IJDCST @December Issue- V-1, I-8, SW-19
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Gesture Recognition Techniques in Natural Expressions on Mobile Devices by Considering OpenCV

M Ashok Kumar[1], Dr.R.R.Tewari[2],
[1]Department of Information Technology,V R Siddhartha Engineering College,Kanuru, Vijayawada, India,
Ashokkumar.munnangi@gmail.com
[2]Professor in Department of Computer Science, J.K.Institute of Applied Physics and Technology,
University of Allahabad-211002

Abstract— The proposed approach is intended to provide input for the analysis of hand gestures and facial expressions that humans utilize while engaged in various conversational states with robots that operate autonomously in public places. It has been integrated into a system which runs in real time on a conventional personal computer which is located on a mobile robot. Experimental results confirm its effectiveness for the specific task at hand. We show experimentally that we can successfully detect face occlusions with an accuracy of 83%. We also demonstrate that we can classify gesture descriptors (hand shape, hand action and facial region occluded) significantly higher than a naive baseline. To our knowledge, this work is the first attempt to automatically detect and classify hand-over-face gestures in natural expressions.

Keywords— facial expression, gesture, recognition, multimodal interface

I. INTRODUCTION

Facial expression has been a focus of research in human behavior for over a hundred. Applications of facial expression analysis include marketing [1], perceptual user interfaces, human-robot interaction [2,3, 4], drowsy driver detection [5], telenursing [6], pain assessment [7], analyzing mother-infant interaction [8], autism [9], social robotics [10, 11], facial animation [12, 13] and expression mapping for video gaming [14] among others.

While there is a considerable body of prior research on automatic facial expression recognition and lip reading, there has been relatively little work examining the possible role of the face in direct, intentional interactions with computers or other machines. This may be partly due totechnological limitations: how can information about motor actions of the mouth be acquired in an non-encumbering, non-invasive fashion? With the extensive work on facial expression recognition over the past decade [13], however, vision-based methods now offer a realistic solution to this obstacle.

The unusual nature of the idea of using the face for intentional interaction may be another factor in the relative dearth of precedent studies, however novelty or lack off amiliarity of a concept should not deter research. In this paper we review several of our projects in this area to support the thesis that facial gesture HCI can be natural, useful, and fun. Recently we have been using vision-based methods to capture movement of the head and facial features and use these for intentional, direct interaction with computers. Two of the projects that will be reviewed below allow textentry involving motion of the head and/or mouth. A further two projects discussed in this work explore the concept of using motion of the mouth for artistic and musical expression. While our primary intention is to suggest that facial actions could provide an input channel for HCI which is parallel to and independent of action of the hands, it could also be of use for motor-impaired computer users.

In this paper, we present an analysis of hand-over-face gestures in a naturalistic video corpus of complex mental states. We define three hand-over-face gesture descriptors, namely hand shape, hand action and facial region occluded and propose a methodology for automatic detection of face occlusions in videos of natural expressions.

We treat the problem as two separate tasks: detection of hand occlusion; and classification of hand gesture descriptors. The main contributions of this paper are:

1. Proposing a mutli-modal fusion approach to detect hand-over-face gestures in videos of natural expressions, based on state-of-the-art spatial and spatio-temporal appearance features.

2. Proposing the first approach to automatically codeand classify hand-over-face gesture descriptors, namely hand shape, hand action and facial region occluded.

3. Demonstrating that multi-modal fusion of spatial andspatio-temporal features outperforms single modalitiesin all of our classification tasks.

Automated facial image analysis confronts a series of challenges. The face andfacial features must be detected in video; shape or appearance information must be extracted and then normalized for variation in pose, illumination and individual differences; the resulting normalized features are used to segment and classify facialactions. Partial occlusion is a frequent

challenge that may be intermittent or continuous(e.g., bringing an object in front of the face, self-occlusion from head turns, eye glassesor facial jewelry). While human observers easily accommodate for changes in pose, scale, illumination, occlusion, and individual differences, these and other sources of variation represent considerable challenges for computer vision. Then there is the machine-learning challenge of automatically detecting actions that require significant training and expertise even for human coders. There is much good research to do.

