MULTI-MODAL EMOTION RECOGNITION – MORE "COGNITIVE" MACHINES

Velina Slavova, Hichem Sahli, Werner Verhelst

***Abstract**: Based on several results related to studies on emotions, we suggest that the process of emotion-recognition is assisted by some internal structure of the cognitive images of emotions, which are at different levels of knowledge representation. We concede that the main proposed in psychology models are in correspondence with these levels and in this sense - complementary. We summarize the state-of-the-art of machine emotion recognition with regards of the used psychological models of emotions. In order to discover amelioration sources of multimodal machine emotion recognition, we propose a scheme of the cognitive process, based on gradual levels of representation. The proposed scheme shows several "strategic" differences with the architectures used in machine emotion recognition. We discuss the questions related to recognition, assisted by two levels of representation that we called "perceptual" and "conceptual".*

***ACM Classification Keywords**: I.2 Artificial Intelligence, 1.2.0.Cognitive simulation*

Introduction and Development of Previous Hypothesis

Recent advances in human-computer interaction (HCI) show the importance of applying knowledge from different disciplines in order to make machines more "intelligent". The wish to produce machines with more and more "intelligence" necessitates providing them the capacity of sensing human emotions. During the last years the research concentrated in all these problems. An example of that is the Human-Machine Interaction Network on Emotion (HUMAINE) – a network of excellence that aims to lay the foundations for European development of systems that can register, model and/or influence human emotional and emotion-related states and processes - 'emotion-oriented systems'. [*http://emotion-research.net/].*

Emotion is a mental[1] and physiological state associated with a wide variety of feelings, thoughts, and behaviour. The associated physiological state is often accompanied by physiological changes which can lead to modifications in persons' observable and measurable manifestations, called expressions. Expressions can be perceived and evaluated by others as evidences of given emotional state.

Emotions are subject of study in several disciplines - psychology, cognitive science, philosophy and computer science. Although there have been numerous studies with regards to both the psychological and the computaional aspect of emotions, it is still not clear how to define and how to categorize them. There are two basic theoretical "models" coming from psychology. The first model is "discrete" (Fig. 1, A). Emotion categories are determined as entities with names and descriptions. Several researchers (see Ekman 1992) argue that a few emotions are basic or primary (*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*). This approach is very convenient for implementation purposes as it provides "ready-on" classes for the machine learning and classifiers. The second model represents emotions as having certain *properties* in a continuous space, on axes as pleasant–unpleasant, active-passive, attention– rejection, simple–complicated, etc. (ex. Schlosberg 1954). Two commonly used dimension-axes are "*valence*" and "*arousal*" (fig. 1, B), where valence describes the pleasantness of the stimuli. Although the consensus of the experts from HUMAINE is that "labelling schemes based on traditional

1 The term *Mental state* is often defined as "state of a person's cognitive processes".

divisions of emotion into ‘basic’ emotions is not relevant", the most part of the systems for emotion recognition are based on the discrete model. There are not still clearly defined strategies for the application of the continuous model. As the performance of automatic recognition system is totally dependent on the number and the degree of differentiation of the emotion categories or dimensions that have to be discriminated, the question about the adequate use of these two models stays central.

Here we raise the hypotheses that the two theoretical models of emotions correspond to different levels of representing the information, or, levels of abstraction. We consider three levels in the information-representation hierarchy: sensory, perceptual and conceptual (figure 1), as it is largely accepted in cognitive science. We assume that the emotion categories *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* etc. are language-names of concepts that humans give to emotional states. In this sense, they correspond to the highest level of abstraction. The "dimensions" model describes emotional states according to their properties. We suppose that, just like the concept of "*snow*" possess perceptual properties like "*cold*" and "*white*", the concept of "*anger*" has properties "*negative*" and "*active*". In correspondence, the "dimensions" model can be associated to the *perceptual* level of abstraction. According to a known work in cognitive science by Barsalou et al. (2003), both levels – the conceptual and the perceptual, are related to conceptual knowledge.

