Face and Body Gesture Recognition for a Vision-Based Multimodal Analyzer

Hatice Gunes, Massimo Piccardi, Tony Jan
Computer Vision Research Group, University of Technology, Sydney (UTS)
PO Box 123 Broadway 2007 NSW – Australia

{haticeg,massimo,jant}@it.uts.edu.au

Abstract

For the computer to interact intelligently with human users, computers should be able to recognize emotions, by analyzing the human's affective state, physiology and behavior. In this paper, we present a survey of research conducted on face and body gesture and recognition. In order to make human-computer interfaces truly natural, we need to develop technology that tracks human movement, body behavior and facial expression, and interprets these movements in an affective way. Accordingly in this paper, we present a framework for a vision-based multimodal analyzer that combines face and body gesture and further discuss relevant issues.

Keywords: facial expression, gesture, recognition, multimodal interface, affective HCI

1. Introduction

There are different ways a human expresses his emotions, as well as expressing them verbally, expressing the emotions also involves non-verbal means and physically sensible actions. When we are face-to-face with another human, no matter what our language, cultural background, or age, we all use our faces, hands and body as an integral part of our communication with others; faces change expressions continuously and spontaneous gestures occur accompanying our speech.

According to Mehrabian 93 percent of our communication is nonverbal and the most expressive way humans display emotions is through facial expressions and body gestures [5]. Considering the effect of the message as a whole, spoken words of a message contributes only for 7 percent, the vocal part contributes for 38 percent, while facial expression of the speaker contributes for 55 percent to the effect of the spoken message [5].

There is good reason to think that non-verbal behavior will play an important role in evoking some social communicative attributions. Cassell's research shows that humans are more likely to consider computers human-like when those computers display appropriate nonverbal communicative behavior [3]. Hence, understanding human emotions through nonverbal means is one of the necessary skills both for humans to interact effectively with each other and for the computers to interact intelligently with their human users.

For the computer to interact intelligently with human users, computers should be able to recognize emotions, by analyzing the human's affective state, physiology and behavior. In order to make human-computer interfaces truly natural, we need to develop technology that tracks human movement, body behavior and facial expression, and interprets these movements in an affective way.

Recent advances in image analysis and machine learning open up the possibility of automatic measurement of face and body signals. For instance, automatic analysis of facial expressions has rapidly become an area of intense interest in computer vision and artificial intelligence research communities.

This paper analyzes various existing systems and techniques used for automatic facial expression and body gesture recognition and discusses the possibility of a vision based multi-modal system that combines face and body signals to analyze human emotion and behavior. The rationale for this attempt of combining face and body gesture for a better understanding of human non-verbal behavior is the recent interest and advances in multi-modal interfaces. Pantic and Rothkrantz in [1] clearly state the importance of a multimodal affect analyzer. The modalities considered are visual, auditory and tactile, where visual mainly stands for facial actions analysis. The interpretation of other visual cues such as body language (natural/spontaneous gestures) is not explicitly addressed in [1]. However, we think that this is an important component of affective communication and this will be a major goal of the proposed system in this paper.

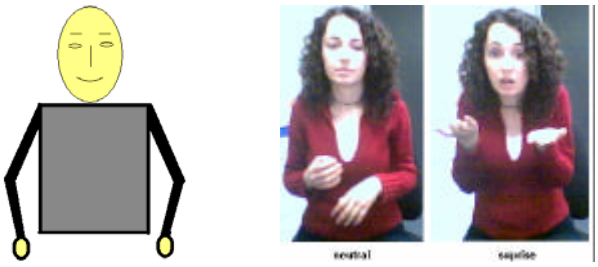
An automated system that senses, processes, and interprets face and body signals has great potential in various research and application areas including video conferencing, video telephony, video surveillance, animation/synthesis of life-like agents and the automated tools for psychological research [1,2,3,4]. An automated multi-modal system combining face and body gesture will find use in creating perceptual user interfaces to facilitate virtual visits to Internet sites. It would have

applications in human-computer interaction and pervasive perceptual man-machine interfaces for developing affective machines and computers that will understand human emotions and will be able to respond intelligently [53,54]. The machine that could understand behavioral cues about various emotional and social situations can be used to assist humans in tasks that require people to make decisions based on a number of social and emotional variables.