2. Proposed Framework

A block diagram of the components that comprise the proposed approach is depicted in Fig. 1.The first block in Fig. 1 is the hand and face tracker. This component is responsible for identifying and tracking hand

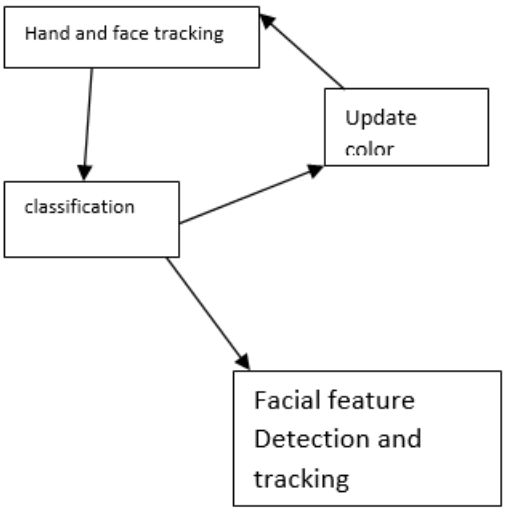


Fig. 1 Block diagram of the proposed system for hands and face tracking

and face blobs based on their color and on the information of whether they lay in the image foreground or not. The second step of the proposed system involves the classification of the resulting tracks into tracks that belong to facial blobs and tracks that belong to hands; left and right hands are also classified separately in this step. Hand trajectories are forwarded to the hand-gesture recognition system (not described in this paper) while facial regions are further analyzed to detect and track-specific facial features (eyes, nose and mouth) and to facilitate facial gestures and expression recognition at a later processing stage of the system (not part of this paper).Blobs classified as faces are also used to update the color distribution of skin-pixels, thus enabling the algorithm to quickly adapt to illumination changes. In the following sections,we describe each of the above-mentioned components in detail.

Figure 2: (a) Proposed human model (b) example input to the system

We propose a multimodal analyzer to recognize face and body gesture using computer vision and machine learning techniques.To our best knowledge there is no such an attempt to combine face and body gesture for nonverbal behavior analysis and recognition. For our multimodal analyzer we will use a human model including the face (eyes, eyebrows, nose, lips and chin) and the upper body (trunk, two arms and two hands) as shown in the Fig. 2. Hence, multi-modalitywill be achieved by combining facial expression and body language.Our system will perform the following tasks respectively: (a) locating human body and face; (b) segmentation of interest points; (c) feature extraction; (d) facial actionrecognition; (e) upper- limb action recognition; (f) fusion of the multimodal data and classification of the actions. Given the fact that we will base our systemimplementation on existing systems and techniques, we give an overview of the previous work on facial expression and gestures and their usage in HCI.

3 Hand and face tracking

In this work, hand and face regions are detected as solid blobs of skin-colored, foreground pixels and they are tracked over time using the propagated pixel hypotheses algorithm[10]. This specific tracking algorithm allows the tracked regions to move in complex trajectories, change their shape, occlude each other in the field of view of the camera and vary in number over time.

Face detection is an initial step in most automatic facial expression recognition systems. For real-time, frontal face detection, the Viola and Jones [125] face detectoris arguable the most commonly employed algorithm. See [135] for a survey of recentadvances in face detection. Once the face is detected two approaches to registrationare common. One performs coarse registration by detecting a sparse set offacial features (e.g., eyes) in each frame. The other detects detailed features (i.e.dense points around the eyes and other facial landmarks) in the video sequence.In this section we will describe a unified framework for the latter, which we referto as Parameterized Appearance Models (PAMs).

Research in psychology has indicated that at least six emotions are universally associated with distinct facial expressions [6,7,8]. Several other emotions, and many combinations of emotions have been studied but remain
unconfirmed as universally distinguishable. The six principal emotions are: happiness, sadness, surprise,

fear,
anger, and disgust.

Most psychological research on facial expressions has been conducted on "mug-shot" pictures. These pictures allow one to detect the presence of static cues (such as wrinkles) as well as the position and shape of the facial features. Few studies have directly investigated the influence of the motion and deformation of facial features on the interpretation of facial expressions. Bassili suggested that motion in the image of a face would allow emotions to be identified even with minimal information about the spatial arrangement of features [8].

3.1. Vision Based Facial Expression Recognition

Within the past decade, analysis of human facial expression has attracted interest in machine vision and artificial intelligence areas to build systems that understand and use this non-verbal form of human communication.

