Towards Affective Sensing

Gordon McIntyre[1] and Roland Goecke[1,2]

[1] Department of Information Engineering, Research School of Information Sciences
and Engineering, Australian National University,
[2] NICTA Canberra Research Laboratory*,
Canberra, Australia
gordon.mcintyre@anu.edu.au, roland.goecke@nicta.com.au
URL: http://users.rsise.anu.edu.au/~gmcintyr/index.html

Abstract. This paper describes ongoing work towards building a multimodal computer system capable of sensing the affective state of a user. Two major problem areas exist in the affective communication research. Firstly, affective states are defined and described in an inconsistent way. Secondly, the type of training data commonly used gives an oversimplified picture of affective expression. Most studies ignore the dynamic, versatile and personalised nature of affective expression and the influence that social setting, context and culture have on its rules of display. We present a novel approach to affective sensing, using a generic model of affective communication and a set of ontologies to assist in the analysis of concepts and to enhance the recognition process. Whilst the scope of the ontology provides for a full range of multimodal sensing, this paper focuses on spoken language and facial expressions as examples.

1 Introduction

As computer systems form an integral part of our daily life, the issue of user-adaptive human-computer interaction systems becomes more important. In the past, the user had to adapt to the system. Nowadays, the trend is clearly towards more human-like interaction through user-sensing systems. Such interaction is inherently multimodal and it is that integrated multimodality that leads to robustness in real-world situations. One new area of research is affective computing, i.e. the ability of computer systems to sense and adapt to the affective state (colloquially 'mood', 'emotion', etc.)[1] of a person.

According to its pioneer, Rosalind Picard, 'affective computing' is computing that relates to, arises from, or deliberately influences emotions [1]. Affective sensing attempts to map measurable physical responses to affective states. Several studies have successfully mapped strong responses to episodic emotions. However, most studies take place in a controlled environment, ignoring the importance that social settings, culture and context play in dictating the display

* National ICT Australia is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

[1] The terms affect, affective state and emotion, although not strictly the same, are used interchangeably in this paper.

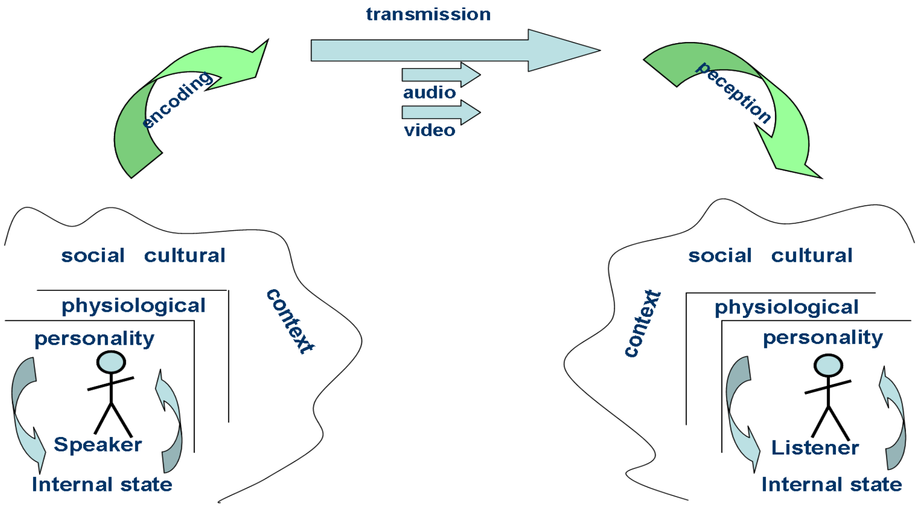


Fig. 1. A generic model of affective communication

rules of affect. In a natural setting, emotions can be manifested in many ways and through different combinations of modalities. Further, beyond niche applications, one would expect that affective sensing must be able to detect and interpret a wide range of reactions to subtle events.

In this paper, a novel approach is presented which integrates a domain ontology of affective communication to assist in the analysis of concepts and to enhance the recognition process. Whilst the scope of the ontology provides for a full range of multimodal sensing, our work to date has concentrated on sensing in the audio and video modalities, and the examples given relate to these.

The remainder of the paper is structured as follows. Section 2 explains the proposed framework. Section 3 gives an overview of the ontologies. Section 4 details the application ontology. Section 5 explains the system developed to automatically recognise facial expressions and uses it as an example for applying the ontologies in practice. Finally, Section 6 provides a summary.