The paper is organized further as follows. Section 2 presents a brief description of our multimodal approach on combining face and body gesture for human emotion recognition, Section 3 covers the previous work on facial expression and the automation of facial expression analysis, while Section 4 explores gestures and their usage in HCI applications. Section 5 presents the possible efforts toward automatic multimodal analyzers of human affective state, Section 6 discusses the potential issues and problems. Finally, Section 7 gives the conclusion.

2. Proposed Framework

Face and body gestures are two of the several channels of nonverbal communication that occur together. Messages can be expressed through face and gesture in many ways. For example, an emotion such as sadness can be communicated through facial expression, a lowered head position, relaxed muscles, and lethargic movement. Thus, various nonverbal channels can be combined for the construction of computer systems that can affectively communicate with humans.



(a) **(b)**

Figure 1: (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. 1. Hence, multi-modality will 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 action recognition; (e) upper- limb action recognition; (f) fusion of the multimodal data and classification of the actions.

Given the fact that we will base our system implementation on existing systems and techniques, we give an overview of the previous work on facial expression and gestures and their usage in HCI applications in Section 3 and Section 4, and present existing multimodal analysis of human affective state in Section 5, respectively. We further discuss the challenges and potential problems we might face in our system implementation in Section 6.

3. Facial Expression

Facial expressions can indicate emotion and pain, regulate social behavior, and reveal brain function. Facial expression measurement provides an indicator of emotion activity and is presently used in a variety of areas of behavioral research.

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 the 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].

Facial expression recognition from image sequences is based on categorizing 5-7 dasses of prototypic facial expressions by tracking facial features and measuring the amount of facial movement. There are various approaches that have been explored. Some of those include analysis of facial motion (Mase [14]; Yacoob and Davis [15], Essa and Pentland [13]) measurements of the shapes and facial features and their spatial arrangements (Lanitis, Taylor, and Cootes [16]), holistic spatial pattern analysis using techniques based on principal components analysis (Padgett and Cottrell [17] ; Lanitis, Taylor, and Cootes [16]) and methods for relating face images to physical models of the facial skin and musculature (Mase [14]; Essa and Pentland [13]). All these methods are similar in that they first extract some features from the images or video, then these features are used as inputs into a classification system, and the outcome is one of the pre-selected emotion categories. They differ mainly in the features extracted and in the classifiers used to distinguish between the different emotions.

3.3. 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].

3.3.1. Facial Action Coding System (FACS)

Ekman and Friesen [6] developed the Facial Action Coding System (FACS) for describing facial expressions by action units (AUs). The system is based on the enumeration of all "action units" of a face that cause facial movements. As some muscles give rise to more than one action unit, the correspondence between action units and muscle units is approximate. Of 44 FACS AUs defined, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are for upper face, and 18 are for lower face. The anatomic basis of the remaining 14 is unspecified. These 14 are referred to in FACS as miscellaneous actions. A FACS coder "dissects" an expression, decomposing it into specific AUs that produced the motion.

3.3.2. Previous Work on Recognizing Facial Actions

Some of the previous work to directly recognize action units has used optical flow across the entire face or facial feature measurement.

Mase [14] and Essa [13] described patterns of optical flow that corresponded to several AUs, but did not attempt to recognize them. Essa and Pentland [13] and Yacoob and Davis [15] proposed methods to analyze expressions into elementary movements using an animation style coding system inspired by FACS. Padgett and Cottrell [17] found that local principal component analysis was superior to full-face Eigenfaces for expression recognition. Cowie et al [19] describe a system to recognize facial expressions by identifying Facial Animation Parameter Units (FAPUs) defined in MPEG-4 standard by feature tracking of Facial Definition Parameter(FDP) points, also defined in MPEG-4 frame-work.

