

Face and Gesture Recognition for Human-Robot Interaction

Dr. Md. Hasanuzzaman¹ and Dr. Haruki Ueno²

¹*Department of Computer Science & Engineering, University of Dhaka*

²*National Institute of Informatics, The Graduate University for Advanced Studies, Tokyo*

¹*Bangladesh, ²Japan*

1. Introduction

This chapter presents a vision-based face and gesture recognition system for human-robot interaction. By using subspace method, face and predefined hand poses are classified from the three largest skin-like regions that are segmented using YIQ color representation system. In the subspace method we consider separate eigenspaces for each class or pose. Face is recognized using pose specific subspace method and gesture is recognized using the rule-based approach whenever the combinations of three skin-like regions at a particular image frame satisfy a predefined condition. These gesture commands are sent to robot through TCP/IP wireless network for human-robot interaction. The effectiveness of this method has been demonstrated by interacting with an entertainment robot named AIBO and a humanoid robot Robovie.

Human-robot symbiotic systems have been studied extensively in recent years, considering that robots will play an important role in the future welfare society [Ueno, 2001]. The use of intelligent robots encourages the view of the machine as a partner in communication rather than as a tool. In the near future, robots will interact closely with a group of humans in their everyday environment in the field of entertainment, recreation, health-care, nursing, etc. In human-human interaction, multiple communication modals such as speech, gestures and body movements are frequently used. The standard input methods, such as text input via the keyboard and pointer/location information from a mouse, do not provide a natural, intuitive interaction between humans and robots. Therefore, it is essential to create models for natural and intuitive communication between humans and robots. Furthermore, for intuitive gesture-based interaction between human and robot, the robot should understand the meaning of gesture with respect to society and culture. The ability to understand hand gestures will improve the naturalness and efficiency of human interaction with robot, and allow the user to communicate in complex tasks without using tedious sets of detailed instructions.

This interactive system uses robot eye's cameras or CCD cameras to identify humans and recognize their gestures based on face and hand poses. Vision-based face recognition systems have three major components: image processing or extracting important clues (face pose and position), tracking the facial features (related position or motion of face and hand poses), and face recognition. Vision-based face recognition system varies along a number of

dimensions: number of cameras, speed and latency (real-time or not), structural environment (restriction on lighting conditions and background), primary features (color, edge, regions, moments, etc.), etc. Multiple cameras can be used to overcome occlusion problems for image acquisition but this adds correspondence and integration problems.

The aim of this chapter is to present a vision-based face and hand gesture recognition method. The scope of this chapter is versatile. Segmentation of face and hand regions from the cluttered background, generation of eigenvectors and feature vectors in training phase, classification of face and hand poses, recognizes the user and gesture. In this chapter we present a method for recognizing face and gestures in real-time combining skin-color based segmentation and subspace-based patterns matching techniques. In this method three larger skin like regions are segmented from the input images using skin color information from YIQ color space, assuming face and two hands may present in the images at the same time. Segmented blocks are filtered and normalized to remove noises and to form fixed size images as training images. Subspace method is used for classifying hand poses and face from three skin-like regions. If the combination of three skin-like regions at a particular frame matches with the predefined gesture then corresponding gesture command is generated. Gesture commands are being sent to robots through TCP-IP network and their actions are being accomplished according to user's predefined action for that gesture. In this chapter we have also addressed multi directional face recognition system using subspace method. We have prepared training images in different illuminations to adapt our system with illumination variation.

This chapter is organized as follows. Section 2 focuses on the related research regarding person identification and gesture recognition. In section 3 we briefly describe skin like regions segmentation, filtering and normalization techniques. Section 4 describes subspace method for face and hand poses classification. Section 5 presents person identification and gesture recognition method. Section 6 focuses on human-robot interaction scenarios. Section 7 concludes this chapter and focuses on future research.

2. Related Work

This section briefly describes the related research on computer vision-based systems that include the related research on person identification and gesture recognition systems. Numbers of approaches have been applied for the visual interpretation of gestures to implement human-machine interaction [Pavlovic, 1997]. Major approaches are focused on hand tracking, hand posture estimation or hand pose classification. Some studies have been undertaken within the context of particular application: such as using a finger as a pointer to control TV, or manipulated Augmented desks. There are large numbers of household machine that can take benefit from the intuitive gesture understanding, such as: Microwave, TV, Telephone, Coffee maker, Vacuum cleaner, Refrigerator, etc. The aged/disabled people can access such kind of machine if its have intuitive gesture understanding interfaces.

Computer vision supports a wide range of human tasks including, recognition, navigation, communication, etc. Using computer vision to sense and perceive the user in an HCI or HRI context is often called vision-based interaction or vision-based interface (VBI). In recent years, there has been increased research on practical vision-based interaction methods, due to availability of vision-based software, and inexpensive and fast enough computer vision related hardware components. As an example of VBI, hand pose or gesture recognition offers many promising approaches for human-machine interaction (HMI). The primary goal

of the gesture recognition researches is to develop a system, which can recognize specific user and his/her gestures and use them to convey information or to control intelligent machine. Locating the faces and identifying the users is the core of any vision-based human-machine interface systems. To understand what gestures are, brief overviews of other gesturer researchers are useful.

2.1 Face Detection and Recognition

In the last few years, face detection and person identification attracts many researchers due to security concern; therefore, many interesting and useful research demonstrations and commercial applications have been developed. A first step of any face recognition or vision-based person identification system is to locate the face in the image. Figure 1 shows the example scenarios of face detection (partly of the images are taken from Rowley research paper [Rowley, 1997]). After locating the probable face, researchers use facial features (eyes, nose, nostrils, eyebrows, mouths, lips, etc.) detection method to detect face accurately [Yang, 2000]. Face recognition or person identification compares an input face image or image features against a known face database or features databases and report match, if any. Following two subsections summarize promising past research works in the field of face detection and recognition.

2.1.1 Face Detection

Face detection from a single image or an image sequences is a difficult task due to variability in pose, size, orientation, color, expression, occlusion and lighting condition. To build a fully automated system that extracts information from images of human faces, it is essential to develop efficient algorithms to detect human faces. Visual detection of face has been studied extensively over the last decade. There are many approaches for face detection. Face detection researchers summarized the face detection work into four categories: template matching approaches, feature invariant approaches, appearance-based approaches and knowledge-based approaches [Yang, 2002]. Such approaches typically rely on a static background, so that human face can be detected using image differencing. Many researches also used skin color as a feature and leading remarkable face tracking as long as the lighting conditions do not varies too much [Dai, 1996], [Crowley, 1997].

Template Matching Approaches

In template matching methods, a standard template image data set using face images is manually defined. The input image is compared with the template images and calculated correlation coefficient or/and minimum distances (Manhattan distance, Euclidian distance, Mahalanobis distance, etc.). The existence of face is determined using the maximum correlation coefficient value and/or minimal distance. For exact matching correlation coefficient is one and minimum distance is zero. This approach is very simple and easy to implement. But recognition result depends on the template images size, pose, orientation, shape and intensity.

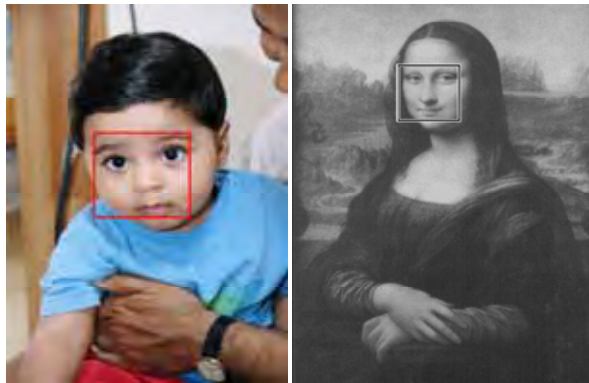
Sakai *et. al.* [Sakai, 1996] used several sub-templates for the eyes, nose, mouth and face contour to model a face which is defined in terms of line spaces. From the input images lines are extracted based on greatest gradient change and then matched against the sub-templates. The correlation between sub-images and contour templates are computed first to locate the probable location of faces. Then matching with the other sub-templates is performed at the probable face location.

Tsukamoto *et. al.* [Tsukamoto, 1994] presents a qualitative model for face [QMF]. In their model each sample image is divided into N blocks and qualitative features ('lightness' and 'edgeness') are estimated for each block. This blocked template is used to estimate "faceness" at every position of an input image. If the faceness measure is satisfied the predefined threshold then the face is detected.