Most of the systems that automatically analyze the facial
expressions can be broadly classified into two categories:

(1) systems that recognize prototypic facial expressions corresponding to basic emotions (happy, angry etc.)

(2) systems that recognize facial actions (eyebrow raise, frown etc.)

There has been a significant amount of research on creating systems that recognize a small set of prototypic emotional expressions, i.e., joy, surprise, anger, sadness, fear, and disgust from static images or image sequences. This focus on emotion-specified expressions follows from the work of Ekman [6,7] who proposed that basic emotions have corresponding prototypic facial expressions.

3.2. Systems that Recognize Prototypic Facial Expressions Automatic facial expression analysis is done in two different ways: from static images or from video frames. The studies based on facial expression recognition from static images are performed by presenting subjects with photographs of facial expressions and then analyzing their relationship between components of the expressions and judgments made by the observers. These judgment
studies rely on static representations of facial expressions
with two facial images: a neutral face and an expressive
face. The use of such stimuli has been heavily criticized
by Bassili since "judgment of facial expression hardly ever takes place on the basis of a face caught in a state similar to that provided by a photograph snapped at 20 milliseconds" [8]

4 Feature Extraction

The first building block of our approach is feature extraction. We chose features that can effectively represent hand gesture descriptors that we want to detect. Therefore, we extract spatial features, namely: Histograms of Oriented Gradients (HOG) and facial landmark alignment likelihood. Moreover, having the detection of hand action in mind, we also extract Space Time Interest Points (STIP) that combine spatial and temporal information. For HOG and STIP features, dimensionality reduction of features is then applied to obtain a more compact feature representation.

4.1 Space Time Features

Local space-time features [17, 18, 9] have become popular motion descriptors for action recognition [24]. Recently, they have been used by Song et al. [27] to encode facial and body micro expressions for emotion detection. They were particularly successful in learning the emotion valence dimension as they are sensitive to global motion in the video. Our methodology for space time interest points feature extraction and representation is based on the approach proposed by Song et al. [27].Space Time Interest Points (STIP) capture salient visual patterns in a space-time image volume by extending the local spatial image descriptor to the space-time domain. Obtaining local space-time features is a two step process: spatio-temporal interest point (STIP) detection followed by feature extraction. Wang et al. [28] reports that using the Harris3D interest point detector followed by a combination of the Histograms of Oriented Gradient (HOG) and the Histogram of Optical Flow (HOF) feature descriptors provide good performance. Thus, we use the Harris3D detector with HOG/HOF feature descriptors to extract local sparse-time features. As we are interested in the face area, we use the face alignmentinput to crop the STIP features and discard any extracted points outside the face region.

The STIP box in the overview diagram in Figure 2 shows how the hand motion is captured by the space-time features (denoted by the yellow circles in the diagram). The local space-time features extracted are dense as they capture micro-expressions. Since we are interested in moresemantic feature representation, we use sparse coding to represent them so that only few salient features are recovered, i.e. features that appear most frequently in the data. Thus, we learn a codebook of features and use it to encode the extracted features in a sparse manner.

The goal of sparse coding is to obtain a compact representation of an input signal using an over-complete codebook, i.e. the number of codebook entries is larger than the dimension of input signal so that only a small number of codebook entries are used to represent the input signal. Given an input signal $x \in RN$ and over-complete codebook $D \in RN \times K$, $K \gg N$, we find a sparse signal $\alpha \in RK$ that minimises the reconstruction error,

4.2 Classifying between hands and faces

The hand and face tracker described in the previous section provides a set of blob tracks that correspond to the location of hands and faces of people that are in front of the robot. To proceed with higher level tasks, like hand gestures and facial expressions recognition, one has to distinguish between tracks that belong to hands and tracks that belong to faces. Moreover, for hand tracks, one has to know which tracks belong to left hands and which tracks belong to right hands.

Towards this goal, we have developed a technique that incrementally classifies a track into one of three classes: faces, left hands and right hands.

The input of the technique is a feature vector Ot which is extracted at each time instant t and is used to update the belief of the robot Bt regarding the class F of each track. The feature vector Ot consists of the following components:

The periphery-to-area ratio rt of the current track's blob. The ratio rt is normalized to the corresponding ratio of a circle and provides a measure of the complexity of the blob's contour. It is expected that hands will generally have more complex contours than faces, i.e. larger values for rt .