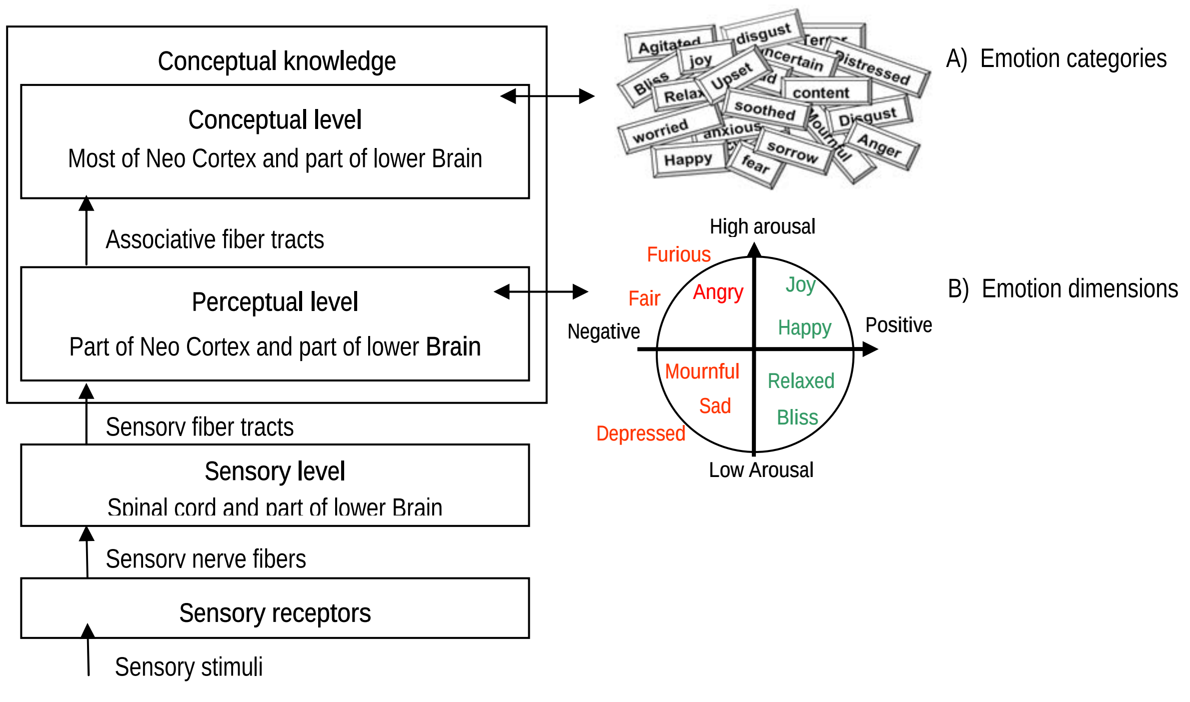

Fig. 1. Levels of abstraction in a cognitive hierarchy and psychological models of emotion (A and B)

In this representation the two models are complementary and dependent. Based on the fact that there are several evidences (see Kim 2004) for biological changes that are correlated with emotion dimensions, we supposed that there exist a conceptual knowledge about the emotion properties. Moreover, the analyses of several results from experiments with human participants from different cultures (Wan et al.,2005) lead to the conclusion that there exists a clear notion about the significance of the axes describing the emotional properties. As an example, Figure 2 shows how human subjects evaluate the concepts of "*anger*" and "*happiness*".

As shown in figure 3, subjects' evaluation shapes clusters that cover more general areas of the plane. We raise the hypothesis about the existence, on the perceptual level of representation, of more general emotional categories, as shown in figure 4. The claim is that the conceptual knowledge about emotions is structured on two levels of representation ("perceptual" and "conceptual") and that its internal structure is highly explored in emotion recognition tasks.