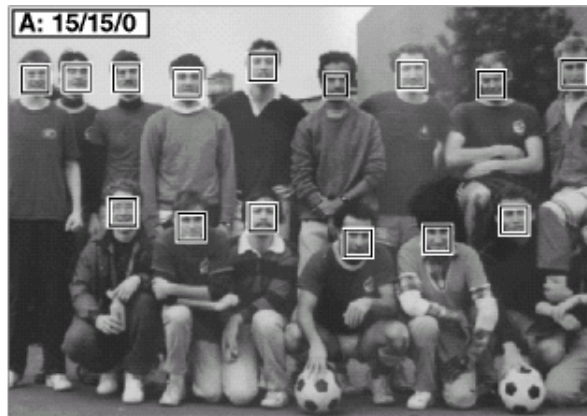
We have developed a face detection method using the combination of correlation coefficient and Manhattan distance features, calculated from multiple face templates and test face image [Hasanuzzaman, 2004a]. In this method three larger skin-like regions are segmented first. Then segmented images are normalized to match with the size and type of the template images. Correlation coefficient is calculated using equation (1),

$$\alpha_t = M_t / P_t \quad (1)$$

where, M_t is total number of matched pixels (white pixels with white pixels and black pixels with black pixels) with the t^{th} template, P_t is total number of pixels in the t^{th} template and t , is a positive number. For exact matching α_t is 1, but for practical environment we have selected a threshold value for α_t ($0 < \alpha_t \leq 1$) through experiment considering optimal matching.



(a) Single face detection



(b) Multiple faces detection

Figure 1. Examples of face detection scenarios [Rowley, 1997]

Minimum distance can be calculated by using equation (2),

$$\delta_i = \left\{ \sum_I^{x \times y} |I - G_i| \right\} \quad (2)$$

where, $I(x,y)$ is the input image and $G_1(x,y), G_2(x,y), \dots, G_t(x,y)$ are template images. For exact matching δ_i is 0, but for practical environment we have selected a threshold value for δ_i through experiment considering optimal matching. If the maximum correlation coefficient and the minimum distance qualifier support corresponding specific threshold values then that segment is detected as face and the center position of the segment is use as the location of the face.

Miao *et. al.* [Miao, 1999] developed a hierarchical template matching method for multi-directional face detection. At the first stage, an input image is rotated from -20° to $+20^\circ$ in step of 5° . A multi-resolution image hierarchy is formed and edges are extracted using Laplacian operator. The face template consists of the edges produced by six facial components: two eyebrows, two eyes, nose and mouth. Finally, heuristics are applied to determine the existence of face.

Yuille *et. al.* [Yuille, 1992] used deformable template to model facial features that fit a priori elastic model to facial features. In this approach, facial features are described by parameterized template. An energy function is defined to link edges, peaks, and valleys in the input image to corresponding parameters in the template. The best fit of the elastic model is found by minimizing an energy function of the parameters.

Feature Invariant Approaches

There are many methods to detect facial features (mouth, eyes, eyebrows, lips, hair-line, etc.) individually and from their geometrical relations to detect the faces. Human face skin color and texture also used as features for face detection. The major limitations with these feature-based methods are that the image features are corrupted due to illumination, noise and occlusion problem.

Sirohey proposed a face localization method from a cluttered background using edge map (canny edge detector) and heuristics to remove and group edges so that only the ones on the face contour are preserved [Sirohey, 1993]. An ellipse is then fit to the boundary between the head region and the background.

Chetverikov *et. al.* [Chetverikov, 1993] presented face detection method using blobs and streaks. They used two black blobs and three light blobs to represent eyes, cheekbones and nose. The model uses streaks to represent the outlines of the faces, eyebrows and lips. Two triangular configurations are utilized to encode the spatial relationship among the blobs. A low resolution Laplacian image is generated to facilitate blob detection. Next, the image is scanned to find specific triangular occurrences as candidates. A face is detected if streaks are identified around the candidates.

Human faces have a distinct texture that can be separated them from other objects. Augusteijn *et. al.* [Augusteijn, 1993] developed a method that infers the presence of face thorough the identification of face like templates. Human skin color has been proven to be an effective feature for face detections, therefore many researchers has used this feature for probable face detection or localization [Dai 1996], [Bhuiyan, 2003], [Hasanuzzaman 2004b].

Recently, many researchers are combining multiple features for face localization and detection and those are more robust than single feature based approaches. Yang and Ahuja [Yang, 1998] proposed a face detection method based on color, structure and geometry.

Saber and Tekalp [Saber, 1998] presented a frontal view-face localization method based on color, shape and symmetry. Darrel *et. al.* [Darrel, 2000] integrated stereo, color and pattern detection method to track the person in real time.

Appearance-Based Approaches

Appearance-based methods use training images and learning approaches to learn from the known face images. These approaches rely on the statistical analysis and machine learning techniques to find the relevant characteristics of face and non-face images. There are many researchers using appearance-based methods.

Turk *et. al.* [Turk, 1991] applied principal component analysis to detect and recognize face. From the training face images they generated the eigenfaces. Face images and non-face images are projected onto the eigenspaces; form feature vectors and clustered the images based on separation distance. To detect the presence of a face from an image frame, the distance between the known face space and all location in the images are calculated. If the minimum distance satisfied the faceness threshold values then the location is identified as face. These approaches are widely used by the many researchers.

Knowledge-Based Approaches

These methods use the knowledge of the facial features in top down approaches. Rules are used to describe the facial features and their relations. For example, a face is always consists of two eyes, one nose and a mouth. The relationship is defined using relative distances and positions among them. For example, the center of two eyes are align on the same line, the center points of two eyes and mouth form a triangular. Yang and Huang [Yang, 1994] used hierarchical knowledge-based method to detect face. In this method they used three layers of rules. At the first level, all possible face candidates are found by scanning a mask window (face template) over the input images, and applying a set of rules at each location. At the second level, histogram equalization and edge detection is performed on candidate faces. At the third level, using rules facial feature are detected individually and using the pre-knowledge of their relation, detect the actual faces. Kotropoulous [Kotropoulous, 1997] and other also presented rule-based face localization method.

2.1.2 Face Recognition

During the last few years face recognition has received significant attention from the researchers [Zhao, 2003] [Chellappa, 1995]. Research on automatic machine- based face recognition has started in the 1970s [Kelly 1970]. Figure 2 shows an example of face recognition scenario. The test face image (preprocessed) is matched with the face images of known persons in the database. If the face is sufficient close (nearest and support predefined threshold) to any one of the face classes, then corresponding person is identified, otherwise the person is unknown. Zhao [Zhao, 2003] *et. al.* have summarized the past recent researches on face recognition methods with three categories: Holistic matching methods, Feature-based matching methods and Hybrid methods.

Holistic Methods

These methods use the whole face region as the raw input for the recognition unit. One of the most widely used representations of the face recognition is eigenfaces, which are based on principal component analysis (PCA). The eigenface algorithm uses the principal component analysis (PCA) for dimensionality reduction and to find the vectors those are best account for the distribution of face images within the entire face image spaces. Using

PCA many face recognition techniques have been developed [Turk, 1991], [Lee, 1999], [Chung, 1999], etc.




Known Face Images	Test Image	Who is the person?
		
		
		
		Person_4
		
		
		

Figure 2. Example of face recognition scenario

Turk and Pentland [Turk, 1991] first successfully used eigenfaces for face detection and person identification or face recognition. In this method from the known face images training image dataset is prepared. The face space is defined by the "eigenfaces" which are eigenvectors generated from the training face images. Face images are projected onto the feature space (or eigenfaces) that best encodes the variation among known face images. Recognition is performed by projecting a test image onto the "facespace" (spanned by the m number of eigenfaces) and then classified the face by comparing its position (Euclidian distance) in face space with the positions of known individuals. Figure 3 shows the example of 8 eigenfaces generated from 140 training face (frontal) images of 7 persons. In this example, the training faces are 60×60 gray images.

The purpose of PCA is to find out the appropriate vectors that can describe the distribution of face images in images spaces and form another face spaces. To form principal components m -numbers of eigenvectors are used based on the eigenvalues distribution. Eigenvectors and eigenvalues are obtained from the covariance matrix generated from training face images. The eigenvectors are sorted based on eigenvalues (higher-to-lower) and selected first m -number of eigenvectors to form principal components.



Figure 3. Example of eigenfaces

Figure 4 shows the example distribution of eigenvalues for 140 frontal face images. This graph explores the eigenvalues spectrum and how much variance the first m -vectors for. In

most cases the number of eigenvectors that account for variance somewhere in the 65%-90% range.

Independent component analysis (ICA) is similar to PCA except that the distributions of the components are designed to be non-Gaussian. The ICA separates the high-order moments of the input in addition to the second order moments utilized in PCA. Bartlett *et. al.* [Bartlett, 1998] used ICA methods for face recognition and reported satisfactory recognition performance.

Face recognition system using Linear Discriminant Analysis (LDA) or Fisher Linear Discriminant Analysis (FLDA) has also been very successful. In Fisherface algorithm by defining different classes with different statistics, the images in the learning set are divided in the corresponding classes [Belhumeur, 1997]. Then, the techniques similar to those used in eigenface algorithm are applied for face classification or person identification.