2 A Framework for Research in Affective Communication

The proposed solution consists of 1) a generic model of affective communication; and 2) a set of *ontologies*. An ontology is a statement of concepts which facilitates the specification of an agreed vocabulary within a domain of interest.

The model and ontologies are intended to be used in conjunction to describe

1. affective communication concepts,
2. affective computing research, and
3. affective computing resources.

Figure 1 presents the base model on the example of emotions in spoken language. Firstly, note that it includes speaker and listener, in keeping with the Brunswikian lens model as proposed by Scherer [2]. The reason for modelling attributes of both speaker and listener is that the listener's cultural and social presentation vis-à-vis the speaker may also influence judgement of emotional content. Secondly, note that it includes a number of factors that influence the expression of affect in spoken language. Each of these factors is briefly discussed and motivated in the following. More attention is given to context as this is seen as a much neglected factor in the study of automatic affective state recognition.

2.1 Factors in the Proposed Framework

Context. Context is linked to modality and emotion is strongly multimodal in the way that certain emotions manifest themselves favouring one modality over the other [3]. Physiological measurements change depending on whether a subject is sedentary or mobile. A stressful context such as an emergency hot-line, air-traffic control, or a war zone is likely to yield more examples of affect than everyday conversation.

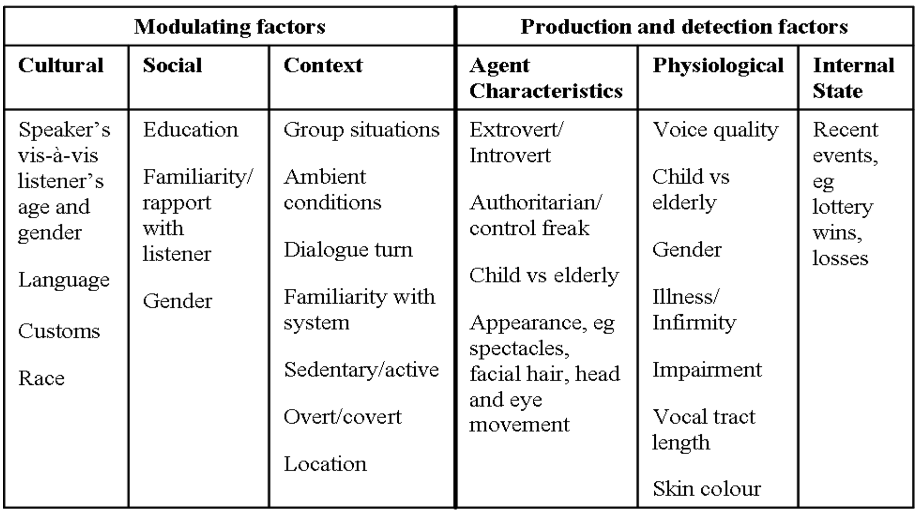
Stibbard [4] recommends *"...the expansion of the data collected to include relevant non-phonetic factors including contextual and inter-personal information."* His findings underline the fact that most studies so far took place in an artificial environment, ignoring social, cultural, contextual and personality aspects which, in natural situations, are major factors modulating speech and affect presentation. The model depicted in Figure 1 takes into account the importance of context in the analysis of affect in speech.

Recently, Devillers *et al.* [5] included context annotation as metadata to a corpus of medical emergency call centre dialogues. Context information was treated as either task-specific or global in nature. The model proposed in this paper does not differentiate between task-specific and global context as the difference is seen merely as temporal, i.e. pre-determined or established at "run-time".

Other researchers have included "discourse context" such as speaker turns [6] and specific dialogue acts of greeting, closing, acknowledging and disambiguation. Inclusion in a corpus of speaker turns would be useful but annotation of every specific type of dialogue act would be extremely resource intensive.