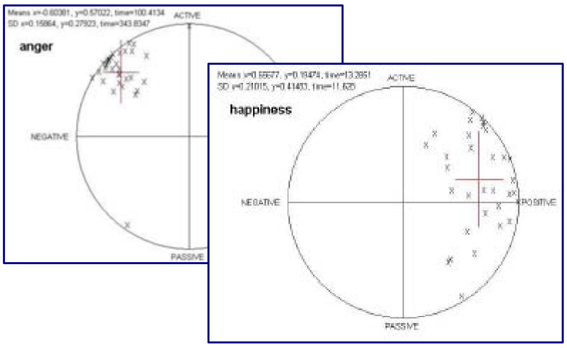

Fig 3. Evaluation of "anger" and "happiness" [Wan et al., 2005]

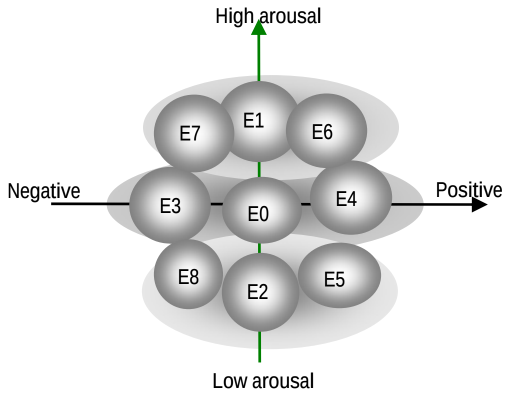

Fig 4. General categories on the "perceptual level"

In the next section we give a short overview of machine emotion recognition with regards of the used models. Section 3 proposes a processing architecture which relates the two levels and the last concludes the paper.

A Trial to Over-fly Machine Emotion Recognition

Extensive surveys in areas such as facial expression analysis (Fasel and Luettin 2003) vocal emotion (Oudeyer 2003), gesture recognition (Turk, 2001; Marcel 2002, Heylen (2006)) head tracking (Jaimes and Sebe 2005) have been published during the last decade. The problem of multimodal emotion recognition approaches are extensively discussed in very recent surveys (Sebe et al. 2005, Jaimes and Sebe 2007, Zeng et al. 2009). Our attempt is to make a parallel between the existing methods of machine emotion classification and a generalized scheme of the process of multimodal emotion recognition by humans.

Facial Expression Recognition. Studies, that are known in the domain (Ekman and Friesen 1978, Ekman 1989, Ekman 1992), have emphasized that facial expressions are universally expressed and recognized by humans. They focused on a set of seven emotions that have associated facial expressions (*fear*, *anger*, *sadness*, *happiness, disgust*, *surprise*, and *contempt*). The so-called Facial Action Coding System was developed to code facial expressions where movements on the face wore described by a set of Action Units. Ekman's work inspired many researchers to use image and video processing in order to classify emotions. Some methods follow feature-based approach (tracking of specific features such as the corners of the mouth, eyebrows, etc.) and other use a region-based approach (facial motions are measured in certain regions such as the eye/eyebrow and the mouth). Two types of classification schemes are used: dynamic and static. Dynamic classifiers (Hidden Markov Models HMM) use several video frames and perform classification by analyzing the temporal patterns of the regions analyzed or features extracted. Static classifiers classify each frame in a video to one of the facial expression categories based on the results of a particular video frame. These methods are similar in the general sense that they first extract some features from the images, then these features are fed into a classification system, and the outcome is one of the pre-selected emotional categories.

Following the analyses by Sebe (2005), the performance of the existing systems varies between 74% and 98% depending on the algorithm, the number of emotional categories etc. More details are given in the survey by Jaimes and Sebe (2007), where the authors summarize that the used approaches suffer from the following limitations: 1) they handle a small set of posed prototypical facial expressions of six basic emotions from portraits or nearly frontal views of faces; 2) they do not perform a context-dependent interpretation of shown facial behavior; 3) they do not analyze extracted facial information on different time scales and can't analyze mood and attitude (larger time scales). On-going works are trying to overcome some of these limitations. For example, Hu et al. 2008 generated multi-view images of facial expressions from 3D data and showed that non-frontal views give better results in the developed system for emotion recognition.