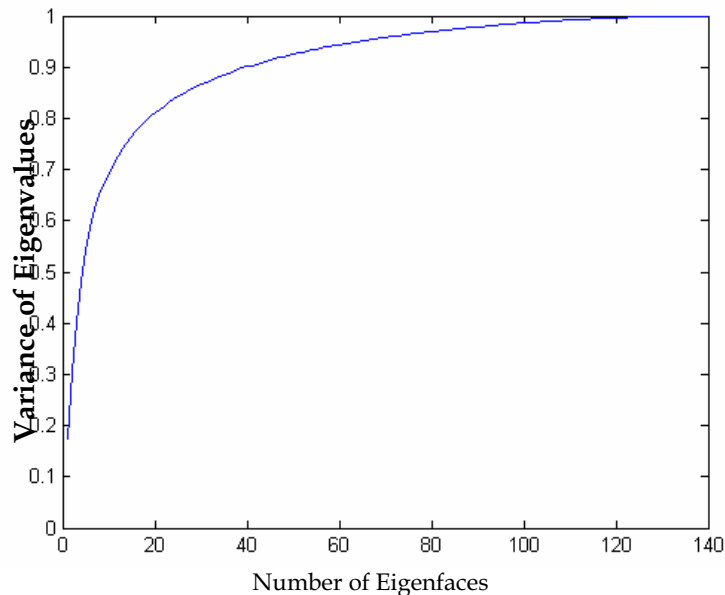


Figure 4. Example of eigenvectors spectrum for 140 eigenfaces

Feature-Based Matching Methods

In these methods facial features such as the eyes, lips, nose and mouth are extracted first and their locations and local statistics (geometric shape or appearance) are fed into a structural classifier. Kanade developed one of the earliest face recognition algorithms based on automatic facial feature detection [Kanade, 1977]. By localizing the corner of the eyes, nostrils, etc., in frontal views, that system compares parameters for each face, which were compared (using Euclidian distance metric) against the parameters of known person faces. One of the most successful of these methods is the Elastic Bunch Graph Matching (EBGM) system [Wiskott, 1997]. Other well-known methods in these systems are Hidden Markov Model (HMM) and convolution neural network [Rowley, 1997]. System based on EBGM approach have been applied to face detection and extraction, pose estimation, gender classification, sketch image based recognition and general object recognition.

Hybrid Approaches

These approaches use both holistic and features based approaches. These methods are very similar to human perception consider whole image and features individually at a time. Chung *et. al.* [Chung, 1999] combined Gabor Wavelet and PCA based approaches for face recognition and reported better accuracy than each of individual algorithm. Pentland *et. al.* [Pentland, 1994] have used both global eigenfaces and local eigenfeatures (eigeneyes, eigenmouth and eigennose) for face recognition. This method is robust against face images with multiple views.

2.2 Gesture Recognition and Gesture-Based Interface

Gestures are expressive meaningful body motions i.e., physical movements of the hands, arms, fingers, head, face or other parts of the body with the intent to convey information or interact with the environment [Turk, 2000]. People all over the world use their hands, head and other parts of the body to communicate expressively. The social anthropologists Edward T. Hall claims 60% of all our communications are nonverbal [Imai, 2004]. Gestures are used for everything from pointing at a person or an object to change the focus of attention, to conveying information. From the biological and sociological perspective, gestures are loosely defined, thus, researchers are free to visualize and classify gestures as these fit. Biologists define “gesture” broadly, stating, “the notion of gesture is to embrace all kinds of instances where an individual engages in movements whose communicative intent is paramount, manifest and openly acknowledged” [Nespoulous, 1986]. Gestures associated with speech are referred to as gesticulation. Gestures, which function independently of speech, are referred to as autonomous gestures. Autonomous gestures can be organized into their own communicative language, such as American Sign Language (ASL). Autonomous gesture can also represent motion commands to use in communication and machine control. Researchers are usually concerned with gestures those are directed toward the control of specific object or the communication with a specific person or group of people.

Gesture recognition is the process by which gestures made by the user are make known to the intelligence system. Approximately in the year 1992 the first attempts were made to recognize hand gestures from color video signals in real-time. It was the year, when the first frame grabbers for color video input were available, that could grab color images in real time. As color information improves segmentation and real time performance is a prerequisite for human-computer interaction, this obviously seems to be the start of development of gesture recognition. Two approaches are commonly used to recognize gestures, one is a gloved-base approach [Sturman, 1994] and another is a vision-based approach [Pavlovic, 1997].

2.2.1 Glove-Based Approaches

A common technique is to instrument the hand with a glove, which is equipped with a number of sensors, which provide information about hand position, orientation and flex of the fingers. The first commercially available hand tracker is the ‘Dataglove’ [Zimmerman, 1987]. The ‘Dataglove’ could measure each joint bend to an accuracy of 5 to 10 degrees, could classify hand pose correctly, but not the sideways movement of the fingers. The second hand tracker, ‘CyberGlove’ developed by Kramer [Kramer, 1989] uses strain gauges placed between the fingers to measure abduction as well as more accurate bend sensing.

Figure 5 shows the example of a 'CyberGlove' which has up to 22 sensors, including three bend sensors on each finger, four abduction sensors, plus sensors measuring thumb crossover, palm arch, wrist flexion and wrist abduction [Blinghurst, 2002]. Once the gloves have captured hand pose data, gestures can be recognized using a number of different techniques. Neural network approaches or statistical template-matching approaches are commonly used to identify static hand poses [Fels, 1993]. Time dependent neural network and Hidden Markov Model (HMM) are commonly used for dynamic gesture recognition [Lee, 1996]. In this case gestures are typically recognized using pre-trained templates, however gloves can also be used to identify natural or untrained gestures. Glove-based approaches provide more accurate gesture recognition than vision-based approaches but they are expensive, encumbering and unnatural.



Figure 5. The 'CyberGlove' for hand gesture recognition [Blinghurst, 2002]

2.2.2 Vision-Based Approaches

Vision-based gesture recognition systems can be divided into three main components: image processing or extracting important clues (hand shape and position, face or head position, etc.), tracking the gesture features (related position or motion of face or hand poses), and gesture interpretation (based on collected information that support predefined meaningful gesture). The first phase of gesture recognition task is to select a model of the gesture. The modeling of gesture depends on the intent-dent applications by the gesture.

There are two different approaches for vision-based modeling of gesture: Model based approach and Appearance based approach.

The Model based techniques are tried to create a 3D model of the user hand (parameters: Joint angles and palm position) [Rehg, 1994] or contour model of the hand [Shimada, 1996] [Lin, 2002] and use these for gesture recognition. The 3D models can be classified in two large groups: volumetric model and skeletal models. Volumetric models are meant to describe the 3D visual appearance of the human hands and arms.

Appearance based approaches use template images or features from the training images (images, image geometry parameters, image motion parameters, fingertip position, etc.) which use for gesture recognition [Birk, 1997]. The gestures are modeled by relating the appearance of any gesture to the appearance of the set of predefined template gestures. A different group of appearance-based model uses 2D hand image sequences as gesture templates. For each gestures number of images are used with little orientation variations [Hasanuzzaman, 2004a]. Images of finger can also be used as templates for finger tracking applications [O'Hagan, 1997]. Some researchers represent motion history as 2D image and use it as template images for different actions of gestures. The majority of appearance-based models, however, use parameters (image eigenvectors, image edges or contour, etc.) to form the template or training images. Appearance based approaches are generally computationally less expensive than model based approaches because its does not require translation time from 2D information to 3D model.

Once the model is selected, an image analysis stage is used to compute the model parameters from the image features that are extracted from single or multiple video input streams. Image analysis phase includes hand localization, hand tracking, and selection of suitable image features for computing the model parameters.

Two types of cues are often used for gesture or hand localization: color cues and motion cues. Color cue is useful because human skin color footprint is more distinctive from the color of the background and human cloths [Kjeldsen, 1996], [Hasanuzzaman, 2004d]. Color-based techniques are used to track objects defined by a set of colored pixels whose saturation and values (or chrominance values) are satisfied a range of thresholds. The major drawback of color-based localization methods is that skin color footprint is varied in different lighting conditions and also the human body colors. Infrared cameras are used to overcome the limitations of skin-color based segmentation method [Oka, 2002].

The motion-based segmentation is done just subtracting the images from background [Freeman, 1996]. The limitation of this method is considered the background or camera is static. Moving objects in the video stream can be detected by inter frame differences and optical flow [Cutler, 1998]. However such a system cannot detect a stationary hand or face. To overcome the individual shortcomings some researchers use fusion of color and motion cues [Azoz, 1998].

The computation of model parameters is the last step of the gesture analysis phase and it is followed by gesture recognition phase. The type of computation depends on both the model parameters and the features that were selected. In the recognition phase, parameters are classified and interpreted in the light of the accepted model or the rules specified for the gesture interpretation. Two tasks are commonly associated with the recognition process: optimal partitioning of the parameter space and implementation of the recognition procedure. The task of optimal partitioning is usually addresses through different learning-from-examples training procedures. The key concern in the implementation of the

recognition procedure is computation efficiency. A recognition method usually determines confidence scores or probabilities that define how closely the image data fits each model. Gesture recognition methods are divided into two categories: static gesture or hand poster and dynamic gesture or motion gesture.

Static Gesture

Static gesture (or pose gesture) recognition can be accomplished by using template matching, eigenspaces or PCA, Elastic Graph Matching, neural network or other standard pattern recognition techniques. Template matching techniques are the simple pattern matching approaches. It is possible to find out the most likely hand postures from an image by computing the correlation coefficient or minimum distance metrics with template images.

