Emotionally-rich Man-machine Interaction Based on Gesture Analysis

A. Drosopoulos, T. Mpalomenos, S. Ioannou, K. Karpouzis and S. Kollias

Image, Video and Multimedia Systems Laboratory,
National Technical University of Athens
9, Heroon Politechniou st., 15780 Zographou, Athens, Greece
{ndroso, tmpal, sivann, kkarpou}@image.ece.ntua.gr, stefanos@cs.ntua.gr

Abstract

Current Man-Machine Interaction (MMI) systems are capable of offering advanced and intuitive means of receiving input and communicating output and other messages to their users. Such interfaces give the opportunity to less technology-aware or handicapped people to use computers more efficiently and thus overcome related fears and preconceptions. In this context, most emotion- and expression-related 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 body gestures can enhance the multilinguality of MMI applications, without the need of text or speech translation. In order to extract emotion-related features through hand movement, we implemented a hand-tracking system. Emphasis was on implementing a near real-time, yet robust enough system for our purposes.

1 Introduction

Despite the progress in related research, our intuition of what a human expression or emotion actually represents is still based on trying to mimic the way the human mind works while making an effort to recognize such an emotion. While a lot of effort has been invested in examining individually different aspects of human expression, recent research has shown that this task can benefit from taking into account multimodal information. Facial and hand gestures as well as body pose constitute a powerful way of non-verbal human communication. Analysing such gestures is a complex task involving motion modelling and analysis, pattern recognition, machine learning, as well as psycholinguistic studies. Besides feeding an emotion analysis system, gestures and pose can also assist multi-user environments, where communication is traditionally text-based. Thus, simple chat or e-commerce applications can be transformed into powerful virtual meeting rooms, where different users interact, with or without the presence of avatars that take part in this process taking into account the perceived expressions of the users.

To achieve natural interactivity, it is required to track and interpret the behaviour of people that interact with a virtual environment, analysing the implicit messages that people convey through their facial expressions, gestures and emotionally coloured speech. As a consequence, it will be possible to generate avatars that follow the behaviour, i.e., the profile, emotional state, actions, choices and reactions of people in a virtual environment, thus modelling their presence in it, in a natural way.

The proposed paper extends work by the authors in the framework of emotion analysis based on facial cues to the context of combined hand and facial gesture analysis for emotionally enriched human-machine interaction; the proposed technologies are tested in real HCI problems, involving interaction of users with their own PC workstation.

Our approach includes detection and tracking of the user’s face, using face position differences from successive frames and pose estimation algorithms. Hand/body gesture extraction is then used for improving the emotion recognition process. Several types of gestures can be recognized: straight-line and turning-in-place motion gestures will be used as an indication of the users' interest in a specific item and its inspection. Also, tactile gestures, such as nudges, may again be viewed as an indication of user's interest, but in this time express the inability or indifference of the users to pick up and inspect the specific item.

2 Gesture Modeling, Analysis and Recognition

The scope of a gestural interface for MMI is directly related to the proper modeling of hand gestures. 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. According to different application scenarios, hand gestures can be classified into several categories: conversational gestures, controlling gestures, manipulative gestures and communicative gestures. Sign language is an important case of communicative gestures (Starner, Weaver & Pentland, 1997).

Hand localization is locating hand regions in image sequences. Skin color offers an effective and efficient way to fulfill this goal. According to the representation of color distribution in certain color spaces, current techniques of skin detection can be classified into two general approaches: nonparametric (Kjeldsen & Kender, 1996) and parametric (Wren, Azarbayejani, Darrel & Pentland, 1997). Hand image features can be geometric features such as points, lines, contours, and silhouettes. (Kuch & Huang, 1995). To capture articulate hand motion in full degree of freedom, both global hand motion and local finger motion should be determined from video sequences. 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, Essa & Pentland, 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 (Kuch & Huang, 1995).

Both temporal hand movements and static hand postures could represent meaningful gestures. Hand postures express certain concepts through hand configurations, while temporal hand gestures represent certain actions by hand movements. Sometimes, hand postures act as special transition states in temporal gestures and supply a cue to segment and recognize temporal hand gestures. Although hand gestures are complicated to model because the meanings of hand gestures depend on people and cultures, a set of specific hand gesture vocabulary can always be predefined in many applications, such as virtual environment (VE) applications, so that the ambiguity can be limited. Different from sign languages, the gesture vocabulary in VE applications is structured and disambiguated. Some simple controlling, commanding, and manipulative gestures are defined to fulfill natural interaction such as pointing, navigating, etc. These gesture commands can be simple in the sense of motion; however, many different hand postures are used to differentiate and switch among the commanding modes. For example, only if we know a gesture is a pointing gesture would it make sense to estimate its pointing direction. View-independent hand posture recognition is also a natural requirement in many VE applications.

3 Hand detection and tracking for MMI

In this work, certain gestures are considered 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. 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.

In order to extract emotion-related features through hand movement, we implemented a hand-tracking system. Emphasis was on implementing a near real-time, yet robust enough system for our purposes. The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames. By tracking the centroid of those skin masks we produce an estimate of the user's movements.

Most hand tracking techniques utilize fused color and motion segmentation algorithms. A region-growing color segmentation step such as RSST or a morphological partitioning tool such as watershed can be used to extract color areas, and combine them with optical flow information. Unfortunately those methods are slow even when combined with multiresolution techniques. In order to implement a computationally lighter system, our architecture (see Figure 1) takes into account a-priori knowledge related to the expected characteristics of the input image. Since the context is MMI applications, we expect to locate the head in the middle area of upper half of the frame and the hand segments near the respective lower corners. In addition to this, we concentrate on the motion of hand segments, given that they are the end effectors of the hand and arm chain and thus the most expressive object in tactile operations.

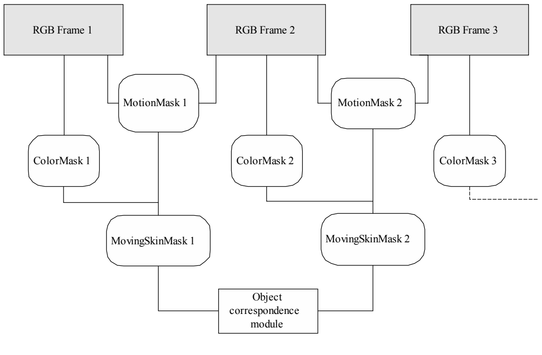
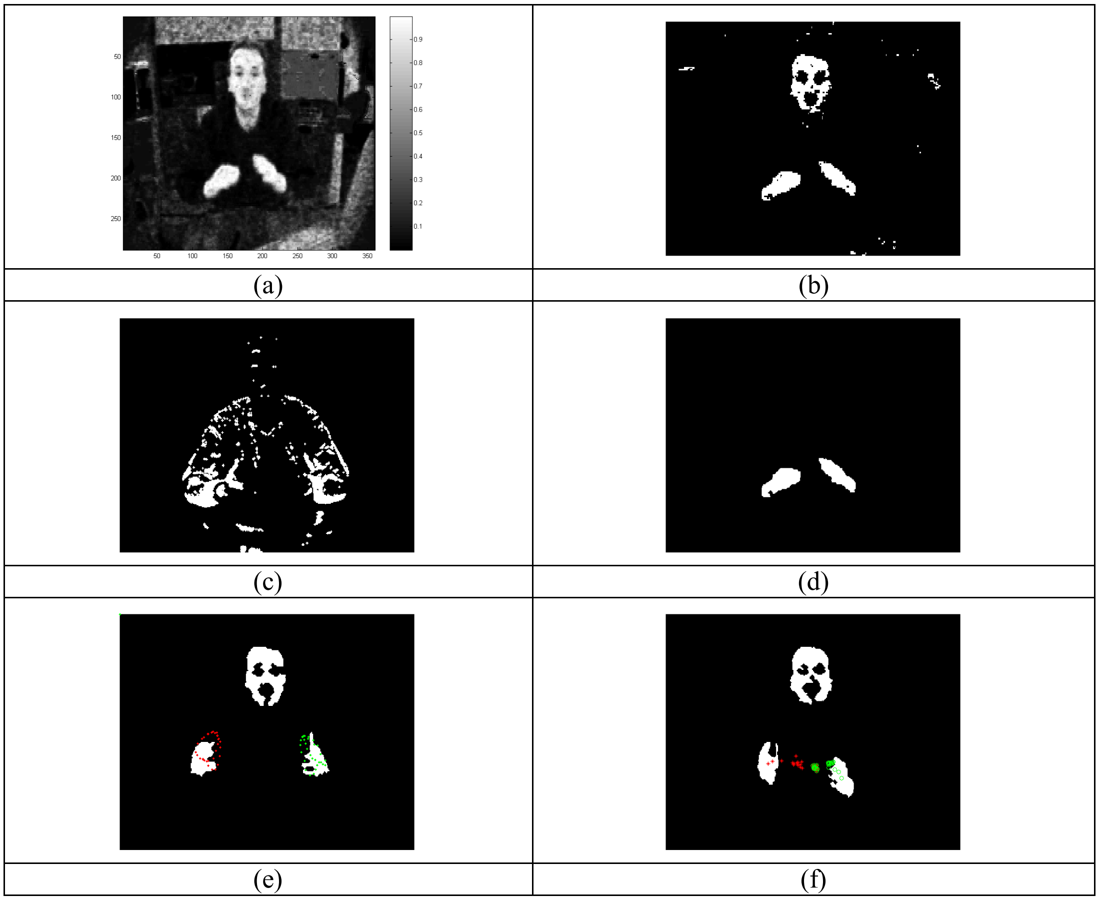