The HUMAINE project [7] included a proposal that at least the following issues be specified:

- Agent characteristics (age, gender, race)
- Recording context (intrusiveness, formality, etc.)
- Intended audience (kin, colleagues, public)
- Overall communicative goal (to claim, to sway, to share a feeling, etc.)
- Social setting (none, passive other, interactant, group)
- Spatial focus (physical focus, imagined focus, none)
- Physical constraint (unrestricted, posture constrained, hands constrained)
- Social constraint (pressure to expressiveness, neutral, pressure to formality)



Modulating factors			Production and detection factors		
Cultural	**Social**	**Context**	**Agent Characteristics**	**Physiological**	**Internal State**
Speaker's vis-à-vis listener's age and gender Language Customs Race	Education Familiarity/ rapport with listener Gender	Group situations Ambient conditions Dialogue turn Familiarity with system Sedentary/active Overt/covert Location	Extrovert/ Introvert Authoritarian/ control freak Child vs elderly Appearance, eg spectacles, facial hair, head and eye movement	Voice quality Child vs elderly Gender Illness/ Infirmity Impairment Vocal tract length Skin colour	Recent events, eg lottery wins, losses

Fig. 2. Use of the model in practice

but went on to say, *"It is proposed to refine this scheme through work with the HUMAINE databases as they develop."* Millar *et al.* [8] developed a methodology for the design of audio-video data corpora of the speaking face in which the need to make corpora re-usable is discussed. The methodology, aimed at corpus design, takes into account the need for *speaker* and *speaking environment* factors.

In contrast, the model presented in this paper treats agent characteristics and social constraints separate to context information. This is because their effects on discourse are seen as separate topics for research. It is evident that

1. Context is extremely important in the display rules of affect;
2. Yet, defining context annotation is still in its infancy.

Agent characteristics. As Scherer [2] points out, most studies are either speaker oriented or listener oriented, with most being the former. This is significant when you consider that the emotion of someone labelling affective content in a corpus could impact the label that is ascribed to a speaker's message.

The literature has not given much attention to the role that agent characteristics such as personality type play in affective presentation which is surprising when one considers the obvious difference in expression between extroverted and introverted types. Intuitively, one would expect a marked difference in signals between speakers. One would also think that knowing a person's personality type would be of great benefit in applications monitoring an individual's emotions.

At a more physical level, agent characteristics such as facial hair, whether they wear spectacles, and their head and eye movements all affect the ability to visually detect and interpret emotions.

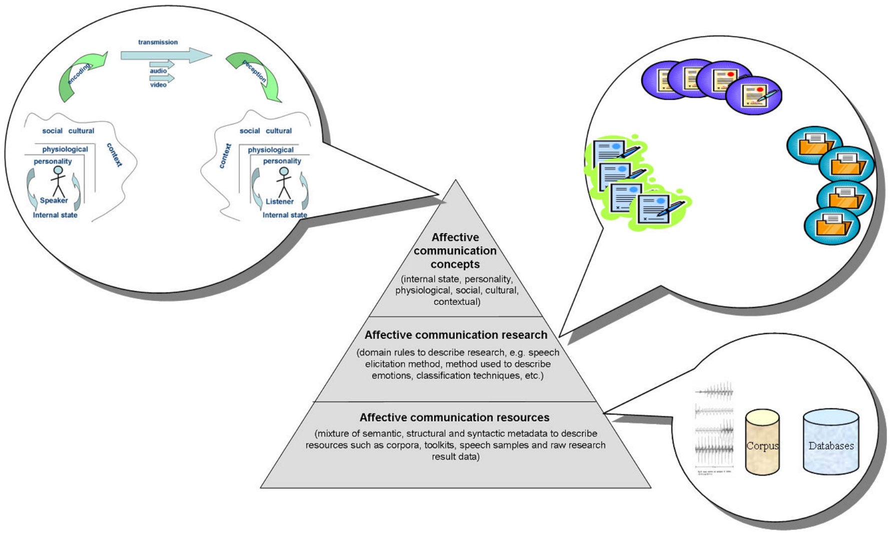

transmission
encoding
audio
video
deception
social cultural
physiological
personality
context
Speaker
Internal state
context
social cultural
physiological
personality
Listener
Internal state
Affective communication concepts
(internal state, personality, physiological, social, cultural, contextual)
Affective communication research
(domain rules to describe research, e.g. speech elicitation method, method used to describe emotions, classification techniques, etc.)
Affective communication resources
(mixture of semantic, structural and syntactic metadata to describe resources such as corpora, toolkits, speech samples and raw research result data)
Corpus
Databases

Fig. 3. A set of ontologies for affective computing

Cultural. Culture-specific display rules influence the display of affect [3]. Gender and age are established as important factors in shaping conversation style and content in many societies.