The CMU/Pittsburgh and UCSD groups are among the most important research groups that have focused on automatic FACS recognition as a tool for behavioral research.

From the CMU/Pittsburgh group, Tian and Kanade [10] developed an automatic AU analysis system using facial features to recognize 16 action units and any combination of those. The shape of facial features like eyes, eyebrow, mouth and cheeks are described by multistate templates. The parameters of these multistate templates are used by a Neural Network based classifier to recognize the action units. The degree of manual pre-processing is reduced by using automatic face detection. However, the system requires that the templates be initialized manually in the first frame of the sequence, which prevents it from being fully automatic. The system has achieved average recognition rates of 96.4 percent for upper face AUs and 96.7 percent for lower face AUs.

Bartlett et al. [12] and Donato et al. [11] from the UCSD group compared holistic spatial analysis, explicit measurement of features (local feature analysis) such as wrinkles, and estimation of motion flow fields and combined them in a hybrid system which classified 6 upper facial actions but no AUs occurring in combinations. The system achieved 91 percent accuracy. However, only results on manually pre-processed image sequences were reported.

The facial gesture recognition system in face profile image sequences is another significant work developed by Pantic and Rothkrantz [20]. Their system analyzes subtle changes in facial expressions based on profile-contour fiducial points in a profile-view video [21]. For tracking the profile face, a profile contour and 10 profile-contour fiducial points are extracted. 20 individual AUs

occurring alone or in a combination are recognized by a rule-based method and the recognition rate of 85 percent is achieved. In addition, Pantic also proposed a self-adaptive facial-expression analyzer that classifies detected facial muscle activity into multiple, quantified, user-defined interpretation categories.

Kapoor and Picard [22] describe a fully automatic framework that requires no manual intervention to analyze facial activity. The work is focused on recognizing upper action units (AUs 1,2,4,5 and 7). The system detects the pupils using an infrared sensitive camera equipped with infrared LEDs. For each frame, the pupil positions are used to localize and normalize eye and eyebrow regions, which are analyzed using PCA to recover parameters that relate to the shape of the facial features. These parameters are used as input to classifiers based on Support Vector Machines to recognize upper facial action units and all their possible combinations. The system achieved a recognition accuracy of 62.5 percent for all possible AU combinations. However, the system breaks when the subjects are wearing glasses. Since the system uses infrared LEDs, it can be confused by the presence of strong direct sunlight. It also needs to be extended to recognize lower facial action units.

From the previous work done on automating FACS coding, the automatic AU analyzers presented by Tian et al. [10] and Pantic [20] perform the best: They code 16 and, respectively, 29 AUs occurring alone or in a combination in face images. Both systems can automatically detect AU4, AU6, AU7, AU9, AU10, AU12, AU20, AU25, AU26, and AU27.

For further reviews of the past attempts to analyze facial expressions and actions, the readers are referred to Donato et al. [11] and Tian *et al.* [10] for a review of techniques for detecting facial actions, and Pantic and Rothkrantz [2] for a survey of current efforts.

3.4. Problem Domain

Facial expression analysis with computer vision and machine learning techniques includes various problems to be considered. Here, we will analyze four of the sub-problem areas in automatic facial expression analysis and classification [2, 9]:

(1) creating/using a facial expression database

(2) detecting and/or tracking the face in a facial image or image sequence

(3) extracting the information from the face

(4) Classifying the facial expression into different categories

3.4.1. Facial Expression Database

Development of robust methods of facial expression analysis requires access to databases that adequately sample from the problem space of facial expression analysis. However, most investigators have used relatively limited data sets, hence the generalizability of the various facial expression analysis methods remained unknown.

There is not one single facial expression database of images that is used commonly by all different facial expression research communities [1,2,4]. In general, each research community has created and used their own facial expression database.

There have been some attempts to create comprehensive test-bed for comparative studies of facial expression analysis. The most famous of these being the Cohn-Kanade AU-Coded Face Expression Image Database [9] and Ekman-Hager Facial Action Exemplars .

