MULTIMODAL EMOTION RECOGNITION AND EXPRESSIVITY ANALYSIS

Stefanos Kollias and Kostas Karpouzis

Department of Computer Science
School of Electrical and Computer Engineering
National Technical University of Athens
Politechnioupoli, Zografou 15773, Greece
e-mail: stefanos@cs.ntua.gr, kkarpou@image.ntua.gr

ABSTRACT

The paper presents the framework of a special session that aims at investigating the best possible techniques for multimodal emotion recognition and expressivity analysis in human computer interaction, based on a common psychological background. The session mainly deals with audio and visual emotion analysis, with physiological signal analysis serving as supplementary to these modalities. Specific topics that are examined include extraction of emotional features and signs from each modality in separate, integration of the outputs of single-mode emotion analysis systems and recognition of the user's emotional state, taking into account emotion models and existing knowledge or demands from both the analysis and synthesis perspective. Various labelling schemes, supply of accordingly labeled test databases, as well as synthesis of expressive avatars and affective interactions, are issues brought up and examined in the proposed framework.

1. INTRODUCTION

In everyday life people express their emotions through multiple modalities, such as their speech, their face and their body. This means that a system that attempts to interact with humans, taking into account their emotional state or attitude, needs to process, extract and analyse a variety of cues provided through humans' speech, facial expressions, hand and body pose. Conversely, all of the cues can be used to convey emotional messages to a user. Additional kinds of information not used in natural communication, but potentially relevant to interfaces, come from emotion-related somatic and cortical changes. The state-of-the-art in emotion research calls for a coherent treatment of all these issues. Work in this area is beginning to develop, and the proposed session brings together leading researchers in it so that the multimedia community can engage with the developments.

Multimodal emotional sign extraction and emotion recognition have been a major limiting factor in the development of emotion-oriented systems. Tackling these issues, the session focuses on extraction of emotional features and signs from each modality in separate, on combining the modality outputs and on recognition of the user's emotional state, taking into account the emotional psychological background and existing knowledge or demands from both the analysis and synthesis perspective. In this way, it foresees a unified framework for human computer interaction (HCI), where the machine is capable of both recognising its user's emotional state and generating expressive avatars that appropriately respond to the extracted user's emotional signs.

The different aspects of multimodal sign extraction involve speech, visual, physiological signal processing, signal segmentation, emotional feature extraction, single-mode classification and statistical analysis, multimodal synchronisation, recognition models, fuzzy reasoning, context analysis. However, the study of emotion is in fact a multidisciplinary endeavour, and consequently the above aspects cannot been considered independently from the emotional psychological theory and the emotional representations, as well as the knowledge about interrelations adopted and used for synthesizing the same signs in embodied computational agents (ECAs). Novel multimodal emotion analysis techniques of wide applicability need to take into account and provide solutions for different emotional models, e.g., discrete, continuous, temporal, for a variety of expressivity parameters, e.g., related to face, gestures, speech, as well as for synchronisation and integration of modalities.

0-7803-9332-5/05/$20.00 ©2005 IEEE

The session aims at investigating the best possible techniques and framework for emotion recognition, while considering expressivity analysis in ECA generation, associated knowledge about the context and dynamics of HCI, as well as theoretical labelling schemes and supply of test databases labelled according to such schemes. The major interest is on audio and visual emotion analysis, with physiological analysis serving as supplementary to external measurements. In sections 2-4, research focuses on each modality in separate, investigating a variety of issues/problems that appear, including processing, feature extraction, analysis of significance of parameters. Section 5 focuses on emotion recognition and associated models and techniques, integrating the information provided by single-mode analysis. Conclusions are given in section 6.

2. EMOTIONAL SPEECH ANALYSIS

Speech is a major channel for communicating emotion. Indeed, in some settings, such as telephone conversations, it is the only available channel. The speech signal conveys a large amount of information: textual, lexical, emotional and gestural information, as well as information related to the identity, age and sex of the speaker. Extracting the information related mainly to emotion is a daunting task that has been examined in many different contexts- different elicitation methods, different labelling schemes, different sizes of databases. In the single-mode form, two main problems are addressed; finding the set of features in the speech signal that are most significant in conveying emotions and finding the best classification algorithm that can indicate emotional expression, based on the above features.

Several key concepts related to the analysis of speech and the information it conveys are presented below; these have to do with the layers of information in the speech signal, and how this is analyzed with respect to emotional content [1-5]. The focus next is on paralinguistic data; detecting emotions in linguistic data can also be performed, using a variety of techniques, comprising Bayesian Networks, N-Grams, and Bunch-of-Words.

2.1. Paralinguistic speech analysis

This is the non-verbal information in the speech signal. Acoustically it is manifested in several aspects of the speech signal: Prosody, which is composed of intonation, duration, and intensity; and speech quality. All of these interact strongly with the verbal component: intensity, for example, is inherently stronger for certain phonemes than for others. Prosodic parameters are the main indicators for punctuation, though they also convey emotional information. Voice quality is influenced by a variety of physiological factors, though it also has a large part in conveying emotion.

2.2. Feature Extraction

The raw speech signal is a stream of samples. Extracting meaningful information from it involves the key phase of feature extraction; in effect, this involves transforming a data stream at a high rate, with a high level of redundancy, into several channels of more informative information at a lower rate. Feature extraction can take place on many levels of abstraction and "density": from phrase markers, words, syllables and phones, down to detailed measures of spectrum, energy and fundamental frequency.

An emotion analysis system should be able to extract and utilize all the information that a human listener has access to, consciously or not. This information includes :

- pitch (F0) contour, range, variance, mean, jitter,
- intensity, shimmer,
- voice quality,
- pauses – duration: pauses, speaking rate,
- Background information on the speaker – age, sex, social / cultural background, personality,
- information on the interaction: social interaction among friends, among strangers, call center, boss/employee conversation, interview.

2.3. Statistical methods/ classification algorithms

Targeting at characterization of emotional speech and implementation of automatic emotion detection requires statistical analysis or classification, based on the feature sets described above. Different methods have been examined in this context, such as Decision Trees (DTs), Hidden Markov Models (HMMs), Neural Networks (NNs), Support Vector Machines (SVMs).

3. VISUAL ANALYSIS AND ECA EXPRESSIVITY

There is a long history of interest in the problem of recognizing emotion from facial expressions, and extensive studies on face perception during the last twenty years. Facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face [10, 15-17]

3.1.Emotional Facial Analysis & the MPEG-4 standard

Most of the above models are based on the well known Facial Action Coding System (FACS). FACS is an anatomically oriented coding system, based on the definition of "action units" of a face that cause facial movements [12]. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard. In particular, the Facial Definition Parameter (FDP) set and the Facial Animation Parameter (FAP) set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, as well as the animation of faces reproducing expressions, emotions and speech pronunciation. Viseme definition has been also included in the standard for synchronizing movements of the mouth related to phonemes with facial animation [11].

Although FAPs provide all the necessary elements for MPEG-4 compatible animation, we cannot use them for the analysis of expressions from video scenes, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, a mapping has been defined between them and the movement of specific FDP feature points (FPs), which correspond to salient points on the human face.

The session's main interest is with the MPEG-4 standard and related feature extraction, since this represents a common framework for emotion analysis and synthesis that can be used in HCI applications.

3.2. Emotional Gesture Analysis

The detection and interpretation of hand gestures has become an important part of HCI in recent years [6-8, 14]. Sometimes, a simple hand action, such as placing a person's hands over his ears, can pass on the message that he has had enough of what he is hearing; this is conveyed more expressively than with any other spoken phrase. Interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and other parts of the human body, be measurable by the machine.

Hand tracking systems are used to extract emotion-related features through hand movement. The general process involves the creation of skin color areas which are tracked between subsequent frames. By tracking the centroid of those skin masks estimates of user's movements can be computed. Most hand tracking techniques utilize fused color and motion segmentation algorithms.

3.3. Targeting Emotion Recognition

A variety of techniques can be used to recognize the underlying emotional states, based on analysis of the FAP features extracted from the user's face, such as neural network classifiers, clustering techniques, SVMs and neurofuzzy networks. Of significant interest is usage of unsupervised hierarchical clustering, since this can form a basis for future merging of different emotional representations (i.e. different hierarchical levels), and categorization in either coarser or more detailed classes (half-plane, quadrants, components, discrete emotions).

Gestures can be utilized to support the outcome of the facial expression analysis subsystem, since in most cases they are too ambiguous to indicate solely a particular emotion. In a given context of interaction, gestures can be associated with a particular expression – e.g. *hand clapping* of high frequency expresses *joy*, *satisfaction* - while others can provide indications for the kind of the emotion expressed by the user. The position of the centroids of the head and the hands over time forms the feature vector sequence that can feed an HMM classifier whose outputs correspond to particular gesture classes.

3.4. Targeting Expressivity of ECAs

3.4.1. Facial Expressivity

A facial expression is characterized by its temporal parameters and its shape, that is the quantity of displacement for all the involved FAPs. Knowing the starting time and duration of an expression, the next step is to calculate the course of its intensity, i.e., the amplitude of time-varying facial movements that compose it. Based on the above-described targets of facial expressivity, features that would be desired to extract are

- FAP general quantity and quality of movement, related to emotional content.
- FAP interaction, i.e., how activation of one FAP is temporally related to activation of others.
- FAP transition, i.e., how FAPs are moved between two consecutive facial expressions.

Due to noise, illumination variations and low resolution capturing devices, the detection of facial features, and consequently, facial feature points, can be inaccurate. For this reason, facial analysis mechanisms should automatically evaluate the quality of each computed mask, assigning a confidence level to it. The emotion recognition system can take advantage of each feature's confidence level when analyzing them. Of particular importance is viseme analysis through time, since they can be the main feature to synchronize emotional facial and speech analysis for multimodal emotion recognition. Moreover, evaluation of FAP changes through time and detected expression in a variety

of examples can compute the facial expressivity parameters described above and form respective rules that can be useful for generation of expressive ECAs in HCI.

3.4.2. Gesture Expressivity

The basic unit of a gesture is the keyframe, inside which there can be the arm configuration in space, i.e., the rotation to apply to the shoulder and elbow to allow the arm to be set up in a certain position, the orientation of the palm. So, starting by tracing the spatial position of wrists trough time it should be possible to determine properties like speed, acceleration, direction variation that contribute to define the expressivity parameters.

4. PHYSIOLOGICAL SIGNAL ANALYSIS

An extensive body of research in psychophysiology has established that visceral signs of emotion show some statistical discrimination between emotional and non-emotional states, and to some extent among emotions [18-21]. Broadly speaking, the research indicates that a few variables carry most of the information about visceral differences between emotional states. They are Heart Rate, Skin Conductance Level, Number of Non-specific Skin Conductance Responses; Finger Temperature, Face Temperature, Muscle Activity, Movement, Systolic Blood Pressure, Diastolic Blood Pressure, Stroke Volume, Cardiac output, Finger Pulse Volume, Respiration, Pulse Transit Time, Blood Volume, Number of Muscle Tension Peaks, Respiration Rate, Hand Temperature, Inspiratory Index, Peripheral Resistance, Respiration Irregularity.

A necessary task is to specify which biosensor type can be deployed to supplement audio/visual measurement with physiological information, e.g. EMG for face recognition and RSP/SC for speech recognition. Some remarkable interactions between human emotion channels have already been observed; for example, the value of skin conductivity increases instantly when user is talking under neutral emotional state. Similarly, the respiration signal changes into an irregular wave form with lower amplitude. The respiration wave also shows particular abrupt changes corresponding to certain facial muscle-activity. EMG sensors on the jaws or forehead can improve accuracy in computation of FAP-variations in facial emotion analysis. They can also be positioned on the body to measure muscle contraction intensity for gesture recognition.

5. MULTIMODAL EMOTION RECOGNITION

5.1. Theoretical Models of Emotion

Emotion recognition includes first the selection of the appropriate emotion model, i.e. discrete, dimensional or appraisal. Issues such as 'which discrete emotions are more often expressed in specific contexts, e.g., educational, games, office', 'how many dimensions would be useful to discriminate as many as possible different emotions', or 'could the appraisal models be useful to refine the identification of emotion provided with the other models' need further research and investigation [9, 13].

5.2. Modality Integration

Multimodal recognition can use different theoretical models, i.e., direct identification (inputs being directly provided to the multimodal recognition system), separate identification (recognition of separate modalities), dominant modality perception, motor space integration (of all modalities). The key problems identified for integration of single-mode emotion analysis techniques are:

- To define the main integration processes and functions for recognition and expression of emotion in the visual, auditory and physiological modalities.
- To identify which modalities will be more relevant for different emotions in the recognition processes (for example, disgust is usually less well recognized in the auditory modality than in the visual modality).
- To define the problems of synchronization and temporal sequence in different modalities (e.g., the importance of temporal sequence in the auditory modality could be different than in visual modality).
- To identify which dimensions increase significantly the perception of realism of emotion at low cost.

5.3. Multimodal Emotional Databases

Although several multimodal databases have already been developed [13] the fusion of the annotations of individual heterogeneous modalities has been seldom tackled. Defining the integrative functions for the different modalities raise several issues some of which have been tackled in multimodal corpora without emotion, such as computing metrics measuring degrees and types of cooperation between modalities and managing intrinsic timing relationships between the modalities.

6. CONCLUSIONS

The special session concentrates on multimodal emotion recognition and expressivity analysis for HCI applications. It investigates models and techniques for extracting and analyzing emotional signs, mainly from speech and faces, but also from gestures and physiological data, using a common psychological background.

Aknowledgement: The presented framework is developed within (and supported by) the EC FP6 HUMAINE (Human-Machine Interaction Network on Emotions) Network of Excellence, 2004-2008 (www.emotion-research.net).

7. REFERENCES

[1] L. Devillers, I. Vasilescu: "Reliability of Lexical and Prosodic cues in two real-life spoken dialog corpora", LREC, Lisbonne, May 2004.

[2] L. Devillers, I. Vasilescu, L. Lamel: "Emotion Detection in a task-oriented Dialog Corpus", IEEE International Confererence on Multimedia, ICME, Baltimore, July 2003

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Noth, "Desperately Seeking Emotion or: Actors, Wizards and Human Beings", Proceedings of ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK, 2000.

[4] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, W. Machiel, S. Sybert, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark", Proceedings of ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK, 2000

[5] ERMIS : Emotion Recognition for Man Machine Interaction Systems, EC FP5 Project, 2001-2004, www.image.ntua.gr/ermis.

[6] A. Wexelblat, An approach to natural gesture in virtual environments, *ACM Transactions on Computer-Human Interaction*, vol. 2, iss. 3, pp. 179 – 200, 1995.

[7] F. Quek, "*Unencumbered gesture interaction*," *IEEE Multimedia*, vol. 3. no. 3, pp. 36-47, 1996.

[8] J. Lin, Y. Wu, and T.S. Huang, "Modeling human hand constraints," in *Proc. Workshop on Human Motion*, Dec. 2000, pp. 121-126.

[9] K. Scherer and P. Ekman, *Approaches to Emotion*, Lawrence Erlbaum Associates, 1984.

[10] M. Davis and H. College, *Recognition of Facial Expressions*, Arno Press, New York, 1975.

[11] A. M. Tekalp, J. Ostermann, "Face and 2-D Mesh Animation in MPEG-4", *Signal Processing: Image Communication, Vol. 15, pp. 387-421,* 2000.

[12] P. Ekman and W. Friesen, *The Facial Action Coding System*, Consulting Psychologists Press, San Francisco, CA, 1978.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, "Emotion Recognition in Human-Computer Interaction", *IEEE Signal Processing Magazine*, January 2001.

[14] Y. Wu and T.S. Huang, "Hand modeling, analysis, and recognition for vision-based human computer interaction", *IEEE Signal Processing Magazine*, vol. 18, iss. 3, pp. 51-60, May 2001.

[15] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4", EURASIP Journal on Applied Signal Processing, Vol. 2002, No. 10, pp. 1021-1038, Hindawi Publishing Corporation, October 2002.

[16] G. Votsis and S. Kollias, "A modular approach to facial feature segmentation on real sequences", *Signal Processing, Image Communication,* vol. 18, pp. 67-89, 2003.

[17] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey", Pattern Recognition, vol. 36, pp. 259-275, 2003.

[18] Davidson R.J., Jackson D.C., Kalin N.H. . "Emotion, plasticity, context and regulation: Perspectives from affective neuroscience", *Psychological Bulletin* 126, 890-909, 2000.

[19] ORESTEIA : Reports D1.5 and D2.5 of EC FET Dissapearing Computer Project Oresteia, www.image.ntua.gr/oresteia , 2003.

[20] Picard R. W., Vyzas E., & Healey J. . "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State", *IEEE Transactions on Patterns Analysis and MachineIntelligence,* 23, 1175-1191, 2001.

[21] Schwartz G. E., Weinberger D. A., & Singer J. A.,*NonlinearBiomedical Signal Processin*g, Vol. II, Dynamic Analysis and Modelling, IEEE Press, New York, ch. 6, pp. 159- 213, 2001.