Eigenspace or PCA is also used for hand pose classification similarly it used for face detection and recognition. Moghaddam and Pentland used eigenspaces (eigenhands) and principal component analysis not only to extract features, but also to estimate complete density functions for localization [Moghaddam, 1995]. In our previous research, we have used PCA for hand pose classification from three larger skin-like components that are segmented from the real-time images [Hasanuzzaman, 2004d].

Triesch *et. al.* [Triesch, 2002] employed the elastic graph matching techniques to classify hand posters against complex backgrounds. They represented hand posters by label graphs with an underlying two-dimensional topology. Attached to the nodes are jets, which are a sort of local image description based on Gabor filters. This approach can achieve scale-invariant and user invariant recognition and does not need hand segmentation. This approach is not view-independent, because it uses one graph for one hand posture. The major disadvantage of this algorithm is the high computational cost.

Dynamic Gesture

Dynamic gestures are considered as temporally consecutive sequences of hand or head or body postures in sequence of time frames. Dynamic gestures recognition is accomplished using Hidden Markov Models (HMMs), Dynamic Time Warping, Bayesian networks or other patterns recognition methods that can recognize sequences over time steps. Nam *et. al.* [Nam, 1996] used HMM methods for recognition of space-time hand-gestures. Darrel *et. al.* [Darrel, 1993] used Dynamic Time Warping method, a simplification of Hidden Markov Models (HMMs) to compare the sequences of images against previously trained sequences by adjusting the length of sequences appropriately. Cutler *et. al.* [Cutler, 1998] used a ruled-based system for gesture recognition in which image features are extracted by optical flow. Yang [Yang, 2000] recognizes hand gestures using motion trajectories. First they extract the two-dimensional motion in an image, and motion patterns are learned from the extracted trajectories using a time delay network.

2.2.3 Gesture-Based Interface

The first step in considering gesture-based interaction with intelligent machine is to understand the role of gesture in human-to-human communication. There are significant amount of researches on hand, arm and facial gesture recognition, to control robot or intelligent machine in recent years. This sub-section summarizes some promising existing gesture recognition system. Cohen *et. al.* [Cohen, 2001] described a vision-based hand gesture identifying and hand tracking system to control computer programs, such as browser of PowerPoint or any other applications. This method is based primarily on color matching and is performed in several distinct stages. After color-based segmentation,

gestures are recognized using geometric configuration of the hand. Starner *et. al.* [Starner, 1998] proposed real-time American Sign Language (ASL) recognition using desk and wearable computer based video. The recognition method is based on the skin color information to extract hands poster (pose, orientation) and locate their position and motion. Using Hidden Markov Models (HMM) this system recognizes sign language words but vocabulary is limited to 40 words. Utsumi *et. al.* [Utsumi, 2002] detected predefined hand pose using hand shape model and tracked hand or face using extracted color and motion. Multiple cameras are used for data acquisition to reduce occlusion problem in their system. But in this process there incurs complexity in computations. Watanabe *et. al.* [Watanabe, 1998] used eigenspaces from multi-input image sequences for recognizing gesture. Single eigenspaces are used for different poses and only two directions are considered in their method. Hu [Hu, 2003] proposed hand gesture recognition for human-machine interface of robot teleoperation using edge features matching. Rigoll *et. al.* [Rigoll, 1997] used HMM-based approach for real-time gesture recognition. In their work, features are extracted from the differences between two consecutive images and target image is always assumed to be in the center of the input images. Practically it is difficult to maintain such condition. Stefan Waldherr *et. al.* proposed gesture-based interface for human and service robot interaction [Waldherr, 2000]. They combined template-based approach and Neural Network based approach for tracking a person and recognizing gestures involving arm motion. In their work they proposed illumination adaptation methods but did not consider user or hand pose adaptation. Torras has proposed robot adaptivity technique using neural learning algorithm [Torras, 1995]. This method is extremely time consuming in learning phase and has no way to encode prior knowledge about the environment to gain the efficiency.

3. Skin Color Region Segmentation and Normalization

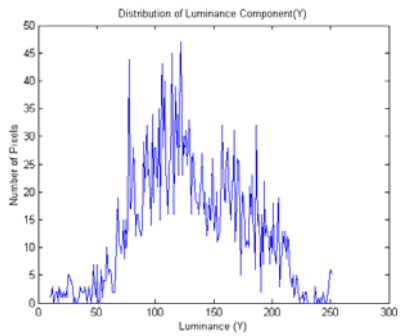
Images containing faces and hand poses are essential for vision-based human-robot interaction. But still it is very difficult to segment face and hand poses in real time from the color images with cluttered background. Human skin color has been used and proven to be an effective feature in many application areas, from face detection to hand tracking. Since face and two hands may present in the images at a specific time in an image frame, three largest skins like regions are segmented from the input images using skin color information. Several color spaces have been utilized to label pixels as skin including RGB, HSV, YCrCb, YIQ, CIE-XYZ, CIE-LUV, etc. However, such skin color models are not effective where the spectrum of the light sources varies significantly. In this study YIQ (Y is luminance of the color and I, Q are chrominance of the color) color representation system is used for skin-like region segmentation because it is typically used in video coding and provides an effective use of chrominance information for modeling the human skin color [Bhuiyan, 2003], [Dai, 1996].

3.1 YIQ-Color Coordinate Based Skin-Region Segmentation

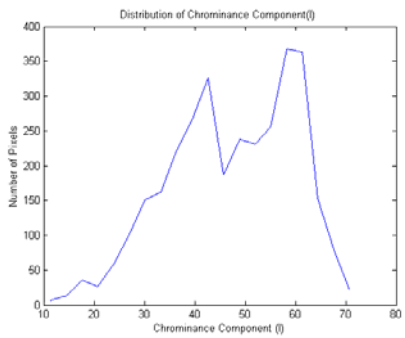
To detect human face or hand, it is assumed that the captured camera images are represented in the RGB color spaces. Each pixel in the images is represented by a triplet $P=F(R,G,B)$. The RGB images taken by the video camera are converted to YIQ color representation system (for detail please refer to Appendix A). Skin color region is determined by applying threshold values $((Y_Low < Y < Y_High) \&\& (I_Low < I < I_High) \&\& (Q_Low < Q < Q_High))$ [Hasanuzzaman, 2005b].



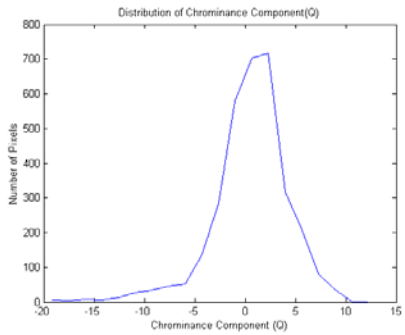
(a) Face Image of User "Cho"



(c) Y-component distributions of face "Cho"



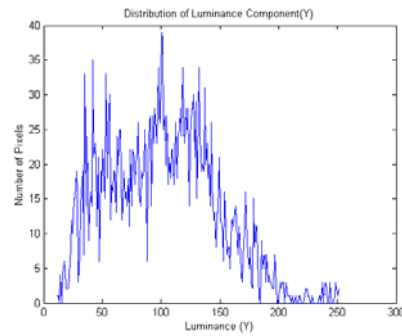
(e) I-component distributions of face "Cho"



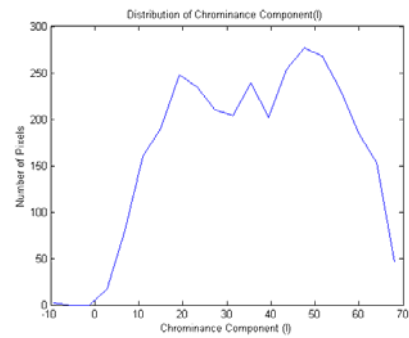
(g) Q-component distributions of face "Cho"



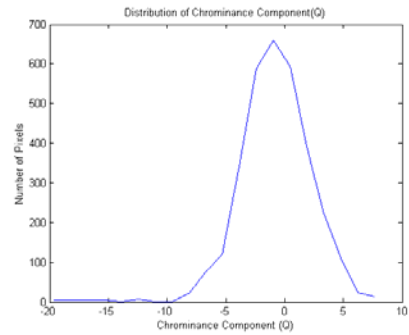
(b) Face Image of User "Hasan"



(d) Y-component distributions of face "Hasan"



(f) I-component distributions of face "Hasan"



(h) Q-component distributions of face "Hasan"

Figure 6. Histograms of Y, I, Q components for different person face images

Figure 6 shows example skin regions and its corresponding Y, I, Q components distributions for every pixels. Chrominance component I, play an important role to distinguish skin like regions from non-skin regions, because it is always positive for skin regions. Values of Y and I increases for more white people and decreases for black people. We have included an off line program to adjust the threshold values for Y, I, Q, if the person color or light intensity variation affect the segmentation output. For that reason we need to manually select small skin region and non-skin regions and run our threshold evaluation program, that will represent graphical view of Y, I, Q distributions. From those distinguishable graphs we can adjust our threshold values for Y, I, Q using heuristic approach.