The database of Ekman–Hager facial action exemplars is not published publicly. The database has been used by Bartlett *et al.* [12], Donato *et al.* [11], and Tian *et al.* [10] to train and test their methods for detecting facial actions from face image sequences. The Cohn–Kanade AU-coded face expression image database is the only facial database made publicly available, and is used only by Tian *et al.* up to now [1]. It is a large, representative facial expression database that can be used as a basis for comparison for efforts in the research area, for use in both training and testing of algorithms for facial expression analysis. However, a larger, well defined, more representative, validated, and commonly used database of images of faces (both still and motion) is still needed as a test-bed with which to evaluate different approaches in an objective manner.

3.4.2. Face Detection

Current facial expression recognition systems assume, in general, that the presence of a face in the scene is ensured and some global location of the face in the scene is known a priori. Moreover, the conditions under which a facial image or image sequence is obtained are controlled having a uniform background and the images mostly contain frontal facial view. Therefore, the problem of locating faces is a segmentation problem (in machine vision) or a detection problem in pattern recognition) [1]. Far from being a challenge in these systems, face detection process highly depends on the type of input image: static face images or face image sequences. Various face detection techniques can be used as mentioned and compared in [27].

In most of the existing systems, images contain portraits of faces with no facial hair or glasses, the illumination is constant, the subjects are young and of the same ethnicity. Few of the current systems deal with rigid head motions (and example is the system proposed Ebine *et al.* [23]) and only the method of Essa and Pentland [13] can handle distractions like facial hair and glasses. None the automated facial affect analyzers proposed in the literature up to date "perceives" a whole face when a part of it is occluded [2]. Also, though the conclusions generated by an automated facial expression analyzer are affected by input data certainty, except for the system proposed by Pantic [20], none existing system for automatic facial expression analysis calculates the output data certainty based upon an input data certainty [1].

3.4.3. Feature Extraction

Feature extraction largely depends on the face representation chosen. Face is represented in three ways: holistic, analytic and hybrid. Analytic face representations is used, where the face is modeled as a set of facial points or as a set of templates fitted to the facial features such as the eyes and the mouth. Feature-based method localizes the features of an analytic face model and template-based method fits a holistic face model. Hybrid face representation is a combination of analytic and holistic approaches. (eg. a set of facial points used to determine an initial position of a template that models the face). The features to be extracted can be either physical (eyes, brows etc) or appearance based (that represent movements and positions of facial feature) .Most of the proposed approaches in facial expression analysis are directed toward automatic, static, analytic, 2-D facial feature extraction [1]. Still, many of the proposed systems do not extract facial information in an automatic way [2,10,12].

Existing systems use various approaches for automatic facial-data extraction. These include *Facial Motion and Optical Flow*[13,14,15]; *Motion Energy Map* [13]; *Feature Measurement* [10,20]; and *Model-based Techniques/ Holistic Analysis* [12] (PCA "Eiegen Actions", Local Feature Analysis (LFA); Fisher Actions;Independent component analysis [24]; Local Representations [17]; Gabor Wavelet representation [16]).

3.4.4. Classification

Facial expression classification categorizes the extracted expression data either as a particular face action or a particular emotion (happiness, surprise, fear, anger, disgust and sadness), or both. Recent studies point out the possibility of the automatic classification of facial expressions into multiple emotion categories [20].

The classification methods used can be broadly categorized as:

Template-based: the facial expression extracted is compared to the templates pre-defined for each category to find the best match.

Rule-based (fuzzy): classifies the facial expression into the basic emotion categories based on the previously encoded facial actions and by finding the category it fits.

Statistical pattern recognition techniques: Uses ANNs for static images and can classify the expression into multiple classes. Further classification techniques used by the existing systems in image sequences include HMM(e.g.,[25]) and Bayesian classification. (e.g., [26]).

The methods mentioned are used both in static images and image sequences.