Studies by Koike *et al.* [9] and Shigeno [10] have shown that it is difficult to identify the emotion of a speaker from a different culture and that people will predominantly use visual information to identify emotion. Putting it in the perspective of the proposed model, cognisance of the speaker and listener's cultural backgrounds, the context, and whether visual cues are available, obviously influence the effectiveness of affect recognition.

Physiological. It might be stating the obvious but there are marked differences in speech signals and facial expressions between people of different age, gender and health. The habitual settings of facial features and vocal organs determine the speaker's range of possible visual appearances and sounds produced. The configuration of facial features, such as chin, lips, nose, and eyes, provide the visual cues, whereas the vocal tract length and internal muscle tone guide the interpretation of acoustic output [8].

Social. Social factors temper spoken language to the demands of civil discourse [3]. For example, affective bursts are likely to be constrained in the case of a minor relating to an adult, yet totally unconstrained in a scenario of sibling rivalry. Similarly, a social setting in a library is less likely to yield loud and extroverted displays of affect than a family setting.

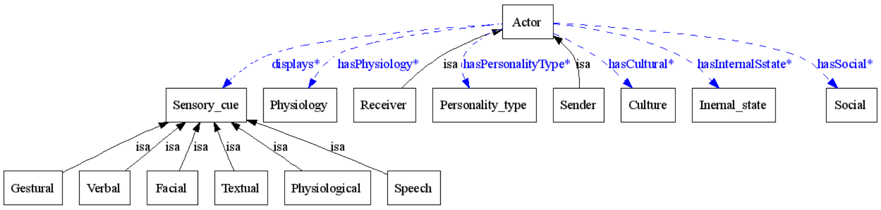


Fig. 4. A fragment of the domain ontology of concepts

Internal state. Internal state has been included in the model for completeness. At the core of affective states is the person and their experiences. Recent events such as winning the lottery or losing a job are likely to influence emotions.

2.2 Examples

To help explain the differences between the factors that influence the expression of affect, Figure 2 lists some examples. The factors are divided into two groups. On the left, is a list of factors that modulate or influence the speaker's display of affect, i.e. cultural, social and contextual. On the right, are the factors that influence production or detection in the speaker or listener, respectively, i.e. personality type, physiological make-up and internal state.

3 A Set of Ontologies

The three ontologies described in this paper are the means by which the model is implemented and are currently in prototype form. Figure 3 depicts the relationships between the ontologies and gives examples of each of them. Formality and rigour increase towards the apex of the diagram. The types of users are not confined solely to researchers. There could be many types of users such as librarians, decision support systems, application developers and teachers.

3.1 Ontology 1 - Affective Communication Concepts

The top level ontology correlates to the model discussed in Section 2 and is a formal description of the domain of affective communication. It contains internal state, personality, physiological, social, cultural, and contextual factors. It can be linked to external ontologies in fields such as medicine, anatomy, and biology. A fragment of the top-level, domain ontology of concepts in shown in Figure 4.

3.2 Ontology 2 - Affective Communication Research

This ontology is more loosely defined and includes the concepts and semantics used to define research in the field. It has been left generic and can be further subdivided into an affective computing domain at a later stage, if needed. It is used to specify the rules by which accredited research reports are catalogued. It includes metadata to describe, for example,

- classification techniques used;
- the method of eliciting speech, e.g. acted or natural; and
- manner in which corpora or results have been annotated, e.g. categorical or dimensional.

Creating an ontology this way introduces a common way of reporting the knowledge and facilitates intelligent searching and reuse of knowledge within the domain. For instance, an ontology just based on the models described in this paper could be used to find all research reports where:

> SPEAKER(internalState='happy',
> physiology='any',
> agentCharacteristics='extrovert',
> social='friendly',context='public',
> elicitation='dimension')

Again, there are opportunities to link to other resources. As an example, one resource that will be linked is the Emotion Annotation and Representation Language (EARL) which is currently under design within the HUMAINE project [11]. EARL is a XML-based language for representing and annotating emotions in technological contexts. Using EARL, emotional speech can be described either using a set of forty-eight categories, dimensions or even appraisal theory. Examples of annotation elements include "Emotion descriptor" - which could be a category or a dimension, "Intensity" - expressed in terms of numeric values or discrete labels, "Start" and "End".

3.3 Ontology 3 - Affective Communication Resources

This ontology is more correctly a repository containing both formal and informal rules, as well as data. It is a combination of semantic, structural and syntactic metadata. This ontology contains information about resources such as corpora, toolkits, audio and video samples, and raw research result data.