It has to be pointed out that the majority of the developed systems use a set of basic emotions. Interestingly, practical results for person independent emotion recognition have shown that the classification rates are significantly higher when considering more general categories such as "*positive*", "*negative*", "*neutral*" and "*surprise*" (Sebe et al. 2002). Recently, some authors have made some attempts to generalise the obtained results to more global categories. E.g, Zeng et al. 2006 have developed a classifier for realistic data which categorizes directly to positive, negative and non-emotional states, simply using labeled training data to these three general categories. That corresponds very much to the general perceptual categories, proposed in figure 4. To our knowledge, there are not theoretical works in the direction of relating facial expressions with emotion dimensions (properties) or with generalized categories.

Speech emotion recognition. The systems for speech emotion recognition use techniques for extraction of relevant characteristics from the raw speech signal. Acoustic correlates of basic emotional categories are investigated in terms of pitch, energy, temporal and spectral parameters, on different time-scales, etc… with the aim to extract emotion-relevant information (for recent advances - consult the site of EU-IST Network of Excellence HUMAINE http://emotion-research.net/). The data-driven approaches for recognition use supervised machine learning algorithms (neural networks, support vector machines, HMM etc.) that are trained on databases of emotional speech, containing labelled utterances with a previously chosen set of basic emotions. One main problem of speech emotion recognition is related to the "training corpus" dependency of the classifiers, as discussed in Slavova, Verhelst and Sahli 2008. Other problems are related to 1) the bad performance of the classifiers in noisy environments and to 2) their speaker dependency.

In general, the reported results (obtained in laboratory conditions, within the corpora used for training) the recognition accuracy of the machine classifiers is comparable with the human "acoustic" categorization capacities (around 66% for 6 basic emotions following Scherer et al. 2001). Several systems use in addition language semantics to improve the results. We provide a few examples: Muler et al. (2004) report that the fuse of acoustic and language information increased the recognition rates from 74% to 92%. Chuang and Wu (2004) applied a keyword spotting system in order to transform the speech signal into textual data and report that the used 500 keywords played a decisive role in the "outside" test, where the acoustic module could not perform satisfyingly. All these strategies require speech recognition and charge the systems with additional complexity.

To our knowledge, the existing systems for speech emotion recognition are based on the discrete model. However, there are results at finding a correspondence between the audio-signal and the emotion-dimensions model. Research in speech synthesis has shown that nearly all acoustic variables show substantial correlations with the emotion dimensions (see for example Schröder M. et al. 2001, Schröder M. 2004).

Recognition using biological signals. Contemporary studies showed that parameters from measurements as electrocardiograms (ECG), electromyography (EMG), electroencephalograms (EEG), respiration and skin conductivity, are highly correlated with the emotional states. The first trial (Fridlund and Izard 1983) to apply pattern recognition to classification of 4 basic emotions using EMG features attained rates of 38-51% accuracy. Since than, the results are considerably better. Takahashi (2004) applied support vector machines in a classifier using data from EEG, pulse, and skin conductance and obtained very promising recognition rates for three of 5 emotions (around 70% for *joy*, *anger*, and *fear*). Nasos et al. (2004) used wireless sensors for measuring temperature, heart rate and galvanic skin response and obtained promising results for *fear*, *sadness* and *anger* (and not so good for *amusement*, *surprise* and *frustration)*. Lisetti and Nasos (2004) used galvanic skin response, heart rate and temperature and showed that three different supervised learning algorithms can generalize from new collections of signals. Despite of the evidences (Kim 2004), showing that biological changes are highly correlated with the emotion dimensions, these systems are conceived on the bases of the discrete model, i.e., the first step is to construct data models (EEG, ECG, temperature etc.) that correspond to basic emotions. Yet, the use of emotion dimensions is possible and gives promising results. Nakasone et al. (2005) developed a system based on the valence-arousal model which realizes real-time recognition of emotions in a game between human user and humanoid agent, using data from EMG and skin conductance. The system, based on Bayesian network, discriminated well different *arousal levels*, indicated by Galvanic skin response, but had difficulties with the *negatively valenced* emotions, indicated by EMG.