The vertical and the horizontal components ut and vt of the speed of a tracked skin-colored blob. The intuition behind this choice is that hands are generally expected to move faster than faces. Moreover, faces are not expected to have large vertical components in their motion.

The orientation θt of the blob. It is expected that faces will tend to have orientations close to $\pi/2$.

The location lt of the blob within the image. This location is relative to the location of each possible head hypothesis and it is normalized according to the radius of this head, as it will be explained later in this section.

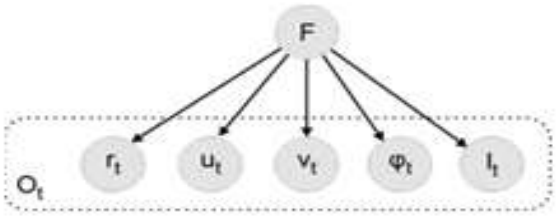
We define the belief Bt of the robot at time instant t to be the probability that the track belongs to class f , given all observations Oi up to time instant t . That is:

$$Bt = P(F = f | O1, \ldots, Ot-1, Ot)$$

By assuming the Markov property and the independence assumptions indicated by Fig. 4, the computation of Bt can be simplified as:

$$Bt = \alpha P(Ot | F = f) Bt-1$$

The above equation defines an incremental way to compute B_t , i.e. to classify the track by incrementally improving the belief B_t based on the previous belief B_{t-1} and the current observations. α is a normalization factor which ensures that the beliefs B_t for all possible values of F sum up to one.
To compute the term $P(O_t | F = f)$ in the right hand of



Eq. (4), we assume the naive Bayes classifier depicted in the graph of Fig. 5, which gives:

$$P(O_t | F) = P(r_t | F) P(u_t | F) P(v_t | F) P(\theta_t | F) P(l_t | F) \quad (5)$$

All the probabilities in the right side of Eq. (5) can be estimated according to training data and encoded and stored
in appropriate look-up tables that permit real-time computations

Fig. 4 Bayes graph encoding the independence assumptions of ourapproach

Fig. 5 The naive Bayes classifier used to compute the P (Ot |F = f)

4.3 Systems that Recognize Facial Actions

The evidence for seven universal facial expressions does not imply that these emotion categories are sufficient to describe all facial expressions [18]. Although prototypic expressions, like happy, surprise and fear, are natural, they occur infrequently in everyday life and provide an incomplete description of facial expression. Emotion is communicated by changes in one or two discrete facial features, such as tightening the lips in anger or obliquely lowering the lip corners in sadness [18]. Further, there are emotions like confusion, boredom and frustration for which any prototypic expression might not exist. To capture the subtlety of human emotion and paralinguistic communication, automated recognition of fine -grained changes in facial expression is needed.

Hence, vision-based systems that recognize facial actions were introduced. Generally, the approaches that attempt to recognize action units (AUs) are motivated by Paul Ekman's Facial Action Coding System (FACS) [6].

5. Gesture

Gesture is the use of motions of the limbs or body as a
means of expression, communicate an intention or feeling
[28]. Gestures include body movements (e.g., palm-down, shoulder-shrug), and postures (e.g., angular

distance) and often occur in conjunction with speech,
thus, the emblematic gestures that can replace speech are

not considered as gesture [3].In noisy situations, humans depend on access to more than one modality, and this is when the non-verbal modalities come in to play [3,28]. It has been shown that when speech is ambiguous or in a speech situation with some noise, listeners do rely on gestural cues [3,59].

The essential nature of gestures in the communicative situation is demonstrated by the extreme rarity of 'gestural errors'. That is, although spoken language is commonly quite disfluent, full of false starts, hesitations, and speech errors, gestures virtually never portray anything but the speaker's communicative intention [3].
According to McNeill [30], speakers may say"left" and mean "right", but they will probably point towards the right. Listeners may correct speakers' errors, on the basis of the speaker's gestures. Thus, gestures serve an important communicative function in face-to-face communication [3,30].