Probable hands and face regions are segmented from the image with the three largest connected regions of skin-colored pixels. The notation of pixel connectivity describes a relation between two or more pixels. In order to consider two pixels to be connected, their pixel values must both be from the same set of values V (for binary images V is 1, for gray images it may be specific gray value). Generally, connectivity can either be based on 4- or 8-connectivity. In the case 4-connectivity, it does not compare the diagonal pixels but 8-connectivity compares the diagonal positional pixels considering 3×3 matrix, and as a result, 8-connectivity component is more noise free than 4-connectivity component. In this system, 8-pixels neighborhood connectivity is employed [Hasanuzzaman, 2006].

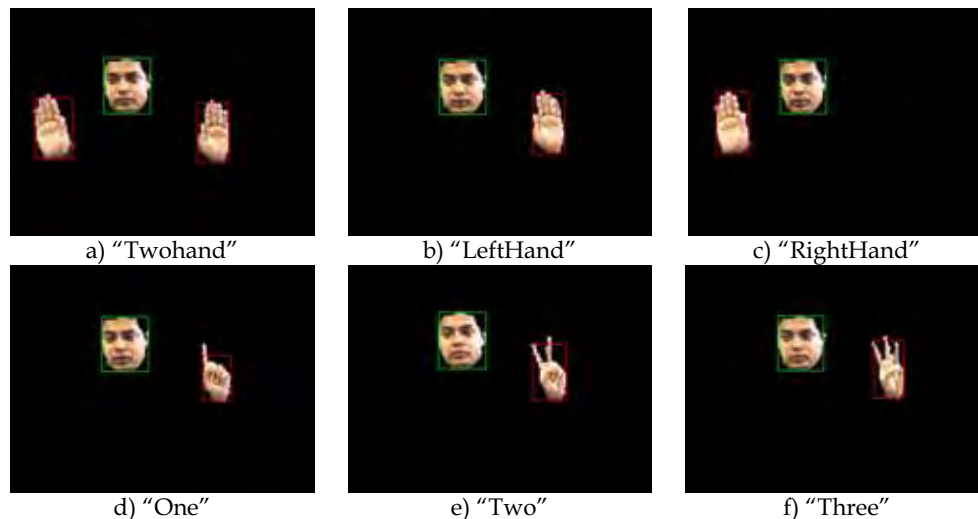


Figure 7. Example outputs of skin-regions segmentation

In order to remove the false regions from the segmented blocks, smaller connected regions are assigned by the values of black-color ($R=G=B=0$). As a result, after thresholding the segmented image may contain some holes in the three largest skin-like regions. In order to remove noises and holes, segmented images are filtered by morphological dilation and erosion operations with a 3×3 structuring element. The dilation operation is used to fill the holes and the erosion operations are applied to the dilated results to restore the shape.

After filtering, the segmented skin regions are bounded by rectangular box using height and width information of each segment: $(M_1 \times N_1)$, $(M_2 \times N_2)$, and $(M_3 \times N_3)$. Figure 7 shows the example outputs of skin like region segmentation method with restricted background. If the user shirt's color is similar to skin color then segmentation accuracy is very poor. If the user wears short sleeves or T-shirt then it needs to separate hand palm from arm. This system assumes the person wearing full shirt with non-skin color.

3.2 Normalization

Normalization is done to scale the image to match with the size of the training image and convert the scaled image to gray image [Hasanuzzaman, 2004a]. Segmented images are bounded by rectangular boxes using height and width information of each segment: $(M_1 \times N_1)$, $(M_2 \times N_2)$, and $(M_3 \times N_3)$. Each segment is scaled to be square images with (60×60) and converted it to as gray images (BMP image). Suppose, we have a segment of rectangle $P[(x^l, y^l) - (x^h, y^h)]$ we sample it to rectangle $Q[(0, 0) - (60 \times 60)]$ using following expression,

$$Q(x^q, y^q) = P\left(x^l + \frac{(x^h - x^l)}{60} x^q, y^l + \frac{(y^h - y^l)}{60} y^q\right) \quad (3)$$

Each segment is converted as gray image (BMP image) and compared with template/training images to find the best match. Using the same segmentation and normalization methods training images and test images are prepared, that is why result of this matching approach is better than others who used different training/template image databases. Beside this, we have included training/template images creation functions in this system so that it can adapt with person and illumination changes. Figure 8 shows the examples of training images for five face poses and ten hand poses.

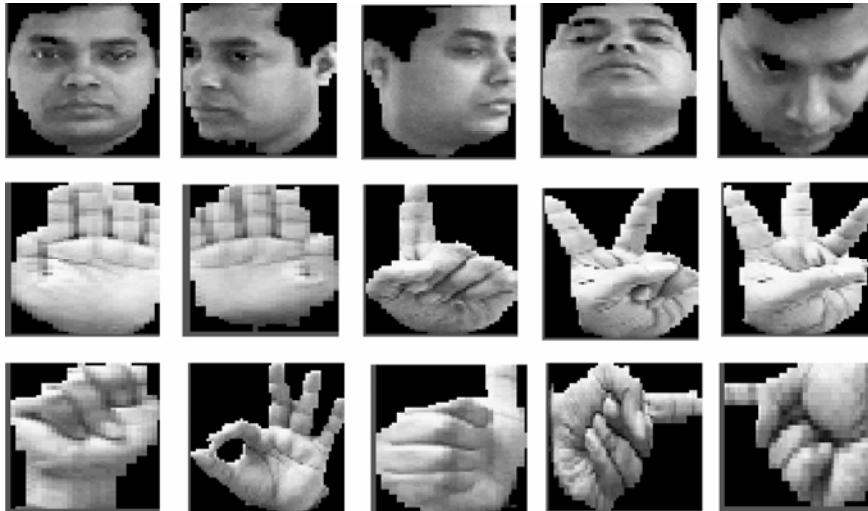


Figure 8. Examples of training images

4. Face and Hand Pose Classification by Subspace method

Three larger skin like regions are segmented from the input images considering that two hands and one face may present in the input image frame at a specific time. Segmented areas are filtered, normalized and then compared with the training images for finding the best matches using pattern-matching method. Principal component analysis (PCA) method is a standard pattern recognition approach and many researchers use it for face and hand pose classification [Hasanuzzaman, 2004d]. The main idea of the principal component analysis (PCA) method is to find the vectors that best account for the distribution of target images within the entire image space. In the general PCA method, eigenvectors are calculated from training images that include all the poses or classes. But for classification a large number of hand poses for a large number of users, need large number of training datasets from which eigenvectors generation is tedious and may not be feasible for a personal computer. Considering these difficulties we have proposed pose-specific subspace method that partition the comparison area based on each pose. In pose-specific subspace method, training images are grouped based on pose and eigenvectors for each pose are generated separately. In this method one PCA is used for each pose [Hasanuzzaman, 2005b] [Hasanuzzaman, 2004c]. In the following subsection we have described the algorithm of pose-specific subspace method for face and hand pose classification, which is very similar to general PCA based algorithm.

Symbols	Meanings
$T_j^{(i)}$	Training images for i^{th} class
$u_m^{(i)}$	m^{th} Eigenvectors for i^{th} class
$\Omega_i^{(i)}$	Weight vector for i^{th} class
$\omega_k^{(i)}$	Element of weight vector for i^{th} class
Φ_i	Average image for i^{th} class
$s_l^{(i)}$	l^{th} Known image for i^{th} class
\mathcal{E}	Euclidean distance among weight vectors
$\mathcal{E}_l^{(i)}$	Element of Euclidean distance among weight vectors for i^{th} class

Table 1. List of symbols used in subspace method

Pose-Specific Subspace Method

Subspace method offers an economical representation and very fast classification for vectors with a high number of components. Only the statistically most relevant features of a class are retained in the subspace representation. The subspace method is based on the extraction of the most conspicuous properties of each class separately as represented by a set of prototype sample. The main idea of the subspace method is similar to principal component

analysis, is to find the vectors that best account for the distribution of target images within the entire image space. In subspace method target image is projected on each subspace separately. Table 1 summarizes the symbols that are used for describing pose-specific subspace method for face and hand poses classification. The procedure of face and hand pose classification using pose-specific subspace method includes following operations:

(I) Prepare noise free version of predefined face and hand poses to form training images $T_j^{(i)} (N \times N)$, where j is number training images of i^{th} class (each pose represent one class) and $j=1,2,\dots, M$. Figure 8 shows the example training image classes: frontal face, right directed face, left directed face, up directed face, down directed face, left hand palm, right hand palm, raised index finger, raised index and middle finger to form "V" sign, raised index, middle and ring fingers, fist up, make circle using thumb and fore fingers, thumb up, point left by index finger and point right by index finger are defined as pose P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14 and P15 respectively.

(II) For each class, calculate eigenvectors ($u_m^{(i)}$) using Matthew Turk and Alex Pentland technique [Turk, 1991] and chose k -number of eigenvectors ($u_k^{(i)}$) corresponding to the highest eigenvalues to form principal components for that class. These vectors for each class define the subspace of that pose [for detail please refer to Appendix B].