Multimodal emotion recognition. The first attempt (Huang et al. 1998) to use an "audio-visual" feature vector increased the performance and some confusions, made by a single-modality classifier wore resolved. The results, obtained by all other realized audio-visual emotion recognition systems are in the same direction. We may cite the works of Chen (2000), Yoshitomi et al. (2000), De Silva and Ng (2000), Go et al. (2003), Schuller at al. (2004), Zeng et al, (2004), Song et al. (2004), Sebe et al. (2006), and Zeng et al. (2007) who investigated the effects of a combined detection of facial and vocal expressions of affective states using different types of techniques and classifiers (HMM, SVM, Bayesian networks etc). In brief, these systems achieve an accuracy of 72–85% when recognising basic emotions from clean audiovisual input (e.g., noise-free recordings, closely placed microphone, non-occluded portraits) from an actor speaking a single word and showing exaggerated facial displays of a basic emotion. In their extended survey, Jaimes and Sebe (2007) note that the developed audio-visual systems *have most of the drawbacks of the unimodal analyzers*.

An number of efforts are reported toward multimodal systems based on other data sources, for example using haptical interaction on a touch-screen, via the mouse etc. (see Schuller et al. (2002) Maat and Pantic (2006)).

Multimodal Problems. Important problems are related to data fusion. First, "architectural" problems, related to the stage on which the fusion should be realised. The survey of Sebe et al. (2005) reports that multi-sensory data are typically processed separately and only combined at the end. In fact, several works recommend early fusion, for example Chen (2000) argues that to realize human-like multimodal analysis, the input data should be processed in a joint feature space. Second - "processing problems" - the size of joint feature space, the different feature formats and timing. These problems are discussed in details in Sebe et al. (2005).

Other problems are related to the multimodal training data. They concern the labelling, the use of unlabeled data, spontaneous or posed collection etc. The survey by Zeng et al.(2009) points that the existing methods typically handle deliberately displayed, exaggerated expressions of prototypical emotions despite the fact that deliberate behaviour differs in visual appearance, audio profile, and timing from spontaneously occurring behaviour.

In conclusion, effective emotion recognition is likely to take place when different input devices [1] are used in combination. Nevertheless, the multimodal machine emotion recognition is still in its infancy. It is clear that human subjects perform the task much better, in real time, in real situation, in noisy environment, looking from different angles. They use additional sources of information such as context and semantics of the produced speech, as well as "knowledge" about their own mind when estimating other peoples' states (a phenomenon, known as mind-reading). Our presentiment is that the improvement of recognition systems necessitates more fundamental analyses in the direction of the used theoretical models of emotions, with regards of the inter-related levels of representation (fig. 1), phases of recognition, interactions between the modality-specific information flows etc.

Multimodal Architecture – Some General Comments

In the proposed approach emotion recognition by human subjects is considered as a process of mapping of obtained from external sources information to existing conceptual knowledge. The process concerns treatment by sensory systems. As known, a *sensory system* is a part of nervous system consisting of sensory receptors that receive stimuli from environment, neural pathways that conduct this information to brain and parts of brain that processes this information. For the modeling purposes, we propose a general scheme of this processing, given on fig. 5.