Figure 1: Abstract architecture of the hand tracking module

For each given frame, a skin color probability matrix is computed (Figure 2(a)). Possible moving areas are found by using the difference pixels between the current frame and the next, resulting to a possible-motion mask is created. This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color (Figure 2(b)) and motion (Figure 2(c)) masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to first unify objects and then remove objects with area smaller than 5% of the input image. A moving skin mask (msm) is then created by fusing the processed skin and motion masks (sm, mm), through the morphological reconstruction of the color mask using the motion mask as marker. The result of this process, after excluding the head object is shown in Figure 2(d). This moving skin mask consists of many large connected areas. For the next frame a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is based on object distance and object area (Figure 2(e)). In this figure, red markers represent the position of the centroid of the detected right hand of the user, while green markers correspond to the left one. In the case of hand object merging and splitting, e.g. in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand (Figure 2(f)).



(a) (b)

(c) (d)

(e) (f)

Figure 2: (a) Skin color probability for the input image; (b) Initial color mask created with skin detection; (c) Initial motion mask after pixel difference; (d) Detected moving hand segments after morphological reconstruction; (e) Results of hand tracking in an "Italianate" gesture; (f) Hand tracking in a "clapping" sequence

Following object matching in the subsequent moving skin masks, the mask flow is computed, i.e. a vector for each frame depicting the motion direction and magnitude of the frame's objects. This vector is used to extract information about each hand's motion speed, direction (horizontal, vertical, rotational, random, etc.) and repetitiveness. The corresponding mask flow magnitude over time for the clapping sequence is shown in Figure 3. The form of this figure shows that we can use the quantitative features of the gesture to track repeated hand movement. The described algorithm is relatively lightweight, allowing a rate of several fps on a usual PC.

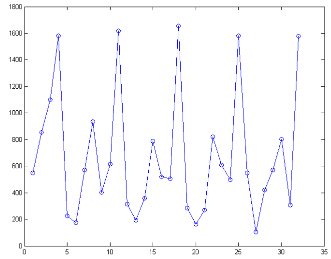


Figure 3: Object flow in the "clapping sequence"

4 Conclusions

Hand gestures and body pose provide another powerful means of communication. 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. We present a work that benefits from hand gestures in order to aid MMI in multi-user environments, where communication is traditionally reduced to text or text-to-speech voice.

References

Bregler, C. Learning and recognition human dynamics in video sequences. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 568-574.

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

Kjeldsen, R., & Kender, J. (1996). Finding skin in color images. *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition*, 312-317.

Kuch, J. J., & Huang, T. S. (1995). Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration. *Proc. IEEE Int. Conf. Computer Vision*, Cambridge, MA, 666-671.

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

Starner, T., Weaver, J., & Pentland, A. (1997). A wearable computer based American sign language recognizer. *Proc. IEEE Int. Symp. Wearable Computing*, 130-137.

Wren, C., Azarbayejani, A., Darrel, T., & Pentland, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Machine Intell.*, 9 (7), 780-785.