3.5. Limitations and Future Research Areas

Current systems that have obtained high recognition rates for recognizing facial action units, use manually preprocessed image sequences, require substantial human intervention or do not recognize more than prototypic expressions. Only few of the current systems acquire images by a mounted camera and only few systems deal with the automatic face detection in an arbitrary scene [2]. In addition to translation and rotation of the head, scaling (*i.e.*, moving away from the camera) is also a major concern. Most of the approaches report results on clean datasets, which are manually pre-processed videos and images of the frontal face of the subjects deliberately making facial actions in front of a camera. Moreover, faces have no facial hair or glasses, the subjects are young (i.e., without permanent wrinkles) and generally of the same ethnicity.In most of the current systems the input is not processed in real time. Categorizing complex facial expressions into one or multiple emotion categories and determining which facial expressions are related to which emotional states is still a problem to be solved. If facial expression analysis is desired in spontaneous and dynamic settings, there is a need to develop a robust system that will address each of these issues.

4. 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) G*esticulation* – spontaneous movements of the hands and arms that accompany speech.

2) *Language-like gestures* – gesticulation that is integrated into a spoken utterance, replacing a particular spoken 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 as American 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.

4.1. Gesture Types in Human-Human Communication

McNeill [29,30] categorized the gestures found in human-human communication into conscious and unconscious gesture categories.

(1)Conscious Gestures

These are consciously produced gestures. There are two types of conscious gestures:

Emblematic Gesture: These gestures are culturally specified in the sense that one single gesture may differ in interpretation from culture to culture and are *consciously* produced and therefore easier to remember [3,29,30]. For example, the American "V-for-victory" gesture can be made either with the palm or the back of the hand towards the listener. In Britain, however, a ‘V’ gesture made with the back of the hand towards the listener is inappropriate in polite society [3].

Propositional gesture: An example is the use of the hands to measure the size of a symbolic space while the speaker says "it was this big" [3].

The conscious gestures do not make up the majority of gestures found in spontaneous conversation.

(2)Unconscious / Spontaneous Gestures

The vast majority of gestures are those that although unconscious and unwitting are the gestural vehicles for our communicative intent, with other humans, and potentially with our computers as well [3,29,30].

Spontaneous (unconscious) gesture accompanies speech in most communicative situations, and in most cultures (despite the common belief to the contrary). People even gesture while they are speaking on the telephone [29]. When referring to spontaneous gestures, mostly gesticulation in the Kendon’s continuum is considered.

Within the spontaneous, speech-associated gesture McNeill (1992) defined four gesture types [29]:

Iconic – representational or pictorial gestures that represent physical entities in the world depicting some feature of the object, action or event being described. An example is "he climbed up the pipe" accompanied by the hand rising upwards to show the path [3].

Metaphoric – gestures that represent an abstract concept or a common metaphor, rather than the object or event directly. An example is "the meeting went on and on" accompanied by a hand indicating rolling motion [3].

Interactive/beats– small, formless gestures, often associated with word emphasis; physically oriented to an interlocutor that play a role in regulating the interaction and/or transitions in discourse. An example is "she talked first, I mean second" accompanied by a hand flicking down and then up on the word “second” [3].

Deictic – pointing gestures that refer to people, objects, or events in space or time. These types of gesture modify the content of accompanying speech and may often help to disambiguate speech – similar to the role of spoken intonation. An example is "Adam looked at Chuck, and he looked back" accompanied by a hand pointing first to the left and then to the right [3].

According to recent findings, the spontaneous gestures (*gesticulation* in Kendon’s Continuum) make up some 90% of human gestures.

4.2. Gesture Recognition in Computer Systems and HCI

Despite the importance of spontaneous gesture in normal human-to-human interaction, most research to date in HCI, and most virtual environment technology, focuses on emblems and sign languages in Kendom’s continuum, where gestures tend to be less ambiguous, less spontaneous and natural, more learned, and more culture-specific [3]. The computer science community mostly has attempted to integrate emblematic gestures (e.g. the thumbs up gesture, or putting one's palm out to mean stop), that are employed in the absence of speech, and emotional facial displays (e.g. smiles, frowns, looks of puzzlement). Emblematic gestures carry more clear semantic meaning and may be more appropriate for the kinds of command-and-control interaction that virtual environments tend to support [3].