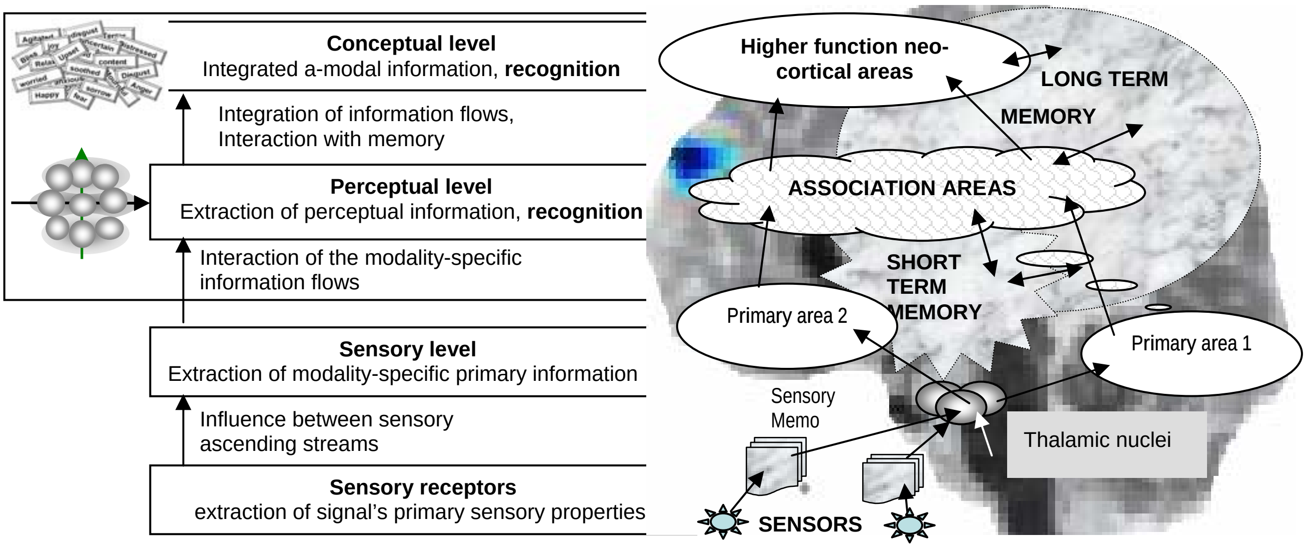

Fig. 5. General scheme of multi-modal processing

The sensors extract some primary features from the stimuli and keep short-time paths of the flow in the sensory memory. The information is conduced to different thalamic nuclei where the sensory flows influence each other. The flows are separately conveyed to the primary (modality specific) neo-cortical areas, which perform to obtain a novel, modality-specific representation. The resulting information flows are transmitted to separate association

[1] Several efforts are directed to machine emotion recognition based on **gestures**, using data from head and body movements, gaze etc. We have not included this direction in the proposed over-flight of the entire domain, we give only some references

(secondary) neo-cortical areas, where they interact to obtain novel, separate representations. On this stage are recognized meaningful perceptual features, in interaction with memory. That corresponds on the "perceptual level", given on fig.1. On the final stage, an a-modal conceptual representation is obtained, in result of integrating relevant information from of all the ascending flows and in interaction with memory.

Once the obtained, by bottom-up processing, information is identified to existing knowledge (on perceptual and conceptual level), a top-down process starts, which, briefly, propagates the features of the identified concepts on all the levels of the sensory hierarchy and the system "concentrates" on "expected" features.

None of the systems for machine emotion recognition is based on principles, similar to the above described scheme. Here we concentrate on the fact that human memory stores at time models of the emotion categories and of the emotion dimensions (properties). Our hypothesis is that two stages of recognition coexist, – at the "perceptual" and at the "conceptual" levels, which is important for the recognition strategy. On the left part of figure 5 we propose a general scheme of the processing. In summary, the information which can be compared to memory models is obtained at several steps. That includes 1. uni-modal feature extraction with memory-storage, 2. classification to general perceptual categories. 3. evaluation of the perceptual image using appropriate weights for the features of the input modalities, 4. identification of the perceptual category, 6. generation of a resulting feature-vector, 7. categorization on conceptual level using context, 8. Inducing of top-down process. Each step necessitates further clarification as well as modeling of interactions with memory.