(III) Calculate corresponding distribution in k -dimensional weight space for the known training images by projecting them onto the subspaces (eigenspaces) of the corresponding class and determine the weight vectors ($\Omega_l^{(i)}$), using equations (4) and (5).

$$\omega_k^{(i)} = (u_k^{(i)})^T (s_l^{(i)} - \Phi_i) \quad (4)$$

$$\Omega_l^{(i)} = [\omega_1^{(i)}, \omega_2^{(i)}, \dots, \omega_k^{(i)}] \quad (5)$$

Where, average image of i^{th} class $\Phi_i = \frac{1}{M} \sum_{n=1}^M T_n$ and $s_l^{(i)} (N \times N)$ is l^{th} known images of i^{th} class.

(IV) Each segmented skin-region is treated as individual test input image, transformed into eigenimage components and calculated a set of weight vectors ($\Omega^{(i)}$) by projecting the input image onto each of the subspace as equations (4) and (5).

(V) Determine if the image is a face pose or other predefined hand pose based on minimum Euclidean distance among weight vectors using equation (6) and (7),

$$\mathcal{E}_l^{(i)} = \|\Omega^{(i)} - \Omega_l^{(i)}\| \quad (6)$$

$$\mathcal{E} = \arg \min \{\mathcal{E}_j^{(i)}\} \quad (7)$$

If \mathcal{E} is lower than predefined threshold then its corresponding pose is identified. For exact matching \mathcal{E} should be zero but for practical purposes this method uses a threshold value obtained from experiment. If the pose is identified then corresponding pose frame will be activated.

5. Face and Gesture Recognition

A number of techniques have been developed to detect and recognize face and gesture. For secure or operator specific gesture-based human machine interaction, user identification or face recognition is important. The meaning of the gesture may differ from person to person based on their culture. Suppose according to his culture, user "Hasan" uses "ThumbUp" gesture to terminate an action of robot, whereas user "Cho" uses this gesture to repeat the previous action. In order to person specific gesture interpret (i.e., gesture is same but different meaning for different users) or person dependent gesture command generation we should map user, gesture and robot action.

5.1 Face Recognition

Face recognition is important for human-robot natural interaction and person dependent gesture command generation, i.e, gesture is same but different meaning for different persons. If any segment (skin-like region) is classified as a face, then it needs to classify the pattern, whether it belongs to a known person or not. The detected face is filtered in order to remove noises and normalized so that it matches with the size and type of the training image. The detected face is scaled to be a square image with 60×60 dimension and converted to be a gray image.

This face pattern is classified using the eigenface method [Turk, 1991], whether it belongs to known person or unknown person. The face recognition method uses five face classes: frontal face (P1), right directed face (P2), left directed face (P3), up state face (P4) and down state face (P5) in training images as shown in Figure 8 (top row). The eigenvectors are calculated from the known persons face images for each face class and k-number of eigenvectors corresponding to the highest eigenvalues are chosen to form principal components for each class. For each class we have formed subspaces and projected known person face images and detected face image on those subspaces using equation (4) and (5). We get weight vectors for known person images and detected face images. The Euclidean distance is determined between the weight vectors generated from the training images and the weight vectors generated from the detected face by projecting them onto the eigenspaces using equation (6) and (7). If minimum Euclidian distance is lower than the predefined threshold then corresponding person is identified other wise result is unknown person [Hasanuzzaman, 2004c]. We have used face recognition output for human robot ('Robovie') greeting application. For example, if the person is known then robot say (" Hi, **person name**, How are you?") but for unknown person robot say ("I do not know you").

We found that the accuracy of frontal face recognition is better than up, down and more left right directed faces [Hasanuzzaman, 2004c]. In this person identification system we prefer frontal and a small left or right rotated faces. Figure 9 shows the sample outputs of face detection method. We have verified this face recognition method for 680 faces of 7 persons, where two are females. Table 2 shows the confusion matrix for the results of face recognition for 7-persons. The diagonal elements represent the correct recognition of corresponding persons. In this table, the 1st column represents the input image classes and other columns represent the recognition results. For example, among 136 face images of person "Hasan", 132 are correctly recognized as "Hasan" and 4 are wrongly recognized as another person "Vuthi".

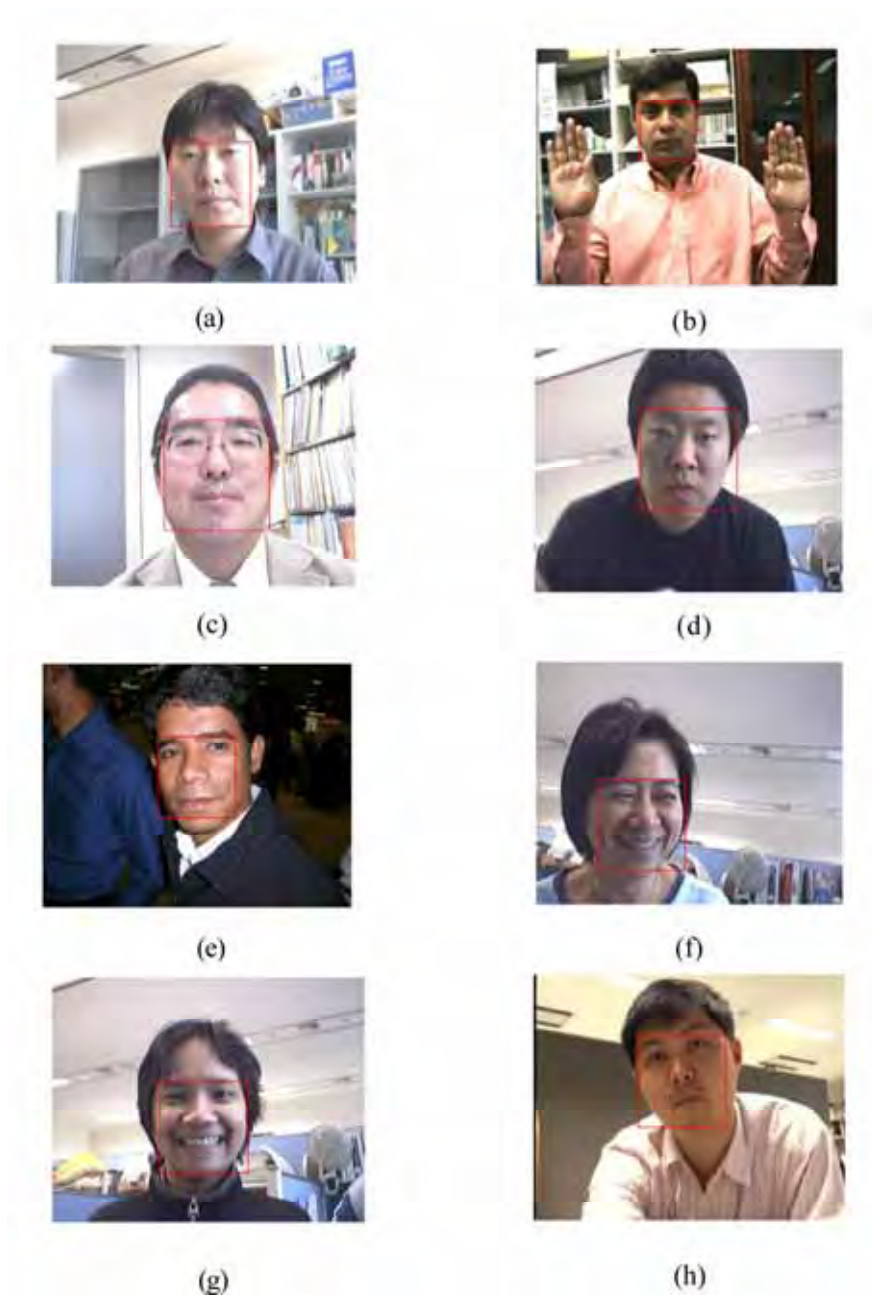


Figure 9. Sample outputs of face detection method

Table 3 presents the precisions (%) and recall rates (%) of face recognition method. The precision (%) is defined by the ratio of the numbers of correct recognition to total numbers

of recognition for each person faces. The recall rate (%) is defined by the ratio of the numbers of correct face recognition to total numbers of input faces for each person. In the case of person "Pattra" (Figure 9(d)), the precision of face recognition is very low because his face has one black spot.

Input	Hasan	Ishida	Pattara	Somjai	Tuang	Vuthi	Cho
Hasan (136)	132	0	0	0	0	4	0
Ishida (41)	0	41	0	0	0	0	0
Pattara (41)	0	0	38	3	0	0	0
Somjai (126)	0	0	5	118	3	0	0
Tuang (76)	0	0	0	10	66	0	0
Vuthi (103)	0	0	7	0	5	91	0
Cho (157)	0	0	0	0	0	0	157

Table 2. Confusion Matrix of face recognition

Person	Precision (%)	Recall (%)
Hasan	100%	97.05%
Ishida	100%	100%
Pattara	76%	92.68%
Somjai	90.07%	93.65%
Tuang	89.18%	86.84%
Vuthi	95.78%	88.34%
Cho	100%	100%

Table 3. Performance evaluation of face recognition method

5.2 Gesture Recognition

Gesture recognition is the process by which gestures made by the user are known to the system. Gesture components are the face and hand poses. Gestures are recognized using rule-based system according to predefined model with the combinations of the pose classification results of three segments at a particular image frame. For examples, if left hand palm, right hand palm and one face present in the input image then recognizes it as "TwoHand" gesture and corresponding gesture command generated. If one face and left hand open palm are present in the input image frame then recognized it as "LeftHand" gesture. Similarly others static gestures as listed in Table 4 are recognized. It is possible to recognize more gesture including new poses and new rules using this system. According to recognized gestures, corresponding gesture commands are generated and sent to interact with robot through TCP-IP network.

