Facial and Body Feature Extraction for Emotionally-Rich HCI

Kostas Karpouzis, Athanasios Drosopoulos, Spiros Ioannou, Amaryllis Raouzaiou,
Nicolas Tsapatsoulis and Stefanos Kollias
Image, Video and Multimedia Systems Laboratory
National Technical University of Athens, Greece

Introduction

Emotionally-aware Man-Machine Interaction (MMI) systems are presently at the forefront of interest of the computer vision and artificial intelligence communities, since they give the opportunity to less technology-aware people to use computers more efficiently, overcoming fears and preconceptions. Most emotion-related facial and body gestures are considered to be universal, in the sense that they are recognized along different cultures; therefore, the introduction of an "emotional dictionary" that includes descriptions and perceived meanings of facial expressions and body gestures, so as to help infer the likely emotional state of a specific user, can enhance the affective nature (Picard, 2000) of MMI applications.

As a general rule, our intuition of what a human expression represents is based on trying to mimic the way the human mind works while making an effort to recognize such an emotion. This means that even though image or video input are necessary to this task, this process cannot come to robust results without taking into account features like hand gestures or body pose. These features are able to convey messages in a much more expressive and definite manner than mere wording, which can be misleading or ambiguous. Sometimes, a simple hand action, such as placing ones' hands over their ears, can pass on the message that they've had enough of what they are hearing more expressively than any spoken phrase.

Background

Emotion Representation

Most emotion analysis applications attempt to annotate video information with category labels that relate to emotional states. However, since humans use an overwhelming number of labels to describe emotion, we need to incorporate a higher-level and continuous representation that is closer to our conception of how emotions are expressed and perceived.

Activation-emotion space (Cowie, 2001) is a simple representation that is capable of capturing a wide range of significant issues in emotion. It rests on a simplified treatment of two key themes:

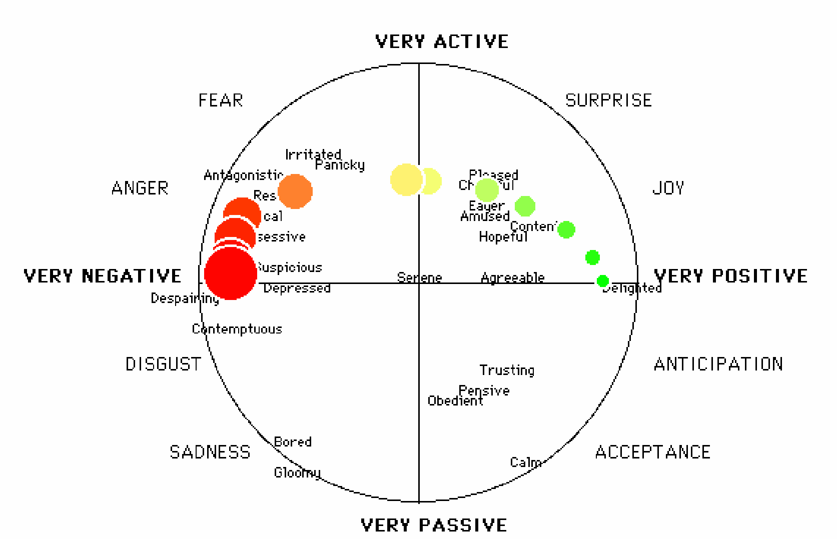
- Valence: The clearest common element of emotional states is that the person is influenced by feelings that are "valenced", i.e. they are centrally concerned with positive or negative evaluations of people or things or events.

- Activation level: Research has recognized that emotional states involve dispositions to act in certain ways. Thus, states can be rated in terms of the associated activation level, i.e. the strength of the person's disposition to take some action rather than none.

The axes of the activation-evaluation space reflect those themes, with the vertical axis showing activation level, while the horizontal axis represents evaluation. This scheme of describing emotional states is more tractable than using words and can still be translated into and out of verbal descriptions. Translation is possible because emotion-related words can be thought of as positions in activation-emotion space.

A surprising amount of emotional discourse can be captured in terms of activation-emotion space. Perceived full-blown emotions are not evenly distributed in activation-emotion space; instead they tend to form a roughly circular pattern. In this framework, the center can be thought of as a natural origin, thus making emotional strength at a given point in activation-evaluation space proportional to the distance from the origin. The concept of a full-blown emotion can then be translated roughly as a state where emotional strength has passed a certain limit. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation. A related extension is to think of primary or basic emotions as cardinal points on the periphery of an emotion circle. Plutchik (1980) has offered a useful formulation of that idea, the "emotion wheel" (see Figure 1).

Figure 1: The Activation-emotion space



Facial Expression Analysis

There is a long history of interest in the problem of recognizing emotion from facial expressions (Ekman, 1978), and extensive studies on face perception during the last twenty years (Ekman, 1973, Davis, 1975, Scherer, 1984). The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others. In most cases, these studies attempt to define the facial expression of emotion in terms of qualitative patterns capable of being displayed in a still image. This usually captures the apex of the expression, i.e. the instant at which the indicators of emotion are most noticeable. More recently, emphasis has switched towards descriptions that emphasize gestures, i.e. significant movements of facial features.

In the context of faces, the task has almost always been to classify examples of the six emotions, considered to be universal, i.e. joy, sadness, anger, fear, disgust and surprise (Ekman, 1978). More recently, morphing techniques have been used to probe states that are intermediate between archetypal expressions. They reveal effects that are consistent with a degree of categorical structure in the domain of facial expression, but they are not particularly large, and there may be alternative ways of explaining them – notably by considering how category terms and facial parameters map onto activation-evaluation space (Karpouzis, 2000).

Analysis of the emotional expression of a human face requires a number of pre-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. Facial features can be viewed (Ekman, 1975) as either static (such as skin color), or slowly varying (such as permanent wrinkles), or rapidly varying (such as raising the eyebrows) with respect to time evolution. Detection of the position and shape of the mouth, eyes, particularly eyelids, wrinkles and extraction of features related to them are the targets of techniques applied to still images of humans. However, in Bassili's (1979) experiments expressions were recognized at above chance levels when based on image sequences, whereas only happiness and sadness were recognized at above chance levels when based on still images. Techniques which attempt to identify facial gestures for emotional expression characterization face the problems of locating or extracting the facial regions or features, computing the spatio-temporal motion of the face through optical flow estimation, and introducing geometric or physical muscle models describing the facial structure or gestures.

Body Gesture Analysis

The detection and interpretation of hand gestures has become an important part of human computer interaction in recent years (Wu, 2001). To benefit from the use of gestures in MMI it is necessary to provide the means by which they can be interpreted by computers. The MMI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and even other parts of the human body, be measurable by the machine. First attempts to address this problem resulted in mechanical devices that directly measure hand and/or arm joint angles and spatial position. The so-called glove-based devices best represent this solutions' group.

Since the processing of visual information provides strong cues in order to infer the states of a moving object through time, vision-based techniques provide at least adequate, alternatives to capture and interpret human hand motion. At the same time, applications can benefit from the fact that vision systems can be very cost efficient and do not affect the natural interaction with the user. These facts serve as the motivating forces for research in the modeling, analysis, animation, and recognition of hand gestures. Analyzing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies.

The first phase of the recognition task is choosing a model of the gesture. The mathematical model may consider both the spatial and temporal characteristic of the hand and hand gestures. The approach used for modeling plays a pivotal role in the nature and performance of gesture interpretation. Once the model is decided upon, an analysis stage is used to compute the model parameters from the image features that are extracted from single or multiple video input streams. These parameters constitute some description of the hand pose or trajectory and depend on the modeling approach used. Among the important problems involved in the analysis are those of hand localization, hand tracking, and selection of suitable image features. The computation of model parameters is followed by gesture recognition. Here, the parameters are classified and interpreted in the light of the accepted model and perhaps the rules imposed by some grammar. The grammar could reflect not only the internal syntax of gestural commands but also the possibility of interaction of gestures with other communication modes like speech, gaze, or facial expressions. Evaluation of a particular gesture recognition approach encompasses accuracy, robustness, and speed, as well as the variability in the number of different classes of hand/arm movements it covers.

Gesture interpretation

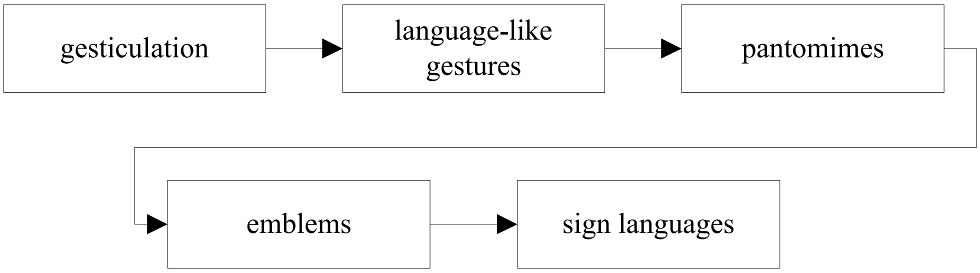
Gesture analysis research follows two different approaches that work in parallel. The first approach treats a hand gesture as a two- or three dimensional signal that is communicated via hand movement from the part of the user; as a result, the whole analysis process merely tries to locate and track that movement, so as to recreate it on an avatar or translate it to specific, predefined input interface, e.g. raising hands to draw attention or indicate presence in a virtual classroom.

The low level results of the approach can be extended, taking into account that hand gestures are a powerful expressive means. The expected result is to understand gestural interaction as a higher-level feature and encapsulate it into an original modal, complementing speech and image analysis in an affective MMI system (Wexelblat, 1995). This transformation of a gesture from a time-varying signal into a symbolic level helps overcome problems such as the proliferation of available gesture representations or failure to notice common features in them. In general, one can classify hand movements with respect to their function as (Cadoz, 1994):

- Semiotic: these gestures are used to communicate meaningful information or indications
- Ergotic: manipulative gestures that are usually associated with a particular instrument or job and
- Epistemic: again related to specific objects, but also to the reception of tactile feedback.

Semiotic hand gestures are considered to be connected, or even complementary, to speech in order to convey a concept or emotion. Especially two major subcategories, namely deictic gestures and beats, i.e. gestures that consist of two discrete phases, are usually semantically related to the spoken content and used to emphasize or clarify it (McNeill, 1992). This relation provides a positioning of gestures along a continuous space. This space is shown in Figure 2 below:

Figure 2: Gesture continuum from (Cadoz, 1994)



Main Thrust of the Chapter

Facial Feature Extraction

Robust and accurate facial analysis and feature extraction has always been a complex problem that has been dealt with by posing presumptions or restrictions with respect to facial rotation and orientation, occlusion, lighting conditions and scaling. These restrictions are being eventually revoked in the literature, since authors deal more and more with realistic environments, while keeping in mind pioneering works in the field. A hierarchical, robust scheme, coping with large variations in the appearance of diverse subjects, as well as of the same subject in various instances within real video sequences (see Figure 3), has been proposed by Votsis (2003). Soft a priori assumptions are made on the pose of the face or the general location of the features in it. Gradual revelation of information concerning the face is supported under the scope of optimization in each step of the hierarchical scheme, producing a posteriori knowledge about it and leading to a step-by-step visualization of the features in search.

Face detection is performed first through detection of skin segments or blobs, merging of them based on the probability of their belonging to a facial area, and identification of the most salient skin color blob or segment. Following this, primary facial features, such as eyes, mouth and nose, are dealt as major discontinuities on the segmented, arbitrarily rotated face (Figure 4). In the first step of the method, the system performs an optimized segmentation procedure. The initial estimates of the segments, also called seeds, are approximated through min-max analysis and refined through the maximization of a conditional likelihood function. Enhancement is needed so that closed objects will occur and part of the artifacts will be removed. Seed growing is achieved through expansion, utilizing chromatic and value information of the input image. The enhanced seeds form an object set, which reveals the in-plane facial rotation through the use of active contours applied on all objects of the set, which is restricted to a finer set, where the features and feature points are finally labeled according to an error minimization criterion (Figure 5).

Figure 3: The original frame from an expressive sequence

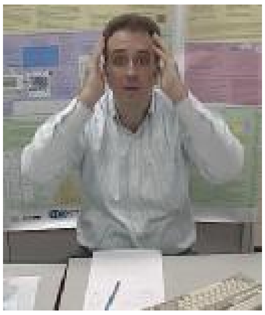


Figure 4: Detected primary facial features

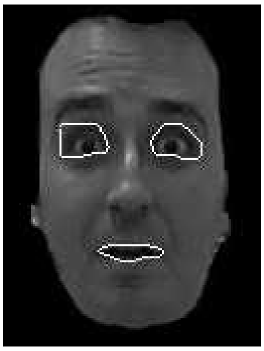
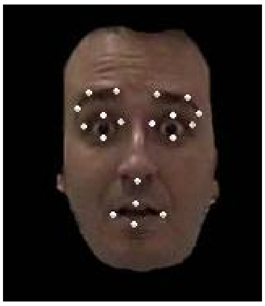


Figure 5: Detected facial features in the apex of an expression



Gesture tracking and recognition

The modeling of hand gestures depends primarily on the intended application within the MMI context. For a given application, a very coarse and simple model may be sufficient. However, if the purpose is a natural-like interaction, a model has to be established that allows many, if not all, natural gestures to be interpreted by the computer.

In general, human hand motion consists of the global hand motion and local finger motion. Hand motion capturing deals with finding the global and local motion of hand movements. Two types of cues are often used in the localization process: color cues (Kjeldsen, 1996) (see Figure 6), and motion cues (Freeman, 1995) (see Figure 7). Alternatively the fusion of color, motion and other visual or non-visual cues like speech or gaze is used (Sharma, 1996, Karpouzis, 2004).

To capture articulate hand motion in full degree of freedom, both global hand motion and local finger motion should be determined from video sequences. Different methods have been taken to approach this problem. One possible method is the appearance-based approaches, in which 2-D deformable hand shape templates are used to track a moving hand in 2-D (Darrell, 1996). Another possible way is the 3-D model-based approach, which takes the advantages of a priori knowledge built in the 3-D models.

In certain applications, continuous gesture recognition is required; as a result, the temporal aspect of gestures must be investigated. Some temporal gestures are specific or simple and could be captured by low detail dynamic models. However, many high detail activities have to be represented by more complex gesture semantics, so modeling the low-level dynamics is insufficient. The HMM (Hidden Markov Model) technique (Bregler, 1997) and its variations (Darrell, 1996) are often employed in modeling, learning, and recognition of temporal signals. Because many temporal gestures involve motion trajectories and hand postures (Figure 8), they are more complex than speech signals. Finding a suitable approach to model hand gestures is still an open research problem. Practical large-vocabulary gesture recognition systems by HMM are yet to be developed.

Figure 6: Skin color probability

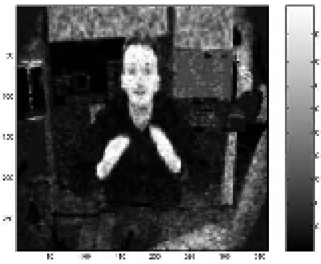


Figure 7: Detected moving hand segments

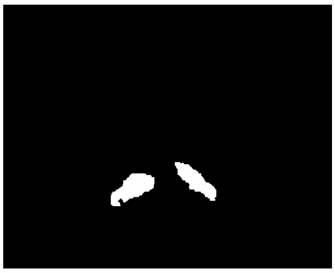
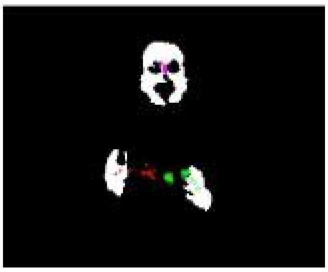


Figure 8: Hand tracking in a "clapping" sequence



Future Trends

General gesture analysis studies consider gestures to be spontaneous, free form movements of the hands during speech (gesticulation), while others, termed emblems, are indicative of a specific emotion or action, such as an insult. An interesting conclusion in (McNeill, 1992) is that the alternative use of gestures and speech in order to comprehend the communicated emotion or idea makes the whole concept of body language obsolete. Indeed, the study shows that instead of being "mere embellishments" of spoken content, gestures possess a number of para-linguistic properties. For example, such gestures convey a specific meaning only when considered as a whole, not as mere collections of low level hand movements. While spoken words are usually unambiguous and can be semantically interpreted only when in a complete sentence or paragraph, gestures are atomic when it comes to conveying an idea and typically their actual form depends on the personality and current emotional state of a specific speaker. As a result, gestures cannot be analyzed with the same tools used to process the other modals of human discourse. In the case of gesticulation, we can regard gestures as functions of hand movement over time; the result of this approach is that the quantitative values of this representation, such as speed, direction or repetition, can be associated to emotion-related values, such as activation. This essentially means that in many cases we do not need to recognize specific gestures to deduce information about the users' emotional state, but merely track the movement of their arms through time. This concept can also help us distinguish a specific gesture from a collection of similar hand movements: for example, the "raise hand" gesture in a classroom or discussion and the "go away" or "I've had enough" gestures are similar when it comes to hand movement, since in both cases the hand is raised vertically. The only way to differentiate them is to compare the speed of the upward movement in both cases: in the latter case the hand is raised in a much more abrupt manner. In our approach, such feedback is invaluable, since we try to analyze the users' emotional state by taking into account a combination of both gesture- and face-related features and not decide based on merely one of the two modals.

Conclusion

In this article we presented a holistic approach to emotion modeling and analysis. Beginning from a symbolic representation of human emotions, based on their expression via facial expressions and hand gestures, we described approaches to extracting quantitative and qualitative feature information from video sequences. While these features can be used for simple representation purposes, e.g. animation or task-based interfacing, this methodology is closer to the target of affective computing and can prove useful in providing feedback on the users' emotional state. Possible applications include human-like agents, that assist everyday chores and react to user emotions or sensitive artificial listeners that introduce conversation topics and react themselves to specific user cues.

References

Bassili, J.N. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology. 37. 2049-2059.

Bregler, C. (1997). Learning and recognition human dynamics in video sequences. Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition. 568-574.

Cadoz, C. (1994). Les realites virtuelles, Dominos, Flammarion.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J., (2001), Emotion Recognition in Human-Computer Interaction, IEEE Signal Processing Magazine, vol. 1. 32-80.

Darrell, T., & Pentland, A. (1996). Active gesture recognition using partially observable Markov decision processes. Proceedings of the IEEE International Conference in Pattern Recognition. vol. 3. 984-988.

Darrell, T., Essa, I., & Pentland, A. (1996). Task-Specific Gesture Analysis in Real-Time Using Interpolated Views. IEEE Transactions in Pattern Analysis and Machine Intelligence. 18(12). 1236-1242.

Davis, M., & College, H. (1975). Recognition of Facial Expressions. Arno Press.

Ekman, P., & Friesen, W. (1975). Unmasking the Face, Prentice-Hall.

Ekman, P., & Friesen, W. (1978). The Facial Action Coding System. Consulting Psychologists Press.

Ekman, P. (1973). Darwin and Facial Expressions. Academic Press.

Freeman, W.T., & Weissman, C.D. (1995). Television Control by Hand Gestures. Proceedings of the International Workshop on Automatic Face and Gesture Recognition. 179-183.

Karpouzis, K., Tsapatsoulis, N., & Kollias, S. (2000). Moving to Continuous Facial Expression Space using the MPEG-4 Facial Definition Parameter (FDP) Set. Proceedings of SPIE Electronic Imaging 2000.

Karpouzis, K., Raouzaiou, A., Drosopoulos, A., Ioannou, S., Balomenos, T., Tsapatsoulis N., & Kollias, S. (2004) Facial expression and gesture analysis for emotionally-rich man-machine interaction. In N. Sarris, M. Strintzis, (eds.), 3D Modeling and Animation: Synthesis and Analysis Techniques. Idea Group Publishers.

McNeill, D. (1992). Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago, USA.

Picard, R.W. (2000). Affective Computing. MIT Press. Cambridge.

Plutchik, R. (1980). Emotion: A psychoevolutionary synthesis. Harper and Row, NY, USA.

Scherer, K., & Ekman, P. (1984). Approaches to Emotion. Lawrence Erlbaum Associates.

Sharma, R., Huang, T.S., & Pavlovic, V.I. (1996). A Multimodal Framework for Interacting With Virtual Environments. C.A. Ntuen and E.H. Park, eds. Human Interaction With Complex Systems. Kluwer Academic Publishers. 53-71.

Votsis, G., Drosopoulos, A., & Kollias, S. (2003). A modular approach to facial feature segmentation on real sequences, Signal Processing: Image Communication. Vol. 18. 67-89.

Wexelblat, A. (1995). An approach to natural gesture in virtual environments. ACM Transactions on Computer-Human Interaction. 2(3). 179 – 200.

Wu, Y., & Huang, T.S., (2001). Hand modeling, analysis, and recognition for vision-based human computer interaction. IEEE Signal Processing Magazine. 18(3). 51-60.

Terms and Definitions

Affective computing: a recent theory that recognizes that emotions play an essential role in perception and learning, by shaping the mechanisms of rational thinking. In order to enhance the process of interaction, we should design systems with the ability to recognize, understand, even to have and express emotions.

Activation-emotion space: a 2-D representation of the emotion space, with the two axes representing the magnitude and the hue of a specific emotion.

Universal emotions: mainly after the influence of Ekman, these six emotions are considered to be universal, in the sense that they are uniformly recognized across different cultures.

Skin color estimation: a breakthrough in face detection and segmentation was the representation of skin color with a Gaussian model in a subset of the CrCb space, irrespectively of the actual skin color of the subject.

Model-based gesture analysis: In this framework, gestures are modeled as, usually finite, states of hand shape or position. This approach captures both the spatial and the temporal nature of the gestures, which is essential for analysis and recognition purposes.

Hidden Markov Model: Statistical models of sequential data utilized in many machine learning applications, e.g. speech and gesture recognition.

Semiotic hand gestures: Gestures used to communicate meaningful information or serving as indications; such gestures convey specific emotions in a more expressive manner, than vague hand movement.