The next section explains the bottom level, application ontology used in our current work in more detail.

4 An Application Ontology for Affective Sensing

Figure 5 shows an example application ontology for affective sensing in a context of investigating dialogues. During the dialogue, various events can occur, triggered by one of the dialogue participants and recorded by the sensor system. These are recorded as time stamped instances of events, so that they can be easily identified and distinguished. In this ontology, we distinguish between two

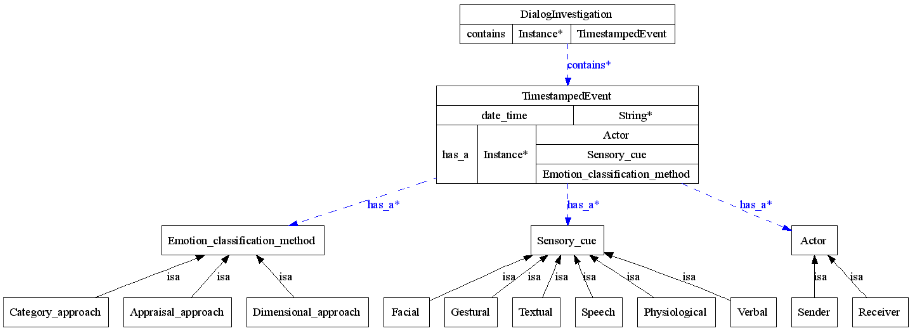


Fig. 5. An application ontology for affective sensing

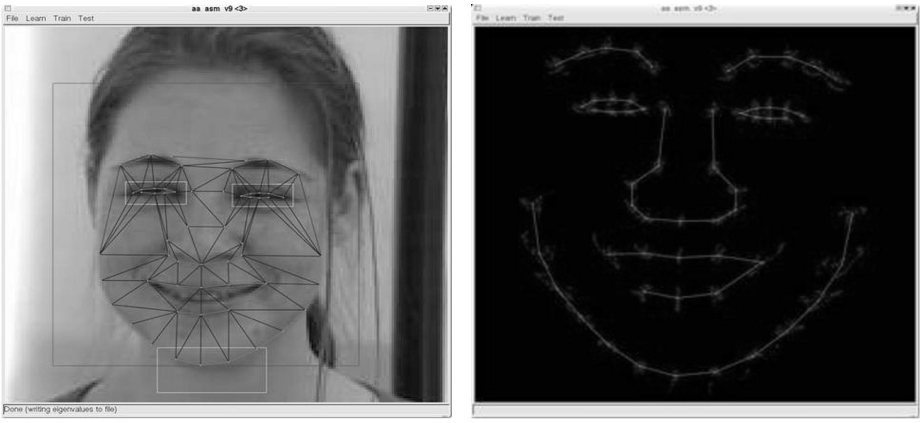
roles for each interlocutor: sender and receiver, respectively. At various points in time, each interlocutor can take on different roles. On the sensory side, we distinguish between facial, gestural, textual, speech, physiological and verbal[3] cues. This list, and the ontology, could be easily extended for other cues and is meant to serve as an example here, rather than a complete list of affective cues. Finally, the emotion classification method used in the investigation of a particular dialogue is also recorded.

We use this ontology to describe our affective sensing research in a formal, yet flexible and extendible way. In the following section, a brief description of the facial expression recognition system developed in our group is given as an example of using the ontologies in practice.

5 Automatic Recognition of Facial Expressions

Facial expressions can be a major source of information about the affective state of a person and they are heavily used by humans to gauge a person's affective state. We have developed a software system – the Facial Expression Tracking Application (FETA) [12] – to achieve automatic facial expression recognition. It uses statistical models of the permissible shape and texture of faces as found in images. The models are learnt from labelled training data, but once such models exist, they can be used to automatically track a face and its facial features. In recent years, several methods using such models have been proposed. A popular method are the Active Appearance Models (AAM) [13]. AAMs are a generative method which model non-rigid shape and texture of visual objects using a low-dimensional representation obtained from applying principle component analysis to a set of labelled data. AAMs are considered to be the current state-of-the-art in facial feature tracking and are used in the FETA system, as they provide a fast and reliable mechanism for obtaining input data for classifying facial expressions.

[3] The difference between speech and verbal cues here being spoken language versus other verbal utterings.



Done (writing eigenvalues to file)

Fig. 6. Left: An example of an AAM fitted to a face from the FGnet database. Right: Point distribution over the training set.

