Extracting emotional relevant features from visual signals *Seminar Paper* [*]

Stefan Egli
University of Fribourg
Department of Computer Science
DIVA Group
CH-1700 Fribourg, Switzerland
stefan.egli@unifr.ch

ABSTRACT

Automated Emotion Recognition is an important domain in the research field of Human Computer Interaction. The visual modality plays a major role when it comes to the part of recognize the emotional states of a human by a machine. There has been an intensive research in this field and the aim of this report is to illustrate the different parts involved in the recognition of emotions based on the visual modality.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: General|Signal processing|; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Psychology*; H.1.2 [**Information Systems**]: User/Machine Systems—*Human Factors, Human information processing*

General Terms

Human Factors, Measurement, Performance

1. INTRODUCTION

Emotion Recognition is an important part in the research field of Human Computer Interaction. It would greatly improve our interactions with machines if they could really understand our emotional state and if machines could communicate back in an appropriate way [1].

The task of building an Automated Emotion Recognition System is very challenging and involves many different disciplines. To understand the problems, methodologies and the research which has been done already, we need to have a profound knowledge in every part of this interdisciplinary research field.

First we need to have a deep understanding of the emotional behaviour of humans. What are emotions? How and when do they appear and which are the signals we can measure to get a clue of the expressed emotions? In the past decades there was a huge improvement in the field of Emotion Recognition. But the fundamental questions how do we recognize emotions, how do humans express emotions and how are emotions related to behavioural patterns are much older. If we talk about history we have to mention several big names like: Aristotle began to classify emotions and explained their qualities, Descartes came up with the idea that human behaviour is based on some basic emotions and Darwin who introduced the idea that emotions are linked to their survival value and that emotions were inherited from animal precursors [6].

In this report we first discuss the terminology used for recognition and classification of emotions in the visual modality. We explain which dimensions are involved, the meaning of cues and how emotions are related to bodily expressions and movement qualities.

Then we give an overview over different coding systems for visual cues and how we can extract features from data, such as images or videos. We show which features are important for the classification of emotions and at last we outline performance and challenges we have to face in the field of emotion recognition based on visual signals.

1.1 Emotions

Emotion research involves several scientific disciplines and there is an ongoing discussion about the definition, categorization and the nature of emotions. In general emotions are defined as affectively valenced states. Emotions are distinguished in terms of time from moods and personalities. Whereas emotions are *short-term* (seconds,minutes), moods are *long-term* (hours,days) and personalities are *very-long term* (years, lifetime) [9].

If we speak about emotions we need a way to classify them, we can do this by defining emotions according to dimensions. Paul Ekman devised a list of some basic universally recognized emotions of the face (Happiness, Sadness, Surprise, Fear, Anger, Disgust). Some researchers argue about the number of basic emotions and there is an ongoing discussion. Nevertheless Ekmans theory of universal basic emotions is the most commonly adopted approach in the field of Automatic Emotion Recognition [7].

[*] The Paper is a deliverable of the Msc Research Seminar *Emotion Recognition* in 2011, DIVA Group University of Fribourg, supervised by Dr. D. Lalanne, Dr. F. Ringeval

There are many dimensions which have been proposed, for an incomplete list see Table 1, but there are three standard dimensions. These dimensions are *Valence*, *Arousal* and *Potency* or also described as evaluation-pleasantness, potency-control and activation-arousal.

A widely used theory is the two dimensional approach from Russell. In his theory, emotions are related to two dimensions, valence and arousal. *Valence* describes if an emotion is positive or negative. *Arousal* is an indicator of how strong an emotion is performed [9]. Nevertheless there are several different opinions on the dimensional corpus, in most research projects the two component model of Russell is used [7].

1.2 From Motion to Emotion

The fundamental process of Emotion Recognition from body movements and postures is still a unresolved problem. Recent psychological studies showed that humans, which judge about others behaviour, rely on combined visual channels of face and body more than any other channel [2]. To extract useful features for classifying emotions, we first have to know what we have to track and how our body expresses emotions. There has been an intensive research in this field and studies have shown that there is a relation between specific movement qualities, postures and emotions [1][2].