Gesture is used for control and navigation in CAVEs [31,32] and in other virtual environments such as smart rooms [33] and virtual work environments. In addition, gesture may be perceived by the environment in order to be transmitted elsewhere, e.g., as a compression technique, to be reconstructed at the receiver. Gesture recognition may also influence a system’s model of the user’s state. For example, a look of frustration may cause the system to slow down its presentation of information, or the urgency of a gesture may cause the system to speed up. Gesture may also be used as a communication *backchannel* to indicate agreement, participation, attention, conversation turn-taking, etc.

For human-computer interface to be truly natural, we need to develop technology to recognize speech with face and body gesture together. Gesture recognition covers the interpretation of tracking data from different devices in order to recognize gestures. Our main interest is on passive sensing from cameras, using computer vision techniques to recognize gestures.

4.3. 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 (feature space) 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.

4.4. Gestures Used in Vision Based Gesture Recognition Systems

(1)Head and face gestures: When people interact with one another, they use an assortment of cues from the head and face to convey information. Some examples of head and face gestures include: head shake, tilt and related macro-head movements, eyebrow lift, direction of eye gaze, raising the eyebrows, opening the mouth to speak, winking, flaring the nostrils, facial expression etc. (e.g. [22]).

(2)Hand and arm gestures: Defined as hand and arm movements generally away from the body, which commonly accompany, and which appear to bear a direct relationship to, speech. (e.g. an upraised and pointed index finger). People naturally use their hands for a wide variety of manipulation and communication tasks. Besides being quite convenient, hands are extremely expressive, with approximately 29 degrees of freedom [36]. In his comprehensive thesis on whole hand input, Sturman [71] analyzed task characteristics and requirements, hand action capabilities, and device capabilities, and discussed important issues in a variety of application domains where hand can be used as a sophisticated input and control device. Many references to gesture recognition in computer vision only consider hand and arm gestures. The vast majority of automatic recognition systems are for deictic gestures (pointing), emblematic gestures (isolated signs) and sign languages (with a limited vocabulary and syntax). Some are components of bimodal systems, integrated with speech recognition [36]. Some produce precise hand and arm configuration while others only coarse motion. Mulder [72] presented an overview of hand gestures in human-computer interaction, discussing the classification of hand movement, standard hand gestures, and hand gesture interface design.

(3) Body gestures: Body gestures include full or partial body motion (e.g. movement of waist or chest, shoulder shrug etc.), body postures (postural shifts, angular distance, upright position with ankles locked etc.) or self-adaptors (e.g. rubbing the chin, scratching the cheek, smoothing the hair etc.). For recognizing body motion Bobick [47] proposed a taxonomy of motion understanding in terms of: (a) Movement – the atomic elements of motion; (b) Activity – a sequence of

movements or static configurations; (c) Action – high-level description of what is happening in context.

4.5. 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 eigenspace 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 the motion 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 recovering three-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.

5. Multimodality

Multimodal systems provide the possibility of combining different modalities that occur together to function in a more efficient and reliable way in diverse human-computer interaction applications [1,4]. Moreover, with multi-modal systems the ambiguities in the interactions can be resolved in a natural manner.

Most of the existing work provides automatic single-modal analysis by analyzing various communication cues separately. Currently, there are very few multi-modal systems introduced attempting to analyze combinations of communication means for human affective state analysis. Some examples are: the integration of gaze and gesture in a robot system showing multi-modal interaction capabilities [48] and a combination of gesture and speech for a multi-modal crisis management system [49]. There are also multimodal systems for affective emotion recognition combining auditory and visual information by processing facial expression and vocal cues. Examples of such bimodal systems are the works of Chen *et al.* [50], De Silva and Ng [51], and Yoshitomi *et al.* [52].

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 rate calculated 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. 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/other auxiliary. 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.