Conclusion, On-going and Future Work

In conclusion, none of the existing systems for machine recognition uses multiple successive stages of fusing information in novel informative representations in order to recognize different levels of knowledge. We suppose that the recognition on the "perceptual" level is an important step which can not be omitted. Moreover, it seams that real-time, person independent, realistic data etc machine recognition, is more reliable when more global categories are used, as in the proposed here structure of the "perceptual level".

With the intention to realize multi-modal emotion recognition, we assume that a lot of efforts have to be performed for conceiving a system, working satisfyingly in realistic environment. When taking into account the provided the presented approach, we conceder that: The uni-modal recognition should be reviewed with concerns of the general categories of the "perceptual" level (the number of dimensions is also to be clarified). Several studies show that the information from audio and from video has different "weights" of recognition importance depending on the emotion displayed. On-going research tries to clarify these dependencies. Our belief is that, when combining the results and with an appropriate use of computational techniques, such general strategy will give more robustness of the system and will lead to a reduction of used feature-vectors. A special effort has to be done in discovering the relations between the two levels of representation. That includes also an appropriate modeling of the top-down influence in attempt to concentrate the computational resources on relevant information. Of course, all these questions cannot be solved without an appropriate model of memory and necessitates being tested and adjusted in implemented machine realizations.

Acknowledgements

The paper is partially financed by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria. www.ithea.org, www.foibg.com

Bibliography

Barsalou L. W., Simmons W. K., Barbey A. K. and Wilson C. D., 2003, Grounding conceptual knowledge in modality-specific systems, in: Trends in Cognitive Sciences, Volume 7, Issue 2, 2003

Chen L. and Huang T. 2000, Emotional expressions in audiovisual human computer interaction, in Proc. International Conference on Multimedia and Expo (ICME), 2000.

Chen L.S. 2000, Joint Processing of Audio–visual Information for the Recognition of Emotional Expressions in Human–computer Interaction, PhD thesis, Univ. of Illinois at Urbana-Champaign (2000).

Chuang Ze-Jing and Chung-Hsien Wu, 2004, Multi-Modal Emotion Recognition from Speech and Text, in: Computational Linguistics and Chinese Language Processing, Vol. 9, No. 2 , 2004

De Silva, L.C. and Ng P.C., 2000, Bimodal emotion recognition, in Proc. Face and Gesture Recognition Conf., 2000.

Ekman Paul and Friesen W. 1978, Facial Action Coding System: Investigator's Guide, Consulting Psychologists Press, 1978.

Ekman, Paul 1992. Are there Basic Emotions? Psychological Review, 99(3): 550-553.

Fasel B.,Luettin J. , 2003, Automatic facial expression analysis: a survey, Pattern Recognition 36 (2003)

Fridlund A. and Izard C., 1983, Electromyographic studies of facial expressions of emotions and patterns of emotion, in : Social Psychophysiology: A Sourcebook, J. Cacioppo and R. Petty, eds., 1983.

Heylen, D.K. (2006), Head gestures, gaze and the principle of conversational structure, International Journal of Humanoid Robotics, Vol: 3 Issue: 3, 2006

Hu Y, Zhihong Zeng, Yin L., Wei X., Tu J. and Huang T.S., 2008, A Study of Non-frontal-view Facial Expressions Recognition, in: proc of the 19th International conference on pattern recognition, 2008.