1.3 Visual Cues

In the visual modality, *cues are non-verbal signals* based on movements of the body. This contains relaxation/contraction of muscles, movements of a body part, or group of body parts. Those cues give us an indication on the subjacent emotions of a human. There are many cues involved and it makes sense to divide them into three main categories which are meaningful for the visual modality [4].

1.3.1 Facial expressions

Facial expressions are the most indicative and most studied expression in the field of behavioural research. They involve facial cues which are displayed using body parts from the head. Those cues form Facial Expressions which are recognized as emotions. For example, raising our lips (facial cues) is a part of a smile (facial expression) which can be recognized as happiness (emotion).

1.3.2 Body postures

Another part of non-verbal communication are body postures which are displayed moving different body parts such as legs, arms, torso, in a specific moment. For example, clenching of the fist and raising it is a indicator of anger. Postures are static, like a snapshot of a gesture.

1.3.3 Gestures

Gestures are another part of non-verbal communication, they are dynamic, which means a gestures involves a sequence of movements. Those actions use also the motion of the limbs to communicate an intention or feeling but commonly they are originated from the face or hand. They include body movements and postures. Those cues are also labelled as motion or motor cues or even more general as kinematic cues [1].

1.4 Variabilities

Until now we were talking about models and definitions which are mostly general and theoretical. In the real world we have to deal with much more complex situations and constantly changing conditions. As all humans are different, in a physical and psychical way, we have to test if the exposed theories apply with those variations. Emotions are not generally understandable and recognizable, there are some variations which make a clear classification or recognition more difficult. In the visual modality we have to take care about variations such as: different cultural behavioural, gender, age, pathological conditions and how they affect the recognition and classification of emotions.

There have been several studies about the cultural impact on emotions. The study on *cross-cultural comparison* showed that the cultural context affects the recognition of emotion but mostly it affects vocal cues and it has a less significant impact on visual information [5]. Another problem is that humans are able to feel multiple emotions at the same time. So there will be an transition from one emotional state to another. It can also be happen that several emotions are blended together at the same time which results in a combined emotional expression which will be hard to track. This process is really complicated and it is still not understood well today [4].

2. CODING SYSTEMS OF VISUAL CUES

To get a better understanding and a standardized description of body movements which are relate to specific emotions researchers developed coding systems. There are several modalities which have been studied, like body movement and gestures, Ekman et al. developed the well known Facial Action Coding System (FACS) which is used to describe changes, contraction or relaxations, of muscles, in the appearance of the face. This system is used to code expressions from static or dynamic pictures [3][6]. The coding system is based on so called *Action Units* (AU) which are a description for component movement or facial actions. The combination of AU leads to facial expressions. There are Coding Systems for hands, body posture and face [8].

2.1 Hand Action Coding System (HACS)

HACS is a hierarchical high-level model for hand motion simulation used in gesture analysis. It codes hand gestures in terms of hand muscle action units. It describes the relationship between hand functions and the corresponding muscle contraction/relaxations.

2.2 Body Posture Coding System (BPCS)

Deal et al. developed a Body Action and Posture coding system [10]. Their approach is based on a three level coding system including an anatomical level for the description of different articulations of body parts, a form level as an indicator of the direction and the movement, and a functional level which indicates communicative and self-regulatory functions.

2.3 Facial Action Coding System (FACS)

The goal of FACS is to develop a comprehensive system which can identify all possible visual distinguishable facial

Table 1: Dimensions in Emotion Recognition

Dimension	Description	+/-	Emotions
Arousal	Grad of activation, strength, intensity	High, Low	Rage, Depression
Valence	How positive or negative the emotion is	Positive, Negative	Anger, Happiness
Potency	The power or sense of control over emotions	High, Low	Fear, Panic
Social	The emotion is directed to oneself or to others	Social, Non-Social	Love, Happiness
Engagement	Grad of Binding	High, Low	Caring, Phlegm
Unpredictability	Urgent reactions, surprise, familiar or unfamiliar situation	High, Low	Surprised, Bored

movements. In other words, FACS is an index of facial expressions. This system is based on so called (AU) which are basic actions of specific single muscles or groups of muscles and *Action Descriptors* which are movements who can involve several muscle groups. Almost every anatomically possible facial expression can be deconstructed into those AU because they are independent and combinable. The system provides also an intensity scoring which indicates the strength of an AU [3][4][6].