As a classifier, the FETA system uses artificial neural networks (ANN). The ANN is trained to recognise a number of facial expressions by the AAM's shape parameters. In the work reported here, facial expressions such as neutral, happy, surprise and disgust are detected. Other expressions are possible but the FGnet data corpus [14] used in the experiments was limited to these expressions. As this paper is concerned with a framework for affective computing, rather than a particular method for affective sensing and a set of experiments, we omit experimental results of the automatic facial expression recognition experiments, which the interested reader can find in [12].

In the domain ontology of concepts, we would list this work as being on the facial sensory cue with one person present at a time. As the data in the FGnet corpus is based on subjects being asked to perform a number of facial expressions, it would be recorded as being acted emotions in the ontology. Recordings were made in a laboratory environment, so the context would be 'laboratory'. One can easily see how the use of an ontology facilitates the capture of important metadata in a formalised way. Following the previous example of an application ontology, we would record the emotion classification method (by the corpus creators) as being the category approach. The resources provided in the FGnet corpus are individual images stored in JPEG format. Due to space limits in this paper, we will not describe the entire set of ontologies for this example. The concept should be apparent from the explanation given here.

6 Conclusions and Future Work

We have presented ongoing work towards building an affective sensing system. The main contribution of this paper is a proposed framework for research in affective communication. This framework consists of a generic model of affective communication and a set of ontologies to be used in conjunction. Detailed descriptions of the ontologies and examples of their use have been given. Using the proposed framework provides an easier way of comparing methodologies and results from different studies of affective communication.

In future work, we intend to provide further example ontologies on our webpage. We will also continue our work on building a multimodal affective sensing system and plan to include physiological sensors as another cue for determining the affective state of a user.

References

1. Picard, R.: Affective Computing. MIT Press, Cambridge (MA), USA (1997)
2. Scherer, K.: Vocal communication of emotion: A review of research paradigms. Speech Communication **40**(1–2) (April 2003) 227–256
3. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond emotion archetypes: Databases for emotion modelling using neural networks. Neural Networks **18**(4) (May 2005) 371–388
4. Stibbard, R.: Vocal expression of emotions in non-laboratory speech: An investigation of the Reading/Leeds Emotion in Speech Project annotation data. PhD thesis, University of Reading, United Kingdom (2001)
5. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks **18**(4) (May 2005) 407–422
6. Liscombe, J., Riccardi, G., Hakkani-Tür, D.: Using context to improve emotion detection in spoken dialog systems. In: Proceedings of the 9th European Conference on Speech Communication and Technology EUROSPEECH'05. Volume 1., Lisbon, Portugal (September 2005) 1845–1848
7. HUMAINE: http://emotion-research.net/, Last accessed 26 October 2006.
8. Millar, J., Wagner, M., Goecke, R.: Aspects of Speaking-Face Data Corpus Design Methodology. In: Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004. Volume II., Jeju, Korea (October 2004) 1157–1160
9. Koike, K., Suzuki, H., Saito, H.: Prosodic Parameters in Emotional Speech. In: Proc. 5th International Conference on Spoken Language Processing ICSLP'98. Volume 2., Sydney, Australia, ASSTA (December 1998) 679–682
10. Shigeno, S.: Cultural similarities & differences in the recognition of audio-visual speech stimuli. In Mannell, R., Robert-Ribes, J., eds.: Proceedings of the International Conference on Spoken Language Processing ICSLP'98. Volume 1., Sydney, Australia, ASSTA (December 1998) 281–284
11. Schröder, M.: D6e: Report on representation languages http://emotionresearch.net/deliverables/D6efinal, Last accessed 26 October 2006.
12. Arnold, A.: Automatische Erkennung von Gesichtsausdrücken auf der Basis statistischer Methoden und neuronaler Netze. Masterthesis, University of Applied Sciences Mannheim, Germany (2006)
13. Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. In Burkhardt, H., Neumann, B., eds.: Proceedings of the European Conference on Computer Vision ECCV'98. Volume 2 of Lecture Notes in Computer Science 1406., Freiburg, Germany, Springer-Verlag (June 1998) 484–498
14. Wallhoff, F.: Facial Expressions and Emotion Database. http://www.mmk.ei.tum.de/∼waf/fgnet/feedtum.html, Last accessed 6 December 2006 Technische Universität München.