7. Conclusion

The paper presented an approach for a vision-based multimodal analyzer that recognizes face and body gesture by firstly presenting various approaches and previous work in automatic facial expression/action analysis, gesture recognition and multimodal interfaces.

A multi-modal interface analyzing face and body gesture will find use in a range of areas such as video surveillance, monitoring of human activity and virtual environments and help in transmitting video for teleconferencing and improve man-machine interaction.

However, due to being a fairly new research area, there still exist problems to be solved and issues to be considered in order to develop a robust, multimodal, adaptive, context-sensitive analyzer of face and body gesture using computer vision and machine learning techniques.

As has been foreseen By Pentland [4] and Nass [54], multimodal, context-sensitive information processing is to become the most widespread research topic of the AI research community and a datum of information processing in future multimedia era.

References

[1] M. Pantic and L.J.M. Rothkrantz, 'Towards an Affect-Sensitive Multimodal Human-Computer Interaction '. In: *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, September 2003

[2] M. Pantic and L.J.M. Rothkrantz,, "Automatic analysis of facial expressions: the state of the art", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1424-1445, Vol: 22, Issue: 12, Dec 2000

[3] 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.

[4]A. Pentland, "Looking at people: Sensing for ubiquitous and wearable computing," *IEEE Trans. Pattern Anal. Machine Intell*., vol. 22, pp. 107–119, Jan. 2000.

[5] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 4, pp. 53–56, 1968.

[6] 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.

[7] P. Ekman. *Emotions in the Human Faces*. Studies in Emotion and Social Interaction. Cambridge University Press, second edition edition, 1982.

[8] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psyschology*, 4:373–379, 1978.

[9]T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. FG*, 2000, pp. 46–53.

[10] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence*, 23(2), February 2001.

[11] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Pattern Analysis and Machine Intelligence*, 21(10):974–989, October 1999.

[12] M. A. Bartlett, J. C. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. Psychophysiology, 36(2):253–263, March 1999.

[13] I. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expressions. Pattern Analysis and Machine Intelligence, 7:757–763, July 1997.

[14] K. Mase. Recognition of facial expressions for optical flow. IEICE Transactions, Special Issue on Computer Vision and its Applications, E 74(10), 1991.

[15] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 70–75. IEEE Computer Society, 1994.

[16] A. Lanitis, C. J.Taylor, T. F. Cootes , Automatic Interpretation and Coding of Face Images Using Flexible Models , Transactions of pattern analysis and machine learning,July 1997,Vol. 19, No. 7, pp. 743-756

[17] C. Padgett and G. Cottrell. *Representing face images for emotion classification*. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems, volume 9, Cambridge, MA, 1997. MIT Press.

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

[19] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):33–80, January 2001.

[20] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expression," *Image Vis. Comput. J.*, vol. 18, no. 11, pp. 881–905, 2000.

[21]M. Pantic, I. Patras and L.J.M. Rothkrantz, 'Facial action recognition in face profile image sequences ', in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, vol. 1, pp. 37-40, Lausanne, Switzerland, August 2002

[22] A. Kapoor and R. W. Picard. Real-time, fully automatic upper facial feature tracking. In *Proceedings of 5th International Conference on Automatic Face and Gesture Recognition*, May 2002.

[23]H. Ebine, Y. Shiga, M. Ikeda, and O. Nakamura, "The recognition of facial expressions with automatic detection of reference face," in *Proc. Canadian Conf. ECE*, vol. 2, 2000, pp. 1091–1099.

[24]M.S. Bartlett and T.J. Sejnowski, ™Viewpoint Invariant Face Recognition Using Independent Component Analysis and Attractor Networks,º Advances in Neural Information Processing Systems, M. Mozer, M. Jordan, and T. Petsche, eds., vol. 9, pp. 817-823, Cambridge, Mass., 1997.

[25] Y. Zhu, L. C. De Silva, and C. C. Ko, "Using moment invariants and HMM in facial expression recognition," *Pattern Recognit. Lett.*, vol. 23, no. 1–3, pp. 83–91, 2002.

