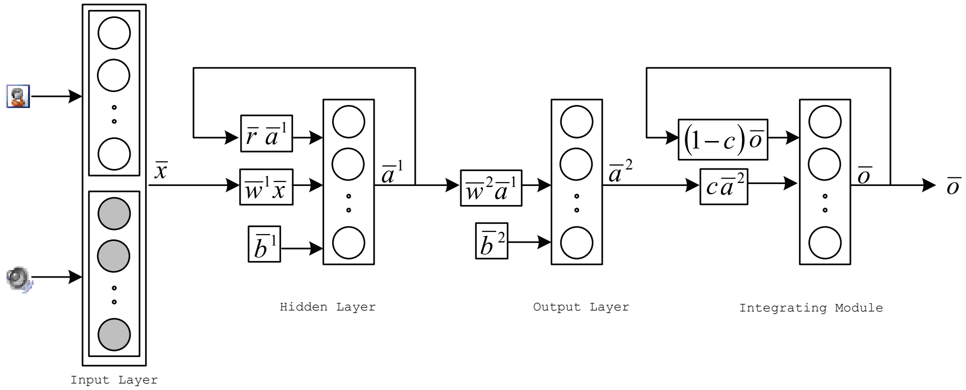

Fig. 6. The Elman net with the output integrator (Cowie, 2008)

In Figure 5, we can see the decision margin when using the weighting integration module at the output of the network. When comparing to Figure 7, we can clearly see that the progress of the margin is smoother, 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.

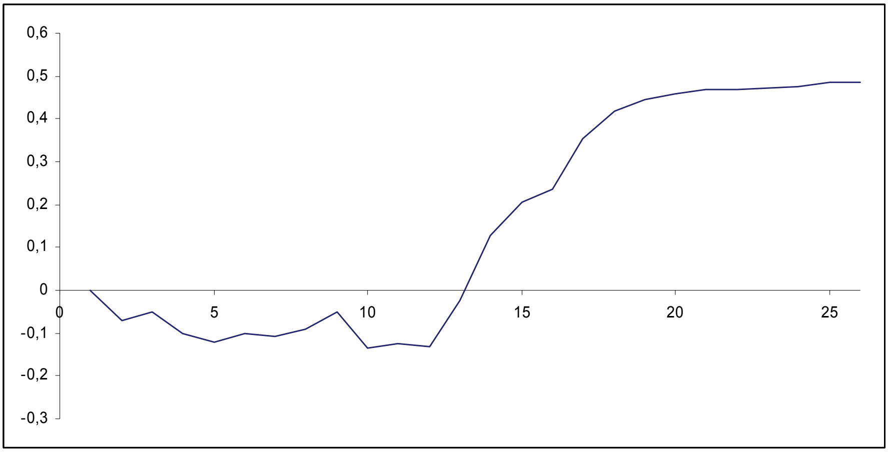

Fig. 7. Decision margin when using the integrator

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

$$\overline{o}\left(0\right)=0 \tag{8}$$

Another meaningful alternative would be to initialize $\overline{o}\left(0\right)$ 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.

It is worth noting that, from a modelling 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 behaviour of the network, as an additional recurrent loop would greatly augment the training time and size and average length of training data required.

5. Application in human-robot interaction

The question of how autonomous robots could be part of our everyday life is of a growing interest. Technology is now at a stage where in the very near future it will be possible for the majority of households to have their own robots, helping with a variety of tasks and even entertaining their owner. However, the problem of deciding on what kind of architectures are going to be embedded in such robots is still not completely answered. Indeed, these architectures require a set of properties that are not yet found in the currently available ones. The robots have to be adaptive in the complex and dynamic environment that we live in. To design such an ideal robot, it is argued that taking an epigenetic approach would be a suited solution (Cañamero et al., 2006).

5.1 Attachment bonds and emotional development

Taking inspiration from psychology, modelling the development of attachment bonds between a robot and its owner or user could offer a promising avenue to improve human-robot interactions. These phenomena would help a robot initiate interactions with the humans in more natural way, without any explicit teaching signal. The interactions and the robot's behaviour would be modulated using the feedback from the current emotional state of the user(s). To that end, an autonomous robot would need to use information from the user's current emotional state, extracting this information from the facial expressions.

A simple application of these idea could be to start to improve already existing models of the imprinting phenomenon. Imprinting was documented by Konrad Lorenz (Lorenz, 1935) and described as the tendency of young birds to follow the first moving object or person seen after hatching. This process is a product of evolution helping the survival of these species which don't have a nest to protect and hide their offspring. A successful architecture has been developed to allow a Koala robot to discover what object or person to follow, and then learn how to follow him/her (Hiolle & Cañamero, 2007) . The robot used the distance between it and the human as a "desired perception" to keep constant, as would be the sight of the mother for a young bird, and then learned how to maintain it using a low-level sensorimotor learning, without any prior knowledge about how to do so. Using facial detection and facial affect recognition would enhance this model since the feedback from the human emotional state could be used as reinforcer for the robot, to learn how and when to follow the human being.

Moreover, applying this kind of architecture to an expressive robot caters for not only developing a following behaviour, but also for allowing it to discover how to respond to the user's facial expression, according to the context (task being handled, number of human present).

From that point, using theories from developmental psychology (Sroufe, 1995), it would be possible for the robot to develop its own set of emotional states, all branching from one precursor of emotion. To clarify, it is believed that infants are born endowed with one proto-emotion, the excitement, which is modulated by fluctuation of endogenous states of the central nervous system and later by external stimulations. During the infant development in the first year of life, these fluctuations of excitement are then correlated with external stimuli, to later build categories of emotional states according to the context. For instance, anger is believed to derive from an increase of the excitement provoked by the inability for the infant to carry out an action because something or someone is preventing it from happening. Giving the possibility for the robot to build these context, and then, the categories of fluctuation of its inner excitement, given the context which has to take into account the other agent (human user or another robot) emotional state, would help the robot developing its own representation from the environment and coping strategies in the various situations it has encountered. This way, each robot would be adapted to the environment they are embedded in, since their development would be a product of their own experiences, and not a set of fixed rules already built-in their architecture.

5.2 Emotional robotic feedback via facial expression

In human-robot interaction, it is of paramount importance that the robot is capable of determining the emotional state of the human user, or on a lesser scale at least able to determine the emotional expression of the user in terms of simple emotional states such as 'stressed', 'relaxed', 'frustrated', or basic emotions such as 'happy', 'sad', 'angry' or 'surprised'. However, there is also a strong case for the reverse of this. That is, allowing the robot to exhibit emotional expressions that relate to some internal state of the robot. The purpose for this would be to allow the human interacting with the robot to assess how their interaction is being handled by the robot in a manner that s/he is familiar with. In order to bridge the gap between humans interacting with robots and in turn robots interacting with humans, the process needs to be as natural as possible without the need for a priori information.

Therefore, we have developed a robot head ERWIN (Emotional Robot with Intelligent Networks) that incorporates several interactive modalities that allows for HRI. ERWIN is designed to be capable of tracking and recognising human faces. The method used for initial detection of a face is based on Viola's (2001) rapid object detection with improvements made by Lienhart and Maydt (2002). In order to be able to detect a face, or any desired object of interest, specially trained classifiers are used. These classifiers are trained as described in (Kuranov *et al.*, 2002) using a sample set of images that contain the particular feature we wish to detect; in our case this was a collection of face images.

Once a set of images has been collated, a separate text file is created which provides the coordinates of a bounding box encapsulating the object within the specified image. These images and the accompanying text file are called the positive training set as they are used to build a classifier that will be used to detect the desired object. However, in order to reduce the chances of the classifier falsely identifying an object during runtime, the classifiers are also presented with what is termed negative training images. These consist of random images of various objects such as cars, trees, and buildings. An important feature of the negative training images is that they must *not* contain a positive image, i.e. the object we wish to detect – a face.