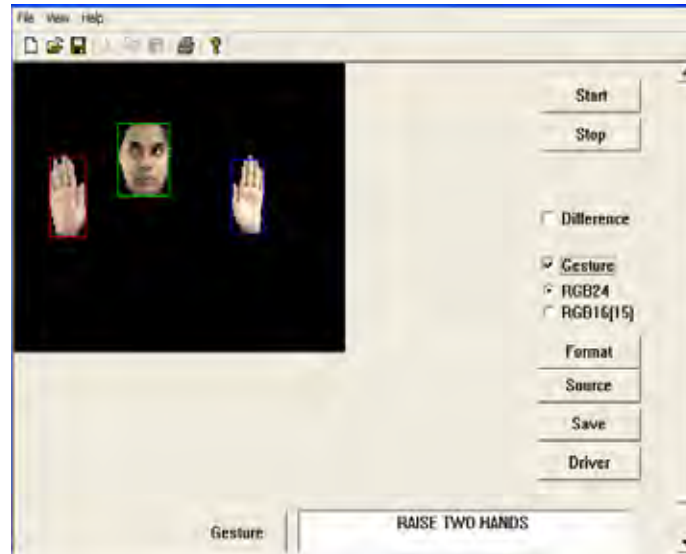


Figure 10. Sample visual output of gesture “TwoHand”

The sample output of our gesture recognition system is shown in Figure 10. This shows gesture command at the bottom text box corresponding to matched gesture, in case of no match it shows “no matching found”. Accuracy of the gesture recognition system depends on the accuracy of the pose detection system. For example: in some cases two hands and one face were present in the image but pose detection method failed to detect one hand due to variation of orientation and output of gesture recognition is then either “LeftHand” or “RightHand”. We use two standard parameters to define accuracy: precision and recall for pose classification method.

Gesture Components			Gesture names
Face	Left hand palm	Right hand palm	TwoHand
Face	Right hand palm	X	RightHand
Face	Left hand palm	X	LeftHand
Face	Index finger raise	X	One
Face	Form V sign with index and middle finger	X	Two
Face	Index, middle and ring fingers raise	X	Three
Face	Thumb up	X/Thumb up	ThumbUp
Face	Make circle using thumb and index finger	X	OK
Face	Fist up	X/Fist up	FistUp
Face/X	Point left by index finger	X	PointLeft
Face/X	Point right by index finger	X	PointRight

Table 4. Three segments combination and corresponding gesture (X=absence of predefined hand poses or face poses)

Table 5 shows the comparison of precisions and recall rates of the pose-specific subspace method and the general PCA method for face and hand poses classification. The precision (%) is defined by the ratio of the number of correct recognition to total number of recognition for each pose. The recall rate (%) is defined by the ratio of the number of correct recognition to total number of input for each pose. From the results, we conclude that precision and recall rates are higher in the subspace method and wrong classification rates are lower than the standard PCA method for majority cases. Wrong classification occurred due to orientation and intensity variation.

For this experiment we have trained the system using 2100 training images of 15 faces and hand poses of 7 persons (140 images for each pose of 7 persons). Figure 8 shows the example of 15 poses. These poses are frontal face (P1), right directed face (P2), left directed face (P3), up directed face (P4), down directed face (P5), left hand palm (P6), right hand palm (P7), raised index finger (P8), raised index and middle finger to form "V" sign (P9), raised index, middle and ring fingers (P10), fist up (P11), make circle using thumb and fore fingers (P12), thumb up (P13), point left by index finger (P14) and point right by index finger (P15). Seven individuals were asked to act for the predefined face and hand poses in front of the camera and the sequence of images were saved as individual image frame. Then each image frame is tested using the general PCA and the pose-specific subspace methods. The threshold value (for minimal Euclidian distance) for the pose classifier is empirically selected so that all the poses are classified.

Pose #	Precision (%)		Recall (%)	
	<i>Pose-specific Subspace</i>	PCA	<i>Pose-specific Subspace</i>	PCA
P1	96.21	90.37	97.69	93.84
P2	100	96.59	98.06	91.61
P3	100	93.28	99.28	99.28
P4	97.33	92.30	99.31	97.95
P5	99.21	90.90	98.43	93.75
P6	100	100	94.28	91.42
P7	97.22	96.47	100	97.85
P8	95.17	94.52	98.57	98.57
P9	97.77	97.67	94.28	90
P10	97.81	93.05	95	95
P11	100	100	92.66	87.33
P12	96.71	96.68	98	97.33
P13	99.31	100	94.66	93.33
P14	94.89	93.28	97.69	93.84
P15	100	100	100	99.33

Table 5. Comparison of pose-specific subspace method and PCA method

6. Implementation Scenarios

Our approach has been verified using a humanoid robot 'Robovie' and an entertainment robot 'Aibo'. This section describes example scenarios, which integrates gestures commands and corresponding robot behaviors. For interaction with an 'Aibo' robot, a standard CCD video camera is attached to the computer (Image analysis and recognition PC) to capture the real-time images. In the case of 'Robovie' robot, its eyes cameras are used for capturing the real time images. Each captured image is digitized into a matrix of 320×240 pixels with 24-bit color. First, the system is trained using the known training images of predefined faces and hand poses of all known persons. All the training images are 60×60 pixels gray images. In the training phase, this system generates eigenvectors and feature vectors for the known users and hand poses. We have considered robot as a server and our PC as a client. Communication link has been established through TCP-IP protocol. Initially, we connected the client PC with robot server and then gestures recognition program was run in the client PC. The result of gesture recognition program generates gesture commands and sends to robot. After getting gesture command robot acted according to user predefined actions. We have considered for human-robot interaction that gesture command will be effective until robot finishes corresponding action for that gesture.



Figure 11. Human robot ('Robovie') interaction scenario

6.1 Example of Interaction with Robovie

Figure 11 shows the example of human interaction with a 'Robovie' robot [Hasanuzzaman, 2005b]. The user steps in front of the eyes camera and raises his two hands. The image analysis and recognition module recognizes the user as 'Hasan' and classifies the three poses as 'FACE', 'LEFTHAND', 'RIGHTHAND'. This module sends gesture command

according to gesture name and user name, and selected robot function will be activated. This system implements person-centric gesture-based human robot interaction. The same gesture can be used to activate different actions for different persons even the robot is same. The robot actions are mapped based on the gesture user relationships ("gesture-user-robot-action") in the knowledge base. In this case, "Robovie" raises its two arms (as shown in Figure 11) and says "Raise Two Arms". This system has considered that gesture command will be effective until the robot finishes corresponding action for that gesture. This method has been implemented on a 'Robovie' for the following scenarios:

<p>User: "Hasan" comes in front of Robovie eyes camera, and the robot recognizes the user as Hasan.</p> <p>Robot: "Hi Hasan, How are you?" (Speech)</p> <p>Hasan: uses the gesture "ThumbUp"</p> <p>Robot: " Oh, sad, do you want to play now?" (Speech)</p> <p>Hasan: uses the gesture "Ok",</p> <p>Robot: "Thanks!" (Speech)</p> <p>Hasan: uses the gesture "TwoHand"</p> <p>Robot: imitate user's gesture "Raise Two Arms" as shown in Figure6.</p> <p>Hasan: uses the gesture "FistUp" (stop the action)</p> <p>Robot: Bye-bye (Speech).</p>	<p>User: "Cho" comes in front of Robovie eyes camera and robot recognizes the user as Cho.</p> <p>Robot: "Hi Cho, How are you?" (Speech)</p> <p>Cho: uses the gesture "ThumbUp".</p> <p>Robot: " Oh, good, do you want to play now?" (Speech)</p> <p>Cho: uses the gesture "Ok".</p> <p>Robot: "Thanks!" (Speech)</p> <p>Cho: uses the gesture "LeftHand"</p> <p>Robot: imitate user's gesture ("Raise Left Arm").</p> <p>Cho: uses the gesture "TwoHand" (STOP)</p> <p>Robot: Bye-bye (Speech)</p>
--	---

The above scenarios show that same gesture is used for different meanings and several gestures are used for the same meanings for different persons. The user can design new actions according to his/her desires using 'Robovie'.

6.2 Example of Interaction with Aibo

Figure 12 shows an example of human robot ('Aibo') interaction scenario. The system uses a standard CCD video camera for data acquisition. The user raises his index finger in front of the camera that is connected to gesture recognition PC. The image analysis and recognition module classifies the poses "FACE" and "ONE" (hand pose) and corresponding pose frames will be activated. Gestures are interpreted using three components. According to the predefined combination gesture is recognized as "One" and corresponding gesture frame is activated. The gesture recognition module recognizes the gesture is "One" and the face recognition module identifies the person as "Hasan". The user selects 'Aibo' robot for the interaction. In this combination activates the 'Aibo' for playing action 'STAND UP'.