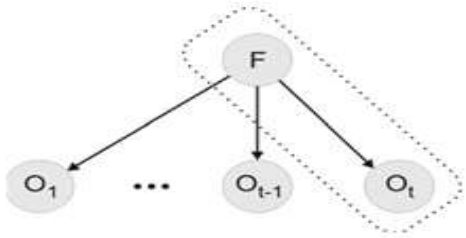
Many of the hand movements speakers make when they speak are unconnected to the content of their speech (e.g., smoothing one's hair). However, the majority of hand gestures produced by speakers are meaningfully connected to speech. Kendon, has situated these communicative hand movements along a "gesture continuum" [28], defining five different kinds of gestures:

1) Gesticulation - spontaneous movements of the hands and arms that accompany speech.

2) Language-like gestures - gesticulation that isintegrated into a spoken utterance, replacing a particular poken word or phrase.

3) Pantomimes - gestures that depict objects or actions,with or without accompanying speech.

4) Emblems - familiar gestures such as "V for victory", "thumbs up", and assorted rude gestures (often culturally specific).



5) Sign languages - Linguistic systems, such asAmerican Sign Language, which are well defined.

Moving from gesticulation to emblems along the continuum, the presence of speech declines; the presence of language-like properties increases; and idiosyncratic gestures are replaced with socially regulated signs, spontaneity decreases, and social regulation increases.

5.1. Vision Based Gesture Recognition Systems

Gesture recognition is the process by which gestures made by the user are made known to the system. During recognition, static position (posture/pose) together with spontaneous gestures is considered.For the past decade, there has been a significant amount of research in the computer vision community on extracting facial motion, interpreting human activity, and recognizing particular hand/arm gestures.

However, the concept of gesture is loosely defined, and depends on the context of the interaction. Gestures can be static, where the user assumes a certain pose or configuration, or dynamic, defined by movement.

McNeill [34] defines three phases of a dynamic gesture: pre-stroke, stroke, and post-stroke. Some gestures have both static and dynamic elements, where the pose is important in one or more of the gesture phases; this is particularly relevant in sign languages. When gestures are produced continuously, each gesture is affected by the gesture that preceded it, and possibly by the gesture that follows it.There are several aspects of a gesture which may be relevant and therefore may need to be represented explicitly in computer vision systems. Hummels and Stappers [35] describe four aspects of a gesture which may be important to its meaning:

(a) Spatial information - where it occurs, locations a gesture refers to; (b) Pathic information - the path which a gesture takes; (c) Symbolic information - the sign that a gesture makes; (d) Affective information - the emotional quality of a gesture.

Automatically segmenting gestures is difficult, and is often finessed or ignored in current systems by requiring a starting position in time and/or space [36].Recognition of natural, continuous gestures requires temporally segmenting gestures by distinguishing intentional gestures from other "random" movements.Since gestures vary, from one person to another, it is essential to capture the invariant properties of gesture and use this for representation.Currently, most computer vision systems for recognizing gestures look similar. Components of a gesture recognition system are [36]:

(1)Sensing human position, configuration, and movement using cameras and computer vision

techniques - the output of initial processing is a time -varying sequence of parameters describing position, velocities, and angles of the relevant body part.

(2)Preprocessing - images are normalized, enhanced, or transformed in some manner

(3) Gesture Modeling and Representation - transforming the input into the appropriate representation (featurespace) and then classifying it from a database of predefined gesture representations ; selection of suitable characteristics that ensure an accurate representation of the gesture; determination of the smallest number of characteristics, so as the recognition task to be accomplished in short time period (a) spatial features from posture and motion (b) temporal features-(preparation, stroke, hold, recovery) [34].

(4) Feature Extraction and Gesture Analysis - Extraction of the features (statistical properties or estimated body parameters); computing the parameters from image features that are extracted from sequences; description of pose and trajectory; localization, tracking and selection of suitable image features.

(5) Gesture Recognition and Classification - classifying gestures by using template matching (from a database of predefined gesture representations); geometric feature classification; using neural networks; time -compressing templates; HMMs or Bayesian networks.

5.2. Overview of Approaches and Techniques Used

An overview of work up to 1995 in hand gesture modeling, analysis, and synthesis is presented by Huang and Pavlovic in [31].