Sound source localisation is another modality that is built into ERWIN to allow for tracking of sound sources of interest within a cluttered environment. In addition to remaining focused on one user allowing for the two-way communication process to be as natural as possible. Thus, with the addition of emotional expressions being possible with robots it is possible for the human to judge how their interaction with the robot is being taken. For example, it would be possible to allow a robot to greet a user with a look of surprise, shock or happiness when it has not seen that user for a long time. Or if interaction with a human isn't going well, the robot is unable to clearly recognise the speech of the user, then the displayed emotion could reflect anger.

ERWIN also has the ability to display simple, but effective emotions by controlling several actuators attached to various parts of the face. ERWIN can move four separate actuators, each controlling a separate feature; these include the left and right eyebrows, and the upper and lower lips. By controlling these features based on responses and interaction from the external influences,s ERWIN can display a range of emotions that in turn can affect the response from a human thus bringing such HRI closer to human-human interactions. The basic emotions include happy, sad, angry, shocked and surprised. For instance, once ERWIN has been called or recognises a familiar face it can respond with generating a happy face. This gives excellent scope for combining the multiple modalities described, allowing the emotions to change if ERWIN detects sound but does not locate a face or changing emotion if ERWIN cannot understand or interpret what a human is saying. This may later provide an opportunity to develop the emotion response using internal states as modelled in artificial immune systems (Neal, 2002) and endocrine systems (Cañamero, 1997; Neal and Timmis, 2003; Avila-García and Cañamero, 2004; Cañamero and Avila-García, 2007), which allows internal states to influence responses to the external acoustic and visual information gathered.

6. Acknowledgment

This work is funded by the EU FP6 project Feelix Growing: FEEL, Interact, eXpress: a Global appRoach to develOpment With INterdisciplinary Grounding, Contract FP6 IST-045169 (http://feelix-growing.org).

7. References

Asteriadis, S., Tzouveli, P., Karpouzis, K., Kollias, S. (2008) Estimation of behavioral user state based on eye gaze and head pose - application in an e-learning environment, *Multimedia Tools and Applications*, accepted for publication.

Averill, J.R. (1975). A semantic atlas of emotional concepts. *JSAS Catalogue of Selected Documents in Psychology*: 5, 330.

Avila-García, O. and Cañamero, L. (2004). Using Hormonal Feedback to Modulate Action Selection in a Competitive Scenario. In *Proc. Eight Intl. Conf. Simulation of Adaptive Behavior (SAB04)*, pp. 243–252. MIT Press, Cambridge, MA.

Cañamero, L.D. (1997). Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In Johnson, W.L. (ed.) Proc. First Intl. Conf. on Autonomous Agents, pp. 148–155. ACM Press, New York.

Cañamero, L. (2005). Emotion Understanding from the Perspective of Autonomous Robots Research. *Neural Networks*, 18: 445-455.

Cañamero, L. and Avila-García, O. (2007). A Bottom-Up Investigation of Emotional Modulation in Competitive Scenarios. In A. Paiva, R. Prada, and R.W. Picard (Eds.), *Proc. Second International Conference on Affective Computing and Intelligent Interaction (ACII 2007),* LNCS 4738, pp. 398–409. Berlin & Heidelberg: Springer-Verlag.

Cañamero, L., Blanchard, A., Nadel, J. (2006), Attachment bonds for human-like robots *International Journal of Humanoıd Robotics*, 3(3), 301–320.

Cañamero, L. and Gaussier, P. (2005). Emotion Understanding: Robots as Tools and Models. In J. Nadel and D. Muir (Eds.), *Emotional Development: Recent research advances*, pp. 235-258. Oxford University Press.

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K. (2006) Modeling naturalistic affective states via facial and vocal expressions recognition, International Conference on Multimodal Interfaces (ICMI'06), Banff, Alberta, Canada, November 2-4, 2006.

Cootes, T., Edwards, G., Taylor, C. (2001) Active appearance models, IEEE Trans PAMI 23 (6), pp. 681-685.

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., & Fellenz, W. (1999). What a neural net needs to know about emotion words. In N. Mastorakis (Ed.), *Computational intelligence and applications,* pp. 109-114. World Scientific Engineering Society.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. (2000) 'Feeltrace': An instrument for recording perceived emotion in real time, in *Proc. ISCA Workshop on Speech and Emotion*, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. (2001) Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*: 18 (1), pp. 32–80.

Cowie, R., Douglas-Cowie, Karpouzis, K., Caridakis, G., Wallace, M., Kollias, S. (2008) Recognition of Emotional States in Natural Human-Computer Interaction, in D. Tzovaras (ed.), *Multimodal User Interfaces - From Signals to Interaction*, pp. 119-153, Springer Berlin Heidelberg.

Cristinacce, D., Cootes, T., Scott, I. (2004) A multi-stage approach to facial feature detection, *15th British Machine Vision Conference*, pp. 231-240.

D' Orazio, T., Leo, M., Cicirelli, G., Distante, A. (2004) An algorithm for real time eye detection in face images, *ICPR*, Vol. 3, 278-281.

Damasio, A. (2003) *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*, Harcourt, Orlando, FL, USA.

Ekman P., Friesen, W.V. (1978) *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, San Francisco, USA.

Ekman, P. (1993) Facial expression and Emotion. *Am. Psychologist*, Vol. 48, pp. 384-392.

Elman, J.L. (1990) Finding structure in time, *Cognitive Science*, 14, pp. 179-211.

Elman, J.L. (1991) Distributed representations, simple recurrent networks, and grammatical structure, *Machine Learning*, 7, 195-224, 1991.

Frijda, N.H. (1986). *The Emotions: Studies in Emotion and Social Interaction*, Cambridge University Press, New York, USA.

Goldie, P. (2004) *On Personality*. Rutledge, New York, USA.

Hammer, B., Tino, P. (2003) Recurrent neural networks with small weights implement definite memory machines, *Neural Computation* 15(8), pp. 1897-1929.

Haykin, S. (1999) *Neural Networks: A Comprehensive Foundation*, Prentice Hall International.

Hiolle, A., Cañamero, L., and Blanchard, A. (2007), Learning to interact with the caretaker: A developmental approach. In Paiva, A., Prada, R., and Picard, R., (Eds.), Proc. of the 2nd Intl. Conf. on Affective Computing and Intelligent Interactions, pages 422–433. Berlin and Heidelberg: Springer-Verlag

Humaine FP6 Network, of Excellence, http://emotion-research.net, Retrieved on Jan. 28, 2008.

Ioannou, S., Caridakis, G., Karpouzis, K., Kollias, S. (2007) Robust Feature Detection for Facial Expression Recognition, *EURASIP Journal on Image and Video Processing*, doi:10.1155/2007/29081.

Jesorsky, O., Kirchberg, K.J., Frischholz, R.W. (2001) Robust face detection using the Hausdorff distance, *3rd Conf. AVBPA*, pp. 90-95.

Kant, I. (1790) *Critique of Judgment*, Trans. Werner S. Pluhar, Hackett Publishing, Indianapolis, USA, 1987.

Kuranov, A., Lienhart R., and Pisarevsky V. (2002). An Empirical Analysis of Boosting Algorithms for Rapid Objects With an Extended Set of Haar-like Features. Intel Technical Report MRL-TR, July 2002.

Lazarus, R.S., (1991). Emotion and Adaptation. Oxford University Press, New York, NY.

Lazarus, R.S., Folkman. S. (1987). Transactional theory and research on emotions and coping - European Journal of Personality.

Leung, S.H., Wang S.L., Lau, W.H. (2004) Lip image segmentation using fuzzy clustering incorporating an elliptic shape function, *IEEE Trans. on Image Processing*, vol.13, No.1.

Lienhart R., and Maydt J. (2002). An Extended Set of Haar-like Features for Rapid Object Detection. In *Proc. IEEE Intl. Conf. Image Processing (ICIP)*, Vol. 1, pp. 900-903.

Lorenz, K. (1935), Companions as factors in the bird's environment. In *Studies in Animal and Human Behavior*, volume 1, pages 101–258. London: Methuen & Co., and Cambridge, Mass.: Harvard University Press.

Neal, M.J. (2002). An Artificial Immune System for Continuous Analysis of Time-Varying Data. In J. Timmis and P. J. Bentley, editors, *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS),* volume 1, pages 76 -- 85, University of Kent at Canterbury, September 2002. University of Kent at Canterbury Printing Unit.

Neal, M., Timmis, J.: Timidity: A Useful Emotional Mechanism for Robot Control? *Informatica* 27, 197–204 (2003)

Ortony, A., Collins A., Clore. G. L. (1988) *The Cognitive Structure of Emotions*, Cambridge University Press.

Pantic, M., Rothkrantz, L.J.M. (2000) Expert system for automatic analysis of facial expressions, *Image and Vision Computing,* Vol. 18, 2000, pp. 881-905.

Picard, R. W. (1997) *Affective Computing,* MIT Press, 0-262-16170-2, Cambridge, MA, USA.

Rusting, C. (1998) Personality, mood, and cognitive processing of Emotional information: three conceptual frameworks, *Psychological Bulletin*, 124, 165-196.

Schaefer, A. M., Zimmermann, H. G. (2006) Recurrent Neural Networks are Universal Approximators, *ICANN 2006*, pp. 632-640.

Scherer, K. R. (1987) Toward a dynamic theory of emotion: The component process model of affective states, *Geneva Studies in Emotion and Communication*, 1, pp. 1–98, Geneva, Switzerland.

Scherer, K. R. (2001) Appraisal considered as a process of multi-level sequential checking, In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research*, pp. 2–120, Oxford University Press, New York, USA.

Scherer, K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In A. S. R. Manstead, N. H. Frijda, & A. H. Fischer (Eds.), *Feelings and emotions The Amsterdam symposium*, pp. 136–157, Cambridge University Press.

Schlosberg, H. (1954) A scale for judgment of facial expressions, *Journal of Experimental Psychology*, 29, pp. 497-510.

Sebe, N., Lew, M.S., Cohen, I., Sun, Gevers, Y. T., Huang, T.S. (2004) Authentic Facial Expression Analysis, *International Conference on Automatic Face and Gesture Recognition (FG'04)*, Seoul, Korea, May 2004, pp. 517-522.

Smith, P., Shah, M., da Vitoria Lobo, N. (2003) Determining Driver Visual Attention with One Camera, *IEEE Trans. Intelligent Transportation Systems*, Vol 4, No. 4, pp. 205-218.

Sroufe, L. A. (1995). *Emotional Development: The Organization of Emotional Life in the Early Years*. Cambridge University Press.

Tekalp, M., Ostermann, J. (2000) Face and 2-D mesh animation in MPEG-4, *Signal Processing: Image Communication* 15, pp. 387-421, Elsevier.

Tian, Y.L., Kanade, T., Cohn, J.F. (2001) Recognizing Action Units for Facial Expression Analysis, *IEEE Transactions on PAMI*, Vol.23, No.2.

Tomasi, C., Kanade, T. (1991) Detection and Tracking of Point Features, *Carnegie Mellon University Technical Report* CMU-CS-91-132.

Viola P., Jones, M. (2001) Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* Vol. 1, pp. 511-518.

Whissel, C.M. (1989) The dictionary of affect in language, in Plutchnik, R. and Kellerman, H. (Eds.): *Emotion: Theory, Research and Experience: The Measurement of Emotions*, , Vol. 4, pp.113–131, Academic Press, New York.

Wundt, W. (1903) *Grundzuge der Physiologischen Psychologie*, vol. 2. Engelmann, Leipzig.