3. FEATURE DETECTION AND DESCRIPTION

After the definition of cues, dimensions and emotional states we need to lead our focus on the detection of interesting regions in the visual data. We define an interesting region as a part of the face and body which give us a hint about the underlying emotion. The coding system above give us a good lead on such regions. The detection of interesting regions is a part of computer vision. It involves the following steps.

3.1 Finding body parts

The first task is to detect or track the body part under consideration. To segregate only the useful information from the input data, well known algorithms, filters and procedures from image processing are used, like: background subtraction, active contours, skin segmentation, camshift tracking. This process is not easy as we have to deal with a lot of variabilities such as, bad input data (noisy, bad lighting condition), physical differences among the suspects (size, skin color). After detecting a body part we need to extract the information based on the displayed body signal. In this process we have to deal with several problems like: is the information static or temporal (picture, movie), does the feature involve the whole body part or just a sub-part (i.e. whole face, only mouth) and is the information view oder volume based (2D/3D) [8].

3.2 Extract data from the displayed signals

After the input data has been preprocessed we are finally able to extract the desired features. Those features can be *geometric*, for example the shape of a body part or the locations of the fiducial markers (point of reference). Or they are *appearance-based* which includes for example the texture of the skin, bulges or furrows. Geometric feature extracting leads to better results than appearance based due to the fact that the latter are more difficult to track. After a successful extraction the information can be classified. [8]

3.3 Signal Descriptors

Extracted features need to be analyzed to determine the underlying emotion. This analysis can be done in two ways either statically or dynamically. If we do it statically we generally use a single image of a face at the apex of the emotional expression. This approach is also called *target-oriented* [11]. The goal is to detect static cues like wrinkles or position, form or shapes of different features.

On the other hand if a dynamic analysis of the signal is done we have to monitor and measure the variations of these features over time. This is based on the timing and duration of different AU. It is also called gesture-oriented analysis as we have to deal with sequences of frames of expressions. A common approach is based on optical-flow method which is used for detection motion of objects or surfaces. In regions of interests motion fields are computed, analyzed (based on motion templates) and then mapped to emotions. The descriptors for such signals are the amplitude, duration, attack and release of actions and occurrences (timing).

For humans the dynamics of body expressions are important since temporal information is essential for recognize complex emotional states such as moods or attitudes [3][6][8].

In other studies [1] they extracted energy and perimeter features and showed that energy features are a better indication for emotion condition than perimeter features. The maximum value of the signal and its duration is an indicator of the impulsiveness of the motion. The shape of the peak determines the emotional condition. The body spatial occupation can be used to discriminate between emotions.

4. RECOGNITION PERFORMANCE

Despite Facial Emotion Recognition is the most widely used and best performing technique it is not perfect and to get better results we can combine different cues to gain better recognition results. Some studies combine different modalities like audio cues and visual cues,which lead to better results [9]. In this report we are interested in how combined visual cues perform. The research done by Gunes et al. showed that combined cues lead to a better recognition accuracy in general. They tested the combination of Facial Expression Cues and Gestures Cues of the upper body. Based on their bi-modal database they first extracted Facial and Body features from data sequences [2].

4.1 Mono-modal

In this approach they used each modality separated to run the classification. The conclusion in this mono-modal approach was that body movements were better recognized due to the fact that facial movements are small and therefore more difficult to recognize [2].

4.2 Bi-modal

Table 2: Emotion Recognition Performance

Modality		Data Source		Fusion	Rate in %
Face	Body	Frame	Movie		
o	o	o		feature level	94.0%
o	o	o		decision level	91.1%
o		o			89.9%
o	o		o		85.0%
o	o	o			82.0%
	o	o			77.0%
	o	o			76.4%
o		o			35.0%

The fusion of different modalities lead to a combined representation of those modalities. The key point is, when to combine the information. There are two variations. *Feature-level fusion*, where features from both modalities are merged into a larger feature vector which is then passed to the classifier. *Decision-level fusion*, which classifies each modality independently before the classification is done on the fused feature vector. Studies showed, that the recognition based on the bi-modal approach with a feature-level fusion leads to a very good performance rate (Table 2) [2][9].