Huang TS, Chen LS, Tao H, Miyasato T, Nakatsu R , 1998 Bimodal emotion recognition by man and machine, in: proc. Workshop on Virtual Communication Environments, 1998

Jaimes A. and Sebe N. 2007, Multimodal human–computer interaction: A survey, in: Computer Vision and Image Understanding, 2007

Jaimes A. and Sebe N., 2005, Multimodal human computer interaction: a survey, IEEE International Workshop on HCI, 2005

Kim, Jonghwa, 2004, Sensing Physiological Information, Applied Computer Science, Workshop Santorini, HUMAINE WP4/SG3, 2004, online available on *http://emotion-research.net/*

Lisetti C.L. and Nasoz F. 2004, Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals, in Journal on Applied Signal Processing, Issue11, 2004

Maat L. and Pantic M., 2006 Gaze-X: adaptive affective multimodal interface for single-user office scenarios, in Proceedings of the 8th international conference on Multimodal interfaces, 2006

Marcel S., 2002, Gestures for multi-modal interfaces: a review, Technical Report IDIAP-RR 02-34, 2002.

Nakasone A., Prendinger H., Ishizuka M. 2005, Emotion Recognition from Electromyography and Skin Conductance, in proc of The Fifth International Workshop on Biosignal Interpretation, 2005 - Citeseer

Nasoz F., Alvarez K, Lisetti C. L. and Finkelstein N. Emotion recognition from physiological signals using wireless sensors for presence technologies; Cognition, Technology & Work; Volume 6, Number 1 / February, 2004

Oudeyer Pierre-Yves, 2003, The production and recognition of emotions in speech: features and algorithms International Journal of Human-Computer Studies 59, 2003

Scherer, Klaus R., Rainer Banse, and Harald G. Wallbott. 2001. Emotion Inferences from Vocal Expression Correlate across Languages and Cultures, Journal of Cross-Cultural Psychology 32/1, 2001

Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M. Gielen & S., 2001, Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis, In: Proceedings Eurospeech, 2001

Schröder, Marc, 2004, Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-extreme Emotions, Publisher: Springer Berlin / Heidelberg

Schuller B., Lang M., Rigoll G, 2002, Multimodal emotion recognition in audiovisual communication, proc. of IEEE International Conference on Multimedia, 2002

Sebe N., Cohen I. , Gevers T. , Huang T.S., 2006, Emotion Recognition Based on Joint Visual and Audio Cues, in: Proc. International Conference on Pattern Recognition (2006),

Sebe N., Cohen I., Gevers T., and Huang T., 2005, Multimodal Approaches for Emotion Recognition: A Survey, in: Proceedings of SPIE, 2005

Slavova, V., Verhelst, W., Sahli H. 2008, A cognitive science reasoning in recognition of emotions in audio-visual speech, in: International Journal of Information Theories and Applications, vol. 15 / 2008,

Takahashi, Kazuhiko 2004; Remarks on Emotion Recognition from Bio-Potential Signals; in proc of 2nd International Conference on Autonomous Robots and Agents, 2004 Palmerston North, New Zealand

Takarishi K. 2004, Comparison of Emotion Recognition Methods from Bio-Potential Signals, in Japanese Journal of Ergonomics, vol. 40 n2, 2004.

Turk M., 2001, Gesture recognition, Handbook of Virtual Environment Technology, K. Stanney (Ed.), 2001.

Wan V., Ordelman R., Moore J., Muller R., (2005) "AMI Deliverable Report, Describing emotions in meetings", internal project report, On line available http://www.amiproject.org/

Zeng Z,, Pantic, M., Roisman, G.I., Huang, T.S., 2009, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, in: Pattern Analysis and Machine Intelligence, 1, 2009

Zeng Z., J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, S. Levinson, 2004, Bimodal HCI-related affect recognition, ICMI, 2004.

Zeng Z., Jilin Tu, Ming Liu, Thomas S. Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson 2007, Audio-Visual Affect Recognition, IEEE Transactions on multimedia, V9, N2, 2007

Authors' Information

Velina Slavova - New Bulgarian University, Department of Computer Science; e mail: vslavova@nbu.bg

Hichem Sahli - Vrije Unversiteit Brussel, Department of Electronics & Informatics; e mail: hsahli@etro.vub.ac.be

Werner Verhelst - Vrije Universiteit Brussel, Department of Electronics & Informatics; wverhels@etro.vub.ac.be