[26] N. Sebe, M. S. Lew, I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition using a cauchy naive bayes classifier," in *Proc. ICPR*, vol. 1, 2002, pp. 17–20.

[27] E. Hjelmas, "Face Detection: A survey.", *Computer Vision and Image Understanding*, 83: 236–274, 2001.

[28] A. Kendon. How gestures can become like words. In F. Poyatos, editor, *Cross-cultural perspectives in nonverbal communication*, New York, 1988. C.J. Hogrefe.

[29] D. McNeill, (Ed.) (2000). *Language and Gesture*, Cambridge, New York, Melbourne and Madrid: Cambridge University Press.

[30] D. McNeill. *Hand and Mind: What Gestures Reveal About Thought*. Univ. of Chicago Press, Chicago, 1992.

[31] V. Pavlovic, R. Sharma, T. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No.7, July 1997

[32] S. Shafer, J. Krumm, B. Brumitt, B. Meyers, M. Czerwinski, and D. Robbins, "The New EasyLiving Project at Microsoft Research," *Proc. Joint DARPA/NIST Smart Spaces Workshop*, Gaithersburg, Maryland, July 30-31, 1998.

[33] A. Pentland, "*Smart rooms*," Scientific Ameri- can, April 1996.

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

[35] 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.

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

[37] H. Ohno and M. Yamamoto, Gesture Recognition using Character Recognition Techniques on Two-Dimensional Eigenspace, Proceedings of ICCV 1999, pp. 151-156

[38] R. Cutler and M. Turk. *View-based interpretation of real-time optical flow for gesture recognition*. In Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition, pages 416--421, Nara, Japan, 1998. IEEE Computer Society Press

[39] P. Peixoto, J. Gonçalves, H. Araújo, Real-Time Gesture Recognition System Based on Contour Signatures", *ICPR'2002 --16th International Conference on Pattern Recognition*, Quebec City, Canada, August 11-15, 2002

[40] J. Triesch, C. Malsburg , Robotic Gesture Recognition by Cue Combination, Proceedings of the Informatik'98, 28th Annual Meeting of the Gesellschaft

[41M. Stark and M. Kohler, "*Video based gesture recognition for human computer interaction*," in W. D.-Fellner (ed.), Modeling - Virtual Worlds - Distributed Graphics, November 1995.

[42] M. Vimplis , K. Kyriakopoulos, Gesture Recognition in Realistic Images: The Statistical Approach , in the Proceedings of International Conference on Image Processing, ICIP 2002

[43] N. Gupta, P. Mittal, S. Dutta Roy, S. Chaudhury, and S. Banerjee. *A Predictive Scheme for Appearance-based Hand Tracking*. In Proc. National Conference on Communications (NCC), pages 513 -- 522, 2002.

[44]T. Starner and A. Pentland, "*Visual Recognition of American Sign Language Using Hidden Markov Models*," Proc. Int'l Workshop Automatic Face- and Gesture-Recognition, 1995.

[45] A. Mulder, "Hand gestures for HCI," Technical Report 96-1, School of Kinesiology, Simon Fraser University, February 1996.

[46]J. Sturman, "Whole-hand Input," Ph.D. Thesis, MIT Media Laboratory, Cambridge, MA, February 1992.

[47] A. Bobick, "Movement, activity, and action: the role of knowledge in the perception of motion," Royal Society Workshop on Knowledge-based Vision in Man and Machine, London, England, February, 1997.

[48] J. Zhang, T. Baier , M. Hueser , Integration of gaze and gesture detection in nature language instructing of robot in an assembly scenario, Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication, 2002.

[49] Krahnstoever, N. Schapira, E. Kettebekov, S. Sharma, R., Multimodal human-computer interaction for crisis management systems, Proceedings. Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002).

[50]L. S. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proc. ICME*, 2000, pp. 423–426.

[51] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. FG*, 2000, pp. 332–335.

[52]Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *Proc. ROMAN*, 2000, pp. 178–183.

[53] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 1175–1191, Oct. 2001.

[54] B. Reeves and C. I. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge Univ. Press, 1996.