5. CONCLUSIONS

Most studies in this research field depend on a data corpus of portrayed emotions from actors. This approach of using a ideal data source is not transferable to real world applications. For humans detecting and analyzing body expressions such as facial expressions is easy, but the development of an Automated Emotion Recognition System which can handle this task, with a good accuracy, is very complex. Such a system would be a great improvement in Human Computer Interaction.Today such systems or research studies still have the same limitations listed below.

- handle only a small set of emotion
- only portrayed, exaggerated expressions of emotions are recognized
- no context-sensitive analysis
- not analyzed according to different time scales (emotion, mood, attitude)
- based on ideal databases, image or videos were recorded under ideal conditions

With this limitations most recognition systems achieve very good results, the best Automated Facial Emotion Recognition system can recognize up to 27 of 44 AU. Future research will focus more on the development of Spontaneous Emotion Recognition Systems [8].

This paper shows just a small part of the whole research field of Emotion Recognition based on visual signals. We realized that this field is huge and it involves many different disciplines which makes it even more difficult. It is not easy to get deep into this topic, there are some definitions and theories which are really confusing in the beginning. Nevertheless the improvements in the last decades shows us the possibilities of this young research field. Current recognition systems still lack the points we mentioned in the chapter before and they are far away from being perfect. We are exited and we are looking forward to see new and better system in the future.

6. REFERENCES

1. Glowinski, D.; Camurri, A.; Volpe, G.; Dael, N.; Scherer, K.; , "Technique for automatic emotion recognition by body gesture analysis," Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on , vol., no., pp.1-6, 23-28 June 2008

2. Hatice Gunes, Massimo Piccardi, Bi-modal emotion recognition from expressive face and body gestures, Journal of Network and Computer Applications, Volume 30, Issue 4, November 2007, Pages 1334-1345

3. Essa, I.A.; Pentland, A.P.; , "Coding, analysis, interpretation, and recognition of facial expressions," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.19, no.7, pp.757-763, Jul 1997

4. Tan, S. C. G., and Nareyek, A. 2009. Integrating Facial, Gesture, and Posture Emotion Expression for a 3D Virtual Agent. In Proceedings of the 14th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGames 2009 USA), 23-31

5. Riviello, M.T.; Esposito, A.; Chetouani, M.; Cohen, D.; , "Inferring emotional information from vocal and visual cues: A cross-cultural comparison," Cognitive Infocommunications (CogInfoCom), 2011 2nd International Conference on , vol., no., pp.1-4, 7-9 July 2011

6. F. Fragopanagos and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," Neural Networks, vol. 18, pp. 389-405, 2005.

7. R. Cowie, N. Sussman, and A. Ben-Ze'ev, "Emotions: concepts and definitions" in Humaine Handbook on Emotion, P. Petta, Ed., Springer, Berlin, Germany, 2010

8. M. Pantic and G. Caridakis, "Image and Video Processing for Affective Applications" Emotion-Oriented Systems: The Humaine Handbook, pp. 101-117, Springer-Verlag, 2011.

9. H. Gunes, M. Piccardi, M. Pantic, From the lab to the real world: affect recognition using multiple cues and modalities, in: J. Or (Ed.), Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition, 2008, pp. 185-218..

10. Dael, N., Mortillaro, M., & Scherer, K. R. (in press). The Body Action and Posture coding system (BAP): Development and reliability. Journal of Nonverbal Bahevior.

11. Sreevatsan, A. N.; Kumar, K. G. Sathish; Rakeshsharma, S. & Roomi, Mohd. Mansoor: Emotion Recognition from Facial Expressions: A Target Oriented Approach Using Neural Network. Allied Publishers Private Limited (2004) , S. 497-502 .