Features representation techniques: Features are represented by analyzing trajectory [37]; motion [38]; color, intensity, edges, silhouettes and contours [40]; or by parametric eigen space representation [37,39]

Feature Detection and Localization Techniques: Features are located by using various techniques such as segmentation, filtering, edge detection, morphological skeletinization [41, 42, 43]; and motion analysis (i.e. recognize themotion of the arm/hand)

Gesture Recognition Techniques: The gesture recognition approaches can be classified into three major categories: (a) model based, (b) appearance based and (c) motion based. Model based approaches focus on recoveringthree-dimensional model parameters of articulated body parts. Appearance based approaches use two -dimensional information such as gray scale images or body silhouettes and edges. And motion based approaches attempt to recognize the gesture directly from the motion without any structural information about the physical body. In all these approaches, the temporal properties of the gesture are typically handled using Dynamic TimeWarping (DTW) or statistically using Hidden Markov Models (HMM).

Static gesture or pose recognition can be accomplished by a straightforward implementation of using template matching, geometric feature classification, neural networks, or other standard pattern recognition techniques such as parametric eigenspace to classify pose. [37,39]. Dynamic gesture recognition requires

consideration of temporal events, typically accomplished through the use of techniques such as time -compressing templates, dynamic time warping, hidden Markov models (HMMs), and Bayesian networks. (e.g. [44]).

Analysis, recognition and synthesis of natural gestures is still an ongoing research [3,42,43]. The latest work on gesture recognition can be found in the upcoming FG 2004 Conference (IEEE Face and Gesture Recognition Conference) held every two years.

The work presented by Picard et al. [53] is the only single work combining different modalities for automatic analysis of affective physiological signals. This work automatically recognizes eight user-defined affective states (neutral, anger, hate, grief, platonic love, romantic love, joy, and reverence) from a set of sensed physiological signals. Five physiological signals have been

recorded: electromyogram from jaw (coding the muscular tension of the jaw), blood volume pressure (BVP), skin conductivity, respiration, and heart ratecalculated from the BVP. For emotional classification, an algorithm combining the sequential floating forward search and the Fisher projection has been used, which achieves an average correct recognition rate of 81.25 percent.

For further reviews of the recent attempts of combining facial expressions and vocal cues, the readers are referred to Pantic and Rothkrantz [1] for a survey of current efforts.

6. Evaluation of facial feature tracking

Different test data sets exist for evaluating algorithms mainly for facial expression or affect recognition. In our case, these data served as an alternative option to further assess individually the method for the detection and tracking of facial features, namely localization accuracy of the

individual features within a given face area. The databases used in our experiments are the Cohn-Kanade (CK) facial expression database [31], the FABO [32] and the BIOID database [33].

The CK database includes 486 greyscale image sequences from 97 posers exhibiting various facial expressions. Each sequence begins with a neutral expression and proceeds to a peak expression in the last frame. The FABO database contains videos in RGB mode (1,024× 768 pixels) of face and body expressions of 23 subjects recorded by face and body cameras simultaneously. In our experiments, 1,010 videos from the face cameras were used for testing. Both databases were used for facial feature localization evaluation based on visual validation. Namely, the eye and mouth regions were detected and tracked in all image sequences and results of successful localization were visually verified by a human

Fig. 6 Filed trial results.

supervisor. High success rates were achieved for the localization of eyes and mouth in all image sequences of both databases. False positives did not occur in any image and both the left and right eye were correctly localized in a 93% of the images in the CK database and the mouth area in 98% of them. In the FABO database only visible facial features were considered. In 95 and 96% of the images, the left/right eye and the mouth were correctly localized, respectively.

The fact that the images were recorded with a high frame rate, led to an increased success rate in feature tracking, since differences in the relative position and shape of features between consecutive frames were not considerably large. Results from selected frames from the CK and FABO data-base are illustrated in Fig. 7.

Fig. 7 Frames from the ck database

7. Discussion

Due to being an uncovered research area, there exist problems to be solved and issues to be considered in order to develop a robust multimodal analyzer of face and body gesture using computer vision and machine learning techniques.

A potential issue to consider in our work is that gesture analysis is even more context-dependent than face action analysis. For this reason, as an initial starting point, we clearly want to distinguish between gesture expressions and gesture actions, as in the evolution process of facial expression to facial action recognition. We aim to build a system which is first of all capable of visually classifying gesture actions such as "crossing arms", "moving hands", and “shrugging shoulders". The affective interpretation of them is later demanded to the interpretation stage which could fuse this information with the other modes. Another issue to consider is that the information content of natural body gestures is reasonably lower than that of the face and is still an ongoing research. Expressions could be detected from face actions alone to a certain level of accuracy [5,6,7,8]. The same level of accuracy may not be achieved by natural gestures alone [3,29,30]. Untangling the grammar of human behavior still represents a rather unexplored topic even in the psychological and sociological research areas [1].