(a) Sample visual output ("One")



(b) AIBO STAND-UP for Gesture "One"

Figure 12. Human robot ('Aibo') interaction scenario

User "Hasan"		User "Cho"	
Gesture	Aibo action	Gesture	Aibo action
One	STAND UP	TwoHand	STAND UP
Two	WALK FORWARD	One	WALK FORWARD
Three	WALK BACKWARD	Two	WALK BACKWARD
PointLeft	MOVE RIGHT	RightHand	MOVE RIGHT
PointRight	MOVE LEFT	LeftHand	MOVE LEFT
RightHand	KICK (right leg)	Three	KICK
TwoHand	SIT	FistUp	SIT
LeftHand	LIE	ThumbUp	LIE

Table 6. User-Gesture-Action mapping for Aibo

But for another user same gesture may be used for another action of 'Aibo'. Suppose user "Cho" defines the action "WALK FORWARD" for gesture "One", i.e. if user is "Cho", gesture is "One" then the 'Aibo' robot will 'Walk Forward'. In a similar way, the user can design 'Aibo' action frames according to his/her desires. The other actions of the 'Aibo' those we have used for interaction, are listed in Table 6. The scenarios in Table 6 demonstrate how the system accounts for the fact that the same gesture is used for different meanings and several gestures are used for the same meanings for different persons. The user can design new actions according to his/her desires and can design corresponding gesture for their desired actions.

7. Conclusions and future research

This chapter describes a real-time face and hand gesture recognition system using skin color segmentation and subspace method based pattern matching technique. This chapter also describes gesture-based human-robot interaction system using an entertainment robot named 'Aibo' and humanoid robot 'Robovie'. In pose-specific subspace method, training images are grouped based on pose and eigenvectors for each pose are generated separately. In this method, one PCA is used for each pose. From the experimental result we have concluded that performance of pose-specific subspace method is better than general PCA method in the same environment.

One of the major constrains of this system is that the background should be non-skin color substrate. If we used infrared camera then it is possible to overcome this problem just by a minor modification of our segmentation technique and other module will remain the same. Since the skin reflects near IR light nicely, active IR sources placed in proximity to the camera in combination with IR pass filter on the lens makes it easy to locate hands those are within the range of light sources.

Considering the reduction of processing time, so far eigenvectors calculations are performed separately in off-line. The eigenvectors do not change during dynamic learning process. The user has to initiate this calculation function to change the eigenvectors or principal components. In future, if faster CPUs are available, these components are then possible to be integrated into on-line learning function.

We could not claim that our system is more robust against new lighting condition and clutter background. Our hope is to make this face and gesture recognition system more robust and capable to recognize dynamic facial and hand gesture.

Face and gesture recognition simultaneously will help us in future to develop person specific and secure human-robot interface. The ultimate goal of this research is to establish a symbiotic society for all of the distributed autonomous intelligent components so that they share their resources and work cooperatively with human beings.

8. Appendix

8.1 Appendix A: CONVERSION FROM RGB COLOR SPACE TO YIQ COLOR SPACE

This system uses skin-color based segmentation method for determining the probable face and hands areas in an image. There are several color coordinate systems, which have come into existence for a variety of reasons. The YIQ is a universal color space used by NTSC to transmit color images using the existing monochrome television channels without increasing the bandwidth requirements. In the YIQ color model a color is described by three attributes: luminance, hue and saturation. The capture color image is represented by the RGB color coordinate system at each pixel. The colors from RGB space are converted into the YIQ space. The YIQ produces a linear transform of RGB images, which generates Y representing luminance channel and I, Q representing two chrominance channels to carry color information. The transformation matrix for the conversion from RGB to YIQ is given below [Jain, 1995],

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Where **R**, **G**, and **B** are the red, green, and blue component values which exist in the range [0, 255]. Using the following equations we can convert the images from RGB color coordinates system to YIQ color coordinate system,

$$Y = 0.299R + 0.587G + 0.114B \quad (\text{A.1})$$

$$I = 0.596R - 0.274G - 0.322B \quad (\text{A.2})$$

$$Q = 0.211R - 0.523G + 0.312B \quad (\text{A.3})$$

Images are being searched in YIQ space depending on the amount of color content of these dominant colors, that is, whether the skin color value is substantially present in an image or not. In order to segment face and hand poses in an image, the skin pixels are thresholded empirically. In this experiment, the ranges of threshold values are defined from the Y, I, Q histograms calculated for a selected skin region.

8.2 Appendix B: EIGENVECTORS CALCULATION

This section describes Eigenvectors calculation method from the training images. The major steps of the Eigenvectors calculation algorithm [Smith, 2002] [Turk, 1991] are,

Step1: Read all the training images $T_i(N \times N)$ those are two-dimensional N by N gray images, where $i=1, 2, \dots, M$.

Step2: Convert each image into a column vector

$$T_i(N^2) = T_i(N \times N) \quad (\text{B.1})$$

Step3: Calculate the mean of all images

$$\Psi = \frac{1}{M} \sum_{i=1}^M T_i \quad (\text{B.2})$$

Step4: Subtract the mean and form a big matrix with all the subtracted image data

$$\phi_i = T_i - \Psi \quad (\text{B.3})$$

$$A = [\phi_1, \phi_2, \phi_3, \dots, \phi_M] \quad (\text{B.4})$$

Step5: Calculate the Covariance of matrix 'A'

$$C = AA^T \quad (\text{B.5})$$

Step6: Calculate the Eigenvectors and Eigenvalues of the Covariance Matrix

$$\lambda_k u_k = C u_k \quad (\text{B.6})$$

Where, the vectors u_k (non-zero) and scalar λ_k are the Eigenvectors and Eigenvalues, respectively, of the Covariance matrix C. The relation between Eigenvectors and Eigenvalues of a Covariance matrix can be written using equation (B.7)

$$\lambda_k = \frac{1}{M} \sum_{n=1}^M (u_k^T \phi_n)^2 \quad (\text{B.7})$$

Using MATLAB function Eigenvectors and Eigenvalues can be calculated,

$$[\text{eigvec}, \text{eigvalue}] = \text{eig}(C) \quad (\text{B.8})$$

Each Eigenvector is of length N^2 , describe an N-by-N images and is a linear combination of the original image. Eigenvalues are the coefficient of Eigenvectors. The Eigenvectors are sorted based on Eigenvalues (higher to lower). According higher order of Eigenvalues k-numbers of Eigenvectors are chosen to form principal components.

9. References

- L. Aryananda, Recognizing and Remembering Individuals: Online and Unsupervised Face Recognition for Humanoid Robot in *Proceeding of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, Vol. 2, pp. 1202-1207, 2002. [Aryananda, 2002]
- H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho and T. Matsui, Socially Embedded Learning of the Office-Conversant Mobile Robot Iijo-2, in *Proceeding of 15th International Joint-Conference on Artificial Intelligence (IJCAI'97)*, pp.880-885, 1997. [Asoh, 1997]
- M. F. Augusteijn, and T.L. Skujca, Identification of Human Faces Through Texture-Based Feature Recognition and Neural Network Technology, in *Proceeding of IEEE conference on Neural Networks*, pp.392-398, 1993. [Augusteijn, 1993]
- R. E. Axtell, *Gestures: The Do's and Taboos of Hosting International Visitors*, John Wiley & Sons, 1990. [Axtell, 1990]
- Y. Azoz, L. Devi, and R. Sharma, Reliable Tracking of Human Arm Dynamics by Multiple Cue Integration and Constraint Fusion, in *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pp. 905-910, 1998. [Azoz, 1998]
- D. H. Ballard, Christopher M. Brown, *Computer Vision*, Prentic-Hall, INC., New Jersey, USA, 1982. [Ballard, 1982]
- M. S Bartlett, H. M. Lades, and, T. Sejnowski, Independent Component Representation for Face Recognition in *Proceedings of Symposium on Electronic Imaging (SPEI): Science and Technology*, pp. 528-539, 1998. [Bartlett, 1998]
- C. Bartneck, M. Okada, Robotic User Interface, in *Proceeding of Human and Computer Conference (Hc-2001)*, Aizu, pp. 130-140, 2001. [Bartneck, 2001]
- P.N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 19, pp. 711-720, 1997. [Belhumeur, 1997]
- M. A. Bhuiyan, V. Ampornaramveth, S. Muto, and H. Ueno, On Tracking of Eye For Human-Robot Interface, *International Journal of Robotics and Automation*, Vol. 19, No. 1, pp. 42-54, 2004. [Bhuiyan, 2004]

- M. A. Bhuiyan, V. Ampornaramveth, S. Muto, H. Ueno, Face Detection and Facial Feature Localization for Human-machine Interface, *NII Journal*, Vol.5, No. 1, pp. 25-39, 2003. [Bhuiyan, 2003]
- M. Billinghurst, Chapter 14: Gesture-based Interaction, *Human Input to Computