Body gesture and facial expression analysis for automatic affect recognition

Ginevra, Castellano[1] and George, Caridakis[2] and Antonio, Camurri[3] and Kostas, Karpouzis[2] and Gualtiero, Volpe[3] and Stefanos, Kollias[2]

1. Department of Computer Science, School of Electronic Engineering and Computer Science, Queen Mary University of London. Mile End Road, London E1 4NS, United Kingdom.
2. Image, Video and Multimedia Systems Lab, National Technical University of Athens. 9 Heroon Polytechniou Str., GR-157 80 Zographou, Greece.
3. InfoMus Lab, Department of Communication, Computer and System Sciences, University of Genoa. Viale Causa 13, 16145, Genoa, Italy.

Ginevra Castellano, ginevra@dcs.qmul.ac.uk

George Caridakis, gcari@image.ece.ntua.gr

Antonio Camurri, antonio.camurri@unige.it

Kostas Karpouzis, kkarpou@cs.ntua.gr

Gualtiero Volpe, gualtiero.volpe@unige.it

Stefanos Kollias, stefanos@cs.ntua.gr

Summary

Affect recognition plays an important role in everyday life. This explains why researchers in the human-computer and human-robot interaction community have increasingly been addressing the issue of endowing machines with affect sensitivity. Affect sensitivity refers to the ability of analysing verbal and non-verbal behavioural cues displayed by the user to infer the underlying communicated affect.

This Chapter describes affect sensitivity as an important requirement for an affectively competent agent. An affectively competent agent should be able to exploit affect sensitivity abilities to successfully interact with human users: the perception and interpretation of affective states and expressions is important for such an agent to act more socially and engage with users in truly natural interaction.

In this Chapter, affective facial and bodily expressions are addressed as channels for the communication of affect that must be taken into account in the design of an affect recognition system. A survey of state of the art computational approaches for affect recognition, based on the automatic analysis of facial and bodily expressions and their combined information, is presented. Relevant contributions in the field are reviewed and affect sensitivity is discussed with respect to key issues that arise in the design of an automatic system for affect recognition. This Chapter draws particular attention to a number of challenges in face and body affect recognition research that have to be addressed in a more

comprehensive manner in order to successfully design an affectively competent agent. First of all, an affectively competent agent should be capable of perceiving naturalistic expressions conveying states different from prototypical emotions. Second, such an agent should be able to analyse multiple modalities of expressions: new fusion methods that take into consideration the relationships and synchronisation of different modalities are required. Another important issue is the dynamic account of affect and affect-related expressions: automatic analysis of temporal dynamics is required for an affective competent agent to be able to monitor the evolution of the affective states displayed by the user over time. Robustness in the real world environment is an issue not to be neglected as, to date, many automatic systems for the analysis of affective facial and bodily expressions only work well in controlled environments. Finally, context sensitivity is currently an underexplored topic that has to be taken into consideration in the design of an affectively competent agent, as the detection of the most subtle affective states can be achieved only through a comprehensive analysis of their causes and effects.

The main contribution of this Chapter is a review of studies on affect recognition from face and body and the role played by a subset of these in addressing the challenges described above. The above issues should be more extensively investigated by the affect recognition community in order to successfully carry out the design of an affectively competent agent.

Introduction

The ability to communicate with humans intelligently and in a sensitive manner is an important requirement for an affectively competent agent. Research in the area of human-computer and human-robot interaction has increasingly been addressing the problem of modelling intelligent agents by exploiting communication channels that are typical of humans. Affect plays a key role in human-human communication: recognition and expression of affective states are vital for people to understand and be understood, to ensure that the communication has succeeded. This explains why there is an increasing interest in the development of paradigms of interaction endowing artificial agents with affect sensitivity (Picard 1997; Zeng 2009).

Affect sensitivity refers to the ability to analyse verbal and non-verbal social affective cues displayed by humans in order to infer their affective states. These range from prototypical emotions (e.g., basic emotions such as joy, anger, sadness, etc.) to more complex affective and mental states such as interest, boredom, frustration, engagement, etc. An affectively competent agent should be able to exploit affect sensitivity abilities to successfully establish an affective interaction with human users: the perception and interpretation of affective states and expressions is important in order for such an agent to be able to provide an appropriate response to the user's behaviour, i.e., to act more socially and engage in a truly natural interaction with users.

Given the key role of affect and emotion in our daily lives, technology for the design of affectively competent agents is attracting increasing interest. It is expected that many applications will benefit from the efforts made by the affect

recognition community to build affect sensitive machines. Applications cover a large range of domains, from security to therapy of people with disabilities, from entertainment to assistive technology.

Several examples of affect sensitive systems are reported in the human-computer and human-robot interaction literature. The affective wearable system by El Kaliouby and colleagues, for example, can help people affected by autism improve their social communication abilities and learn to express emotions (El Kaliouby et al. 2006; Madsen et al. 2008). Kapoor et al. (Kapoor et al. 2007) designed a system for the detection of frustration in students using a learning companion based on automatic analysis of multimodal non-verbal cues including facial expressions, head movement, posture, skin conductance and mouse pressure data. In the robotics domain, one of the most famous examples of an affect sensitive robot is Kismet, which is endowed with an attention system based on low-level processing of perceptual stimuli (Breazeal et al. 2001; 2003). There are several other examples of affectively competent agents. All of them require several types of abilities, but the understanding of the nature of human affect is a basic requirement for scientists to be able to design affect sensitive machines.

There is no doubt that the design of an affect recognition system should be addressed from a multidisciplinary perspective. The development of new technologies must be accompanied by critical studies of psychological issues that arise from these developments. In this respect, the inclusion of affect representation into a framework for affect recognition is of primary importance.

Incorporating models and paradigms developed by psychologists for the classification of affective states (Scherer 2000; chapter by Scherer a) is a pressing need and still a challenging issue.

According to the *Component Process Model of Emotion (CPM)* proposed by Scherer (Scherer 1984; Chapter by Scherer b), a necessary condition for an emotional episode to occur is the synchronisation of different processes, as a consequence of a situation/event appraised as highly relevant for an individual's well-being. These processes include appraisal, physiological arousal, action tendency, subjective feeling and motor expression.

People communicate emotions consciously or unconsciously and can use some channels of expression more than others. Some expressions can be directly perceived, such as facial expressions and voice intonation, while others, such as changes in physiological parameters, can be detected only by using specific sensors. Another channel of affective information is the body: people can express emotions through gestures, postures, and head and full-body movements. The human sensory system performs multimodal integration of all this information, allowing people to recognise emotions by exploiting different channels (Meeren et al. 2005; Stein and Meredith 1993). Similarly, an affectively competent agent should be endowed with the ability to analyse different types of affective expressions and fuse them together to infer emotions. At the same time, such an agent must be able to evaluate how different affective expressions influence each other, as well as the amount of information each of them provides about emotion.

This Chapter focuses on facial and bodily expressions and their combined information as channels of affect expression. The ability for a system to recognise affective states requires successful association with their patterns of expression (Picard, 1997). Nevertheless, a lot of work still has to be conducted in order to establish which affective expressions are the best indicators of an affective state. Several researchers, mainly from psychology, have investigated the relationship between emotion, facial expressions, and body gestures (see, for example, Boone and Cunningham 1998; De Meijer 1989; Ekman, 1994; Ekman and Friesen 1975; Wallbott 1998). Face and body appear to be the most important channels for communicating behavioural information, including affective states (Ambady and Rosenthal 1992).

While research on affect recognition based on face and body gesture has been extensively addressed in the literature (Zeng et al. 2009), several challenges still have to be properly investigated.

Spontaneous and non-prototypical affective expressions and states. Research on automatic affect recognition needs to move toward systems sensitive to spontaneous, non-prototypical affective expressions and states. The design of many existing affect recognition systems was based on databases of acted affective expressions displaying basic emotions (Zeng et al. 2009). Although acted affective expressions present several advantages that recordings in a controlled environment can provide (e.g., precise definition, many expressions recorded from the same individual, very high quality of recordings,

etc.), they are often exaggerated and decontextualised (Bänziger and Scherer 2007) and mainly reflect stereotypes rather than genuine affective states.

Multimodal affective expressions. Humans can rely on different channels of information to understand the affective messages communicated by others. Similarly, it is expected that an automatic affect recognition system should be able to analyse different types of affective expressions. In this respect, an important issue to be addressed is the fusion of different modalities of expression, which must be designed by taking into account the relationship and correlation across different modalities.

Dynamic account of affective expressions. The dynamics of affective expressions is an important factor in the understanding of human behaviour. Affect and its expressions vary over time: analysis of static affective expressions cannot account for the dynamic changes in the behavioural responses characterising an affective state and in the affective state itself.

Robustness to real world scenarios. Real world applications require affect recognition systems built upon face and body detectors and facial and body features tracking systems able to efficiently function in uncontrolled environments: this requires, for example, robustness to noisy backgrounds, occlusions, rigid head motions, etc.

Context sensitivity. As suggested by the Component Process Model of Emotion (Scherer 1984; Chapter by Scherer b), appraisal is as important as

behavioural responses to characterise affect. This highlights the need for an automatic affect recognition system to take into account the conditions that elicited an emotional response. A context sensitive affect recognition system must be sensitive to several types of contextual information, such as individual differences in expressing affect, personality of the person expressing affect, preferences, goals, underlying mood, task, environment, etc.

This Chapter presents an overview of computational models and techniques reported in the literature for the analysis of facial and bodily expressions and the automatic recognition of affective states. Relevant contributions in the field are discussed and examples of studies addressing the challenges highlighted above are reported.

This Chapter is organised as follows. The next two sections present an overview of affect recognition methods based on the analysis of facial and bodily expressions. Subsequently, a survey of studies and techniques for achieving a fusion of face and body information is provided. Different fusion strategies are discussed. The Chapter ends with an overview of the limitations and challenges in the field of automatic affect recognition from face and body.

Affect recognition from face

Most of the work on affect recognition reported in the literature focuses on the automatic analysis of facial expressions (Zeng et al. 2009).

Two main streams of research in affect recognition from the face can be identified: facial affect recognition and facial muscle action recognition (for an extensive survey see Pantic and Bartlett 2007; Zeng et al. 2009). While the first aims to map facial expressions directly onto affective categories, the second attempts to recognise facial signals. These two streams are associated with two major approaches for behaviour measurement adopted in psychology: message judgement, which tries to infer what is behind a displayed facial behaviour, and sign judgement, which merely provides an objective description of facial signals (Cohn 2006).

As far as the latter is concerned, the Facial Action Coding System (FACS) introduced by Ekman and colleagues (Ekman and Friesen 1978) allows for an objective and comprehensive description of facial expressions and is the most frequently used method to describe facial behaviour. FACS describes expressions in terms of Action Units (AUs), which relate to the contractions of specific facial muscles. Once AUs are detected, they can be mapped onto affective categories using high-level rules such as FACSAID (FACS Affect Interpretation Database) (Ekman et al. 2002). An alternative method to model facial expressions is that based on MPEG-4 metrics (Tekalp and Ostermann 2000). MPEG-4, which mainly focuses on facial expression synthesis and animation, defines the Facial Animation Parameters (FAPs), which are strongly related to the AUs.

Most of the systems for automatic analysis of facial expressions are based on 2D spatio-temporal features. These include *geometric features*, such as the positions of salient facial points (e.g., mouth corners, etc.) or the shape of face

components (e.g., eyes, mouth, etc.), and *appearance features*, which represent the texture of the facial skin. Systems based on methods using geometric features include, for example, those of Pantic and colleagues (e.g., see Pantic and Patras 2006; Valstar et al. 2007), Gunes and Piccardi (Gunes and Piccardi 2009), Chang and colleagues (Chang et al. 2006). Examples of methods using appearance features are those of Bartlett and colleagues (e.g., see Bartlett et al. 2003) and Anderson and McOwan (Anderson and McOwan 2006).

The majority of the affect recognition systems based on automatic analysis of facial expressions have focused on the recognition of basic emotions (Zeng et al. 2009). A few attempts to detect non basic affective states have been reported in the literature. El Kaliouby and Robinson (El Kaliouby and Robinson 2004), for example, developed a computational model that detects, in real-time, complex mental states such as agreeing, concentrating, disagreeing, being interested, being unsure and thinking from facial expressions and head movement in video. Littlewort et al. (Littlewort et al. 2007) used an automatic system for facial expression recognition to detect expressions of pain. Yeasin et al. (Yeasin et al. 2006) developed a system that recognises six universal facial expressions and uses them to compute levels of interest.

Furthermore, most of the approaches are based on databases of acted affective expressions. Naturalistic data goes beyond extreme emotions and concentrates on more natural affective episodes that happen more frequently in everyday life. To date a few efforts to develop systems for the automatic detection of spontaneous affective expressions have been reported. Examples

include the neurofuzzy system for emotion recognition by Ioannou et al. (Ioannou et al. 2005), which allows for the learning and adaptation to specific users' naturalistic facial expressions, and the works by Valstar et al. (Valstar et al. 2007) and Littlewort et al. (Littlewort et al. 2007), who reported results on the automatic discrimination between posed and spontaneous facial expressions.

Another key challenge in facial affect recognition is the dynamics of affective expressions. Littlewort et al. (Littlewort et al. 2006) suggest that natural and posed expressions are inherently different in terms of temporal dynamics, providing arguments from psychological research (Ekman and Friesen 1982; Frank and Ekman 1993). Other studies show that the dynamics of expressions is a key factor in the discrimination between posed and spontaneous facial behaviour (Cohn and Schmidt 2004; Littlewort et al. 2007; Valstar et al. 2007). Valstar et al. (Valstar et al. 2007), for example, showed the important role of the temporal dynamics of face, head and shoulder expressions in discriminating posed from spontaneous smiles. Other efforts towards a dynamic account of affective expressions include the work by Pantic and Patras (Pantic and Patras 2006), which deals with facial actions dynamics recognition, and the work by Cowie et al. (Cowie et al. 2008), who proposed a dynamic approach for affect recognition based on a recurrent neural network whose short-term memory and approximation capabilities allow for a dynamic modelling of events and classification of input patterns into affective states.

Context sensitivity in facial affect recognition is still an underexplored challenge. Systems for facial affect recognition need to take into account and

adapt their knowledge to the specific user or context of interaction. Adaptation in terms of environment variables or personalised expressivity has to be considered. Neural networks are well suited to fulfilling the adaptation requirement. Caridakis and colleagues (Caridakis et al. 2008), for example, proposed an approach that uses neural network architectures to detect the need for adaptation of their knowledge.

Another example in which context is taken into account in facial affect recognition is the work by Pantic and Rothkrantz (Pantic and Rothkrantz 2004), who proposed a case-based reasoning system capable of classifying facial expressions into the emotion categories learned from the user.

Finally, robustness in real world scenarios is an important factor in facial affect recognition. Most existing approaches in facial feature extraction are either designed to cope with a limited diversity of video characteristics or require manual initialisation or intervention. Examples of exceptions are the system by Pantic and Patras (Pantic and Patras 2006), which is robust to illumination changes, and the system by Anderson and McOwan (Anderson and McOwan 2006), characterised by its robustness to rigid head motions.

Affect recognition from body

Although most of the studies on affect recognition reported in the literature have focused on the automatic analysis of facial expressions, some attempts have also been made toward the design of systems capable of analysing

expressive body movement to infer human affect. Different streams of research in affect recognition from body movement can be identified depending on the type of information analysed to recognise affect. Some systems, for example, base the prediction of the conveyed affective content on the type of gesture performed, others on the way a gesture is performed, others analyse body postures.

As far as systems based on analysis of the type of gesture performed are concerned, examples include: the system by Balomenos and colleagues (Balomenos et al. 2005), which uses the position of the centroid of the head and hands to recognise gesture classes and map them onto emotions; the work by Gunes and Piccardi (Gunes and Piccardi 2009), who presented a method for the recognition of acted affective states based on analysis of affective body displays and automatic detection of their temporal segments; and the work by Shan et al. (Shan et al. 2007), who used spatio-temporal features for modelling affective body gestures.

Other studies addressed automatic affect recognition from the perspective of movement expressivity, attempting to predict affect based on the way gestures are performed (Bernhardt and Robinson 2007; Camurri et al. 2003, 2004; Castellano et al. 2007, 2008). Camurri and colleagues (Camurri et al. 2005) developed a multilayered conceptual model for multimodal analysis of affective, emotional content in human full-body movement and gesture based on movement qualities. The model (see Figure 1) is based on four different layers, following a bottom-up approach (Camurri et al., 2004). Layer 1 includes

techniques for pre-processing of data from different kinds of sensors such as video cameras, on-body (e.g., accelerometers), and environmental sensors. Layer 2 extracts from sensor data a collection of expressive motion features describing the movement being performed (Volpe 2003). Features are derived from research in psychology (e.g., Wallbott 1998; Boone and Cunningham 1998) and human sciences. For example, an important set of features are those relating to the *Theory of Effort* by choreographer Rudolf Laban (Laban 1947; 1963). EyesWeb XMI, a platform for synchronised analysis of multimodal data streams developed by Camurri and colleagues (Camurri et al. 2007; www.eyesweb.org), allows for the extraction of a wide collection of motion features from video and sensor data streams. Layer 3 deals with two major issues: segmentation of movement in its composing gestures, and representation of such gestures in suitable spaces. Layer 4 is conceived as a conceptual network mapping the extracted features and gestures into conceptual structures. The model was tested in a study aiming to automatically discriminate between four emotions (anger, fear, grief, and joy) in dance performances (Camurri et al., 2003; 2004).

Insert Figure 1 about here.

Work on affect recognition from movement expressivity includes the study by Bernhardt and Robinson (Bernhardt and Robinson 2007), who proposed a framework for affect recognition in knocking motions using motion-captured data. They computed motion cues (e.g., velocity and acceleration of the

hand, etc.) over motion primitives obtained using a method based on segmentation by motion energy and clustering of motion segments. Support Vector Machines were then trained for each motion primitive using statistical measures of the extracted motion cues.

The literature provides some examples of work on affect recognition from postures. Bianchi-Berthouze and Kleinsmith (Bianchi-Berthouze and Kleinsmith 2003) proposed an approach to self-organise postural features into affective categories to provide robots with the ability to incrementally learn to recognise affective human postures through interaction with human partners. Mota and Picard (Mota and Picard 2003) explored how sequences of postures can be used to predict affective states related to a child's interest level during a learning task performed with the computer. Based on posture data collected through pressure sensors mounted on a chair, Hidden Markov Models were used to predict the affective state related to sequences of postural behaviour. D'Mello and Graesser (D'Mello and Graesser 2009) investigated the effectiveness of detecting affective states (boredom, confusion, delight, flow, and frustration) of learners interacting with an intelligent tutoring system using automatically extracted body pressure data carrying information about body position and arousal.

Despite the fact that studies on affect recognition from body movement are less numerous than those based on facial expressions, the literature includes a few studies that address some of the challenges in affect recognition research. Castellano et al. (Castellano et al. 2008), for example, proposed an approach based on movement expressivity analysis to classify non-prototypical

expressions in music performance into different emotionally expressive intentions (i.e., sad, allegro, serene, personal and overexpressive). Varni et al. (Varni 2008; Varni et al. 2008) reported initial results in the attempt to measure emotional synchronisation and identify leadership relationships in musicians using analysis of head motion expressivity.

As far as attempts to take into consideration the dynamics of affective expressions are concerned, Castellano (Castellano 2008) proposed an approach for affect recognition based on the dynamics of movement expressivity, arguing that human affect can be recognised not only from the type of gestures performed, but also from the way they are performed. The approach was inspired by recent theories from psychology (Scherer 2001) claiming that emotional expression is reflected to a greater extent in the timing of the expression than in absolute or average measures. By focusing on the dynamics of expressive motion cues, the proposed approach addresses how motion qualities vary at different temporal levels. A mathematical model that allows for the extraction of information about the dynamics of movement expressivity was proposed. The idea behind this model is to use information about temporal series of expressive motion cues in a format that is suitable for feature vector-based classifiers. The model provides features conveying information about the dynamics of movement expressivity, i.e., information about fluctuations and changes in the temporal profiles of cues (Castellano 2008; Castellano et al. 2008).

Based on this model, Castellano et al. (Castellano et al. 2008) analysed upper-body and head gestures in musicians expressing emotions and reported

that features related to the timing of expressive motion cues such as the quantity of motion and the velocity of the head are more effective for the discrimination of affective states than traditional statistical features, such as the mean or the maximum. Figure 2 shows a measure of the quantity of motion of a pianist's upper body and the tracking of the head.

Finally, a noteworthy effort is that of Gunes and Piccardi, who explored how the modelling and detection of temporal phases of body displays can improve the accuracy of affect recognition.

Insert Figure 2 about here.

Affect recognition through multiple modalities: face and body

An affectively competent agent should be endowed with the ability to analyse different types of affective expressions: fusing different affective cues can allow for a better understanding to be achieved of the affective message communicated by the user. Hence the need for a multimodal affect recognition system.

While unimodal systems have received a thorough investigation, studies taking into account the multimodal nature of the affective communication process are still not numerous (Zeng et al. 2009). Some efforts combining the face and body modalities of expressions for the purpose of affect recognition have been reported in

the literature. El Kaliouby and Robinson (El Kaliouby and Robinson 2004), for example, proposed a vision-based computational model to infer mental states from head movements and facial expressions. Their approach uses Hidden Markov Models for real-time head and facial actions recognition and Dynamic Bayesian Networks to model mental states over time. Other examples include: the work by Balomenos et al. (Balomenos et al. 2005), who used gestures, recognised with Hidden Markov Models, to support the output of the facial expression analysis in a bimodal emotion recognition system; the system by Gunes and Piccardi (Gunes and Piccardi 2009), that allows for the recognition of affective states via synchronisation of temporal phases of face and body displays; the work by Shan et al. (Shan et al. 2007), who combined face and body features for bimodal affect recognition using Canonical Correlation Analysis to establish the relation between the two modalities; and the work by Valstar et al. (Valstar et al. 2007), who combined multimodal information conveyed by facial expressions, head and shoulders movement, to discriminate between posed and spontaneous smiles.

Other studies have considered the use of multiple channels of information in addition to face and gesture. Kapoor et al. (Kapoor et al. 2007), for example, designed a system capable of detecting frustration using multimodal non-verbal cues such as facial expressions, blink, head movement, posture, skin conductance and mouse pressure. Castellano et al. (Castellano et al. 2008) proposed an approach in which facial expressions, body gesture and speech data is fused at the feature and decision level to predict eight affective states in a speech-based interaction scenario.

A few of the proposed multimodal systems integrate contextual information for the purpose of affect recognition. The system by Kapoor and Picard (Kapoor and Picard 2005) allows for the detection of interest in a learning environment by combining non-verbal cues and information about the learner's task (e.g., the level of difficulty and the state of the game). Castellano et al. (Castellano et al. 2009) proposed an approach to predict the level of engagement of children playing chess with an iCat robot in a naturalistic scenario. Their approach is based on the fusion of non-verbal behaviour (i.e., eye gaze and smiles) and contextual features such as the state of the game and the display of facial expressions by the robot.

An important issue in multimodal affect recognition is the fusion of different modalities. Features from different modalities of expressions can be fused at different levels (Wu et al. 1999). *Feature-level fusion* can be performed by merging features from different modalities and inputting them into a single classifier. In this framework, correlation between modalities can be taken into account during the learning phase. In general, feature-level fusion is more appropriate in case of closely coupled and synchronised modalities such as speech and lip movements, but tends not to generalise very well if the temporal characteristics of features from different modalities differ substantially, as is the case for speech and facial expression or gesture. Moreover, due to the high dimensionality of input features, large amounts of data are required for training purposes.

Decision-level fusion is preferred when the integration of asynchronous, but temporally correlated modalities, is needed. With this approach each modality is classified independently and the outputs of different modalities are combined to

obtain the final classification. Although several approaches have been proposed (Kittler et al. 1998), designing suitable strategies for decision-level fusion is still a challenge. The knowledge of the characteristics of each modality could inform the choice of different classification schemes for different modalities and the dynamic interconnection among them.

A major issue for both decision-level and feature-level fusion is the synchronisation of the different modalities. As far as the feature extraction process is concerned, raw data can be collected at a different sampling rate and different modalities can be processed at different time scales. On the other hand, in the case of decision-level fusion, different classifiers may provide classification results at a different frequency. Multimodal synchronisation remains a challenging issue that requires further investigation.

Results from studies in neurology (Stein and Meredith 1993) suggest that the integration of different perceptual signals occur at an early stage of human processing of stimuli. This seems to suggest that different modalities should be processed in a joint feature space rather than combined with the results of a late fusion. Features from different modalities are often complementary and redundant, and their relationship is often unknown. The development of novel methods for multimodal fusion should take into consideration the underlying relationships and correlations between the feature sets in different modalities (Shan et al. 2007; Zeng et al. 2006), how different affective expressions influence each other and how much information each of them provides about the communicated affect.

Conclusion

This Chapter investigated affect sensitivity as an important requirement for an affectively competent agent. Specifically, face and body gestures were addressed as channels for the communication of affect. A survey of the state of the art in computational approaches for affect recognition based on the automatic analysis of facial and bodily expressions and their combined information was presented. Affect sensitivity was discussed with respect to key issues that arise in the design of an automatic system for affect recognition.

Although affect recognition from face and body movement has been increasingly investigated as of late (Zeng et al. 2009), several challenges still have to be addressed in a more comprehensive manner by the affect recognition community.

First of all, despite a few attempts (El Kaliouby and Robinson 2004; Kapoor et al. 2007; Littlewort et al. 2007), many proposed approaches have focused on the detection of prototypical emotions, while it is expected that an affectively competent agent is sensitive to affective signals conveying more subtle states. Moreover, the majority of the existing systems have been trained using acted and exaggerated expressions rather than with spontaneous samples of human behaviour. The design of real world applications will require affectively competent agents that are sensitive to spontaneous, real-life affective expressions and states. This will necessitate advances in the definition of new methodologies for labelling naturalistic affective data, currently a non-trivial issue in affect recognition research.

A second important issue is the need for new systems and methods that analyse multiple modalities for the purpose of affect recognition. New challenges to be addressed by the affect recognition community include the development of new methods for fusing different modalities that take into consideration their intrinsic relationships and synchronisation.

Third, the temporal dynamic of affective expressions has been shown to play an important role in the interpretation of affective displays (Cohn and Schmidt 2004; Castellano et al. 2008; Gunes and Piccardi 2009; Valstar et al. 2007). The segmentation and analysis of the temporal phases of facial and bodily expressions are still challenging issues, as well as their relationships with the dynamics of the underlying affect.

Fourth, the design of an affectively competent agent requires further advances in the development of methods robust in uncontrolled environments. Expression detectors must be robust to noisy backgrounds, changes in the illumination conditions, rigid head motions, occlusions, etc.

Fifth, despite a few efforts (e.g., see Kapoor and Picard 2007; Castellano et al. 2009), context sensitivity is still an underexplored topic. A recommendation for the design of an affectively competent agent is that of taking into consideration information about the context in which an affective state or expression is displayed, as the interpretation of behavioural signals and the underlying affective states is context dependent.

The current state of the art in affect recognition represents a valuable resource for the design of an affectively competent agent. One of the drawbacks of the richness of studies in the field is that results are often not comparable due to the different experimental conditions, the different databases of affective expressions used to train the affect detectors, the different data used, and so on. Research on an affect sensitive agent should aim to establish common guidelines for the design of affect recognition frameworks suitable for real world applications.

Finally, we believe that the inclusion of affect representation into a framework for affect recognition is an important issue for the design of an affectively competent agent. Strengthening the connection with psychological models, although still challenging, would allow for the first steps towards the detection of more complex affective states and their components (e.g., appraisals, blends of emotions, preferences, mood, etc.) to be undertaken.

References

Ambady, N. and Rosenthal, R. (1992). Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin,* 111(2), 256-274.

Anderson, K. and McOwan P.W. (2006). A real-time automated system for recognition of human facial expressions. *IEEE Transactions on Systems, Man and Cybernetics – Part B*, 36(1), 96-105.

Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A. I., Karpouzis, K. and Kollias, S. (2005). Emotion analysis in man-machine interaction systems, in S. Bengio and H. Bourlard (ed.), *Machine Learning for Multimodal Interaction*, pp. 318–28, LNCS, vol. 3361, Berlin: Springer Verlag.

Bänziger, T, and Scherer, K. (2007). Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon.

Bartlett, M.S., Littlewort, G., Braathen, P., Sejnowski, T.J., and Movellan, J.R. (2003). A Prototype for Automatic Recognition of Spontaneous Facial Actions. *Advances in Neural Information Processing Systems*, 15, pp. 1271-1278.

Bernhardt, D. and Robinson, P. (2007). Detecting affect from non-stylised body motions, in A. Paiva, R. Prada, and R. W. Picard (ed.), *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, LNCS, 4738, pp. 59–70, Berlin: Springer-Verlag.

Bianchi-Berthouze, N. and Kleinsmith, A. (2003). A categorical approach to affective gesture recognition. *Connection Science*, 15(4), 259–69.

Boone, R. T. and Cunningham, J. G. (1998). Children's Decoding of Emotion in Expressive Body Movement: The Development of Cue Attunement. *Developmental psychology*, 34(5), 1007–16.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2), 119-55.

Breazeal, C. , Edsinger, A., Fitzpatrick, P. and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on Systems, Man and Cybernetics-Part A*, 31(5), 443–453.

Camurri, A., Coletta, P., Varni, G. and Ghisio, S. (2007). Developing multimodal interactive systems with EyesWeb XMI. *Proceedings of the 2007 Conference on New Interfaces for Musical Expression*, pp. 305–08.

Camurri, A., De Poli, G., Leman, M. and Volpe, G. (2005). Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications. *IEEE Multimedia*, 12(1), 43-53.

Camurri, A., Lagerlöf, I. and Volpe, G. (2003). Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques. *International Journal of Human-Computer Studies*, 59(1-2), 213-25.

Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R. and Volpe, G. (2004). Multimodal analysis of expressive gesture in music and dance performances, in A. Camurri, G. Volpe (ed.), *Gesture-based Communication in Human-Computer Interaction,* pp. 20-39, (Heidelberg: Springer Verlag).

Caridakis, G., Karpouzis, K. and Kollias, S. (2008), User and Context Adaptive Neural Networks for Emotion Recognition, *Neurocomputing*, Elsevier, 71(13-15), pp. 2553-62.

Castellano, G. (2008). *Movement Expressivity Analysis in Affective Computers: From Recognition to Expression of Emotion.* Ph.D. Dissertation, Faculty of Engineering, University of Genova, April 2008.

Castellano, G., Kessous, L. and Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech, in C. Peter and R. Beale (ed.): *Affect and Emotion in Human-Computer Interaction*. LNCS, 4868. Springer, Heidelberg.

Castellano, G., Mortillaro, M., Camurri, A., Volpe, G. and Scherer, K. (2008). Automated Analysis of Body Movement in Emotionally Expressive Piano Performances. Music Perception, 26(2), 103-19, University of California Press.

Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P.W. (2009). Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features. ACM International Conference on Multimodal Interfaces (ICMI'09), Cambridge, Massachusetts, USA.

Castellano, G., Villalba, S. D. and Camurri. A. (2007). Recognising Human Emotions from Body Movement and Gesture Dynamics, in A. Paiva, R. Prada, and R. W. Picard (ed.), *Affective Computing and Intelligent Interaction, Second*

International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings, LNCS, 4738, pp. 71–82, Berlin: Springer-Verlag.

Chang, Y., Hu, C., Feris, R., and Turk, M. (2006). Manifold Based Analysis of Facial Expression. *Journal of Image and Vision Computing*, 24(6), pp. 605-614.

Cohn, J.F. (2006). Foundations of Human Computing: Facial Expression and Emotion. ACM International Conference on Multimodal Interfaces (ICMI '06), pp. 233-238.

Cohn, J.F. and Schmidt, K.L. (2004). The Timing of Facial Motion in Posed and Spontaneous Smiles. International Journal of Wavelets, Multiresolution and Information Processing, 2, pp. 1-12.

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

De Meijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4), 247–68.

D'Mello, S. K. and Graesser, A. C. (2009). Automatic Detection of Learners' Emotions from Gross Body Language. *Applied Artificial Intelligence,* 23(2), 123-50.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115, 268-87.

Ekman, P. and Friesen, W.V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues*. Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P. and Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, Calif.: Consulting Psychologists Press.

Ekman, P. and Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior* 6(4), 238-52.

Ekman, P., Friesen, W. V. and Hager, J. C. (2002). *Facial Action Coding System*. A Human Face. Salt Lake City, USA.

El Kaliouby, R. and Robinson, P. (2004). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. Workshop on Real Time Computer Vision for Human Computer Interaction, IEEE International Conference on Computer Vision and Pattern Recognition.

El Kaliouby, R., Teeters, A. and Picard, R.W. (2006). An Exploratory Social-Emotional Prosthetic for Autism Spectrum Disorders. International Workshop on Wearable and Implantable Body Sensor Networks, pp. 3, April 3-5, 2006, MIT Media Lab, Cambridge, MA.

Frank, M. and Ekman, P. (1993). Not all smiles are created equal: The differences between enjoyment and non enjoyment smiles. *Humor: International Journal of Humor Research* 6(1), 9-26.

Gunes, H. and Piccardi, M. (2009). Automatic Temporal Segment Detection and Affect Recognition from Face and Body Display. *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, 39(1), pp. 64-84.

Kapoor, A., Burleson, W., and Picard, R. W (2007). Automatic prediction of frustration. International Journal of Human-Computer Studies, 65(8), pp.724-736.

Kapoor, A. and Picard R. W. (2005). Multimodal affect recognition in learning environments. ACM International Conference on Multimedia, pp. 677-682.

Kittler, J., Hatef, M., Duin, R. P. W. and Matas, J. (1998). On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), 226-39.

Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K. and Kollias, S. (2005). Emotion recognition through facial expression analysis based on a neurofuzzy method. *Neural Networks*, 18, 423-35.

Laban, R. and Lawrence, F.C. (1947). *Effort*. London: Macdonald & Evans Ltd.

Laban, R. (1963). *Modern Educational Dance*. London: Macdonald & Evans Ltd.

Littlewort, G., Bartlett, M., Fasel, I., Susskind, J. and Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing* 24(6), 615-25.

Littlewort, G.C., Bartlett, M.S. and Lee, K. (2007). Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. International Conference of Multimodal Interfaces, pp.15-21.

Madsen, M., el Kaliouby, R., Goodwin, M. and Picard, R.W. (2008). Technology for Just-In-Time In-Situ Learning of Facial Affect for Persons Diagnosed with an Autism Spectrum Disorder. Proceedings of the 10th ACM Conference on Computers and Accessibility (ASSETS), October 13-15, 2008, Halifax, Canada.

Meeren, H., Heijnsbergen, C. and Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. Proceedings of the National Academy of Sciences of USA, 102(45), pp. 16518-23.

Mota S. and Picard R.W. (2003). Automated posture analysis for detecting learner's interest level. Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI, June 2003.

Pantic, M. and Bartlett, M.S. (2007). Machine Analysis of Facial Expressions, in K. Delac and M. Grgic (ed.), *Face Recognition*, pp. 377-416, Vienna, Austria: I-Tech Education and Publishing.

Pantic, M., and Patras, I. (2006). 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*, 36(2), pp.433-49.

Pantic, M. and Rothkrantz, L.J.M. (2004). Case-Based Reasoning for User-Profiled Recognition of Emotions from Face Images. IEEE International Conference on Multimedia and Expo, pp. 391-394.

Picard, R. W. (1997). *Affective Computing*. The MIT Press.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach, in K. R. Scherer and P. Ekman (ed.), *Approaches to emotion*, pp. 293–317. Hillsdale, NJ: Erlbaum.

Scherer, K. R. (2000). Psychological models of emotion, in J. Borod (ed.) *The neuropsychology of emotion*, pp. 137–62, Oxford/New York: Oxford University Press.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking, in K. R. Scherer, A. Schorr, and T. Johnstone (ed.), *Appraisal*

processes in emotion: Theory, Methods, Research, pp. 92–120, New York and Oxford: Oxford University Press.

Scherer, K.R. (In press a). Emotional competence: Conceptual and theoretical issues for modelling agents. Chapter to be published in K. R.Scherer, T. Bänziger, & E. Roesch (ed.) *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford: Oxford University Press.

Scherer, K.R. (In press b). The Component Process Model: Architecture for a comprehensive computational model of emergent emotion. Chapter to be published in K. R.Scherer, T. Bänziger, & E. Roesch (ed.) *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford: Oxford University Press.

Shan, C., Gong, S. and McOwan, P. W. (2007). Beyond Facial Expressions: Learning Human Emotion from Body Gestures. *Proceedings of the British Machine Vision Conference (BMVC'07)*, Warwick, UK.

Stein, B. and Meredith, M.A. (1993). *The Merging of Senses*. MIT Press, Cambridge, USA.

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

Valstar, M.F., Gunes, H. and Pantic, M. (2007). How to Distinguish Posed from Spontaneous Smiles using Geometric Features. ACM International Conference on Multimodal Interfaces (ICMI'07), pp. 38-45, Nagoya, Japan.

Varni, G. (2008). *Multimodal non-verbal interaction based on sound and music. Toward enactive and social interfaces*. Ph.D. Dissertation, University of Genova, 2009.

Varni, G., Camurri, A., Coletta, P. and Volpe, G. (2008). Emotional Entrainment in Music Performance, in *Proceedings 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG2008)*, Amsterdam, The Netherlands, September 2008.

Volpe, G. (2003). *Computational models of expressive gesture in multimedia systems*. Ph.D. Dissertation, Faculty of Engineering, University of Genova, April 2003.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28(6), 879–96.

Wu, L., Oviatt, S. and Cohen, P. (1999). Multimodal integration-a statistical view. *Multimedia, IEEE Transactions on* 1(4), 334-341.

Yeasin, M., Bullot B., and Sharma R. (2006). Recognition of facial expressions and measurement of levels of interest from video, *IEEE Transactions on Multimedia*, 8(3), pp. 500-507.

Zeng, Z., Hu, Y., Liu, M., Fu, Y., and Huang, T.S. (2006). Training Combination Strategy of Multi-Stream Fused Hidden Markov Model for Audio-Visual Affect Recognition. ACM International Conference on Multimedia, pp.65-68.

Zeng, Z., Pantic, M., Roisman, G. I. and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39-58.

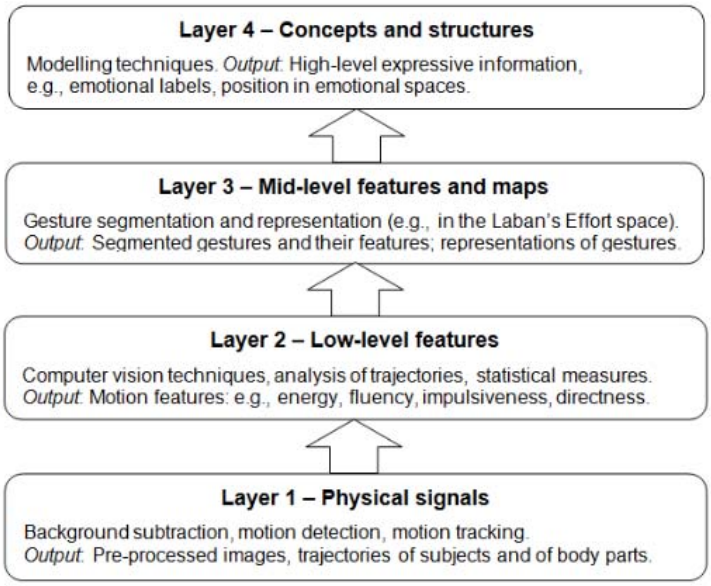
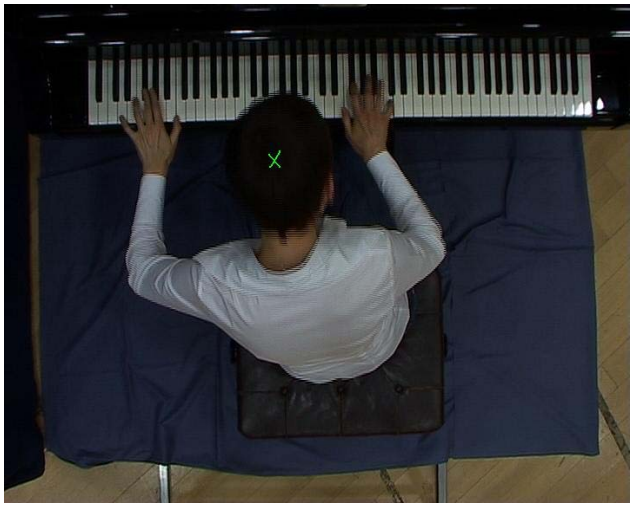


Figure 1. A snapshot of the multilayered conceptual model for analysis of expressive gesture worked out in (Camurri et al., 2005).

(a)



(b)

Figure 2. (a) A measure of quantity of motion using Silhouette Motion Images (the shadow along the pianist's body) and (b) the tracking of the head. From Castellano et al. 2008.