The issue that makes this problem even more difficult to solve is that detection of gesture actions could be technically more challenging than face actions. There is a greater intrinsic visual complexity, facial features never occlude each other and they are not deformable; instead, limbs are subject to occlusions and deformations. This expected lower detection accuracy might even worsen expression recognition rather than improve it. However, the use of gesture actions could be an auxiliary mode to be used only when expressions from the remaining modes are classified as ambiguous. Moreover, fusing the information from the different modes is still an open problem in general. According to Pantic when different modalities are coupled for usage in multimodal HCI, fusion of the data can be accomplished at three levels: data, feature and decision level (see [1]). Thus, fusion could be (a) done early or late in the interpretation process; (b) some mode could be principal/otherauxiliary. Most likely, this cannot be modeled explicitly but rather found out by statistical decomposition methods such as PCA.

A further potential issue to consider is that gestures might be more context (speaker)-dependent than facial actions. Different speakers might use different gesture actions to express a same emotion, to a higher degree of variance than they would do with face actions. Our body language has higher variance than our face language, at a parity of ethnicity, age, culture, and also has dependency to the grammar of a person’s behavioral actions/reactions, to his context (i.e., to where he is and to what he is doing at this point), and to the current scenario. Machine learning can be used as a source of help to potentially learn application-, user-, and context-dependent

rules by watching the user's behavior in the sensed context [1].

Besides these standard visual-processing problems, there is another cumbersome issue typical for multimodality: Development of robust multimodal methods requires access to databases that combine face and body gesture with possible other modalities such as vocal and tactile information. However, no readily accessible common database of test material that combines different modalities has been established yet.

Due to the potential greater context - dependency of gesture actions and issues discussed above, our system will explicitly separate the layer of gesture action detection from that of interpretation. The interpretation layer will explicitly consider the input of context information to add the detected gestures with a correct semantic. How to generate the context information will be considered as an external and independent problem

8. Conclusion and further work

The real-time facial gesture recognition system we have developed consists of two modules running in parallel; a Face Tracker and a Gesture Recogniser. The face tracking module fuses information from the vision system with information derived from a two-dimensional model of the face using multiple Kalman filters.

Our system is able to track the features without special
illumination or makeup. It can track features change shade, deform or even disappear. We have had experimental success in all situations where a person
turns their head 60 degrees (it is physically difficult to turn further!). Further rotation increases the risk of
loosing all features since the initial templates are all
taken from images with the person looking straight into
the camera. Our system does recover from such
situations by using dynamic search. However, the recovery can take several seconds. In future work we plan
to introduce a 3D-model of the face which will allow us
to more precisely predict the position of the face.

Another improvement we are considering is to grab
templates of the features dynamically while the system is
tracking the face. This would not only improve the
tracking., but the system would also cope with much greater ranges of changing illumination. We are planning to create a dynamic face model that adapts to the gathered data. Such a dynamic system would learn how to track the face of a unknown person. The system would be initially provided with several generic faces including startup templates and face geometries. It selects the most similar model for the unknown person and then learns the exact templates and geometry.

Our Gesture Recogniser module which runs in parallel with the face tracking module is capable of recognising a wide variety of gestures based on head movements. Gesture recognition is robust due to the statistical approach we have adopted. If future we plan to record and analyse the head gestures of a large sample of people. The statistical parameters of head motion will be incorporated into our program

References

[1]G.H. Shergill, H. Sarrafzadeh, O. Diegel, and A. hekar. Computerized sales assistants: Theapplication of computer technology to measure consumer interest;a conceptual framework.Journal of Electronic Commerce Research, 9(2):176–191, 2008.

[2] A. van Dam. Beyond wimp. Computer Graphics and Applications, 20(1):50–51, 2000.

[3] V.W. Zue and J.R. Glass. Conversational interfaces: Advances and challenges. Proceedingsof the IEEE, 88(8):1166–1180, 2002.

[4] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. PatternAnalysis and Machine Intelligence, 22(1):107–119, 2000.

[5] E. Vural, M. Bartlett, G. Littlewort, M. Cetin, A. Ercil, and J. Movellan. Discrimination ofmoderate and acute drowsiness based on spontaneous facial expressions. In ICPR, 2010.

[6] Y. Dai, Y. Shibata, T. Ishii, K. Hashimoto, K. Katamachi, K. Noguchi, N. Kakizaki, andD. Ca. An associate memory model of facial expressions and its application in facial expressionrecognition of patients on bed. In ICME, pages 591 – 594, 2001.

[7] P. Lucey, J.F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K.M. Prkachin.Automatically Detecting Pain in Video Through Facial Action Units. IEEE Transactions onSystems, Man, and Cybernetics, Part B: Cybernetics, (99):1–11, 2010.

[8] Erika E. Forbes, Jeffrey F. Cohn, Nicholas B. Allen, and Peter M. Lewinsohn. Infant affectduring parent-infant interaction at 3 and 6 months: Differences between mothers and fathersand influence of parent history of depression. Infancy, 5:61–84, 2004.

[9] M. Madsen, R. el Kaliouby, M. Eckhardt, M. Hoque, M. Goodwin, and R.W. Picard. Lessonsfrom participatory design with adolescents on the autism spectrum. In Proc. Computer HumanInteraction, 2009.

[10] M. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facialexpression recognition: Development and

applications to human computer interaction. InCVPR Workshops for HCI, 2003.

[11] V. Bruce. What the human face tells the human mind: Some challenges for the robot-humaninterface. In IEEE Int. Workshop on Robot and Human Communication, 1992.

[12] H. Lo and R. Chung. Facial expression recognition approach for performance animation. InInternational Workshop on Digital and Computational Video, 2001.

[13] B. J. Theobald and J. F. Cohn. Facial image synthesis. Oxford University Press., 2009.

[14] D. Huang and F. De la Torre. Bilinear kernel reduced rank regression for facial expressionsynthesis. In ECCV, 2010.

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. InCVPR, 2001.

[16] C. Zhang and Z. Zhango. A survey of recent advances in face detection. In Technical Report.MSR-TR-2010-66 Microsoft Research., June 2010.

[17] I. Laptev. On space-time interest points. International

Journal of Computer Vision, 64(2-3):107-123, 2005.

[18] B. de Gelder. Why bodies? twelve reasons forincluding bodily expressions in affective neuroscience.Phil. Trans. of the Royal Society B, 2009.

[19] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie.

Behavior recognition via sparse spatio-temporalfeatures. In Visual Surveillance and PerformanceEvaluation of Tracking and Surveillance, 2005.

[20] R. Poppe. A survey on vision-based human action

recognition. IVC, 2010.

[21] Y. Song, L.-P. Morency, and R. Davis. Learning a

sparse codebook of facial and body microexpressionsfor emotion recognition. In ICMI, 2013.

[22] H. Wang, M. M. Ullah, A. Klaser, I. Laptev,C. Schmid, et al. Evaluation of local spatio-temporalfeatures for action recognition. In BMVC, 2009.

[23] Joseph C. Hager and Paul Ekman,Essential Behavioral Science ofthe Face and Gesture that Computer Scientists Need to Know, 1995.

[24] P. Ekman and W. V. Friesen. The Facial Action Coding System: A Technique for Measurement of Facial Movement. Consulting Psychologists Press, San Francisco, CA, 1978.

[25] D. McNeill, (1985). So you think gestures are nonverbal Psychological Review, 92, 350-371.

[26] C. Hummels and P. Stappers, "Meaningful gestures for human computer interaction: beyond hand gestures," Proc. Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, Apr. 1998.

[27] M. Turk, Gesture Recognition, To appear in the Handbook of Virtual Environment Technology. Stanney, K. Ed., Lawrence Erlbaum Associates, Inc.

[28] J. Cassell. A framework for gesture generation and interpretation. In R. Cipolla and A. Pentland, editors, Computer vision in human-machine interaction . Cambridge University Press, 2000.

[29] Baltzakis, H., Argyros, A.: Propagation of pixel hypotheses formultiple objects tracking. In: Proceedings of the International Sym posium on Visual Computating (ISVC), Las Vegas, Nevada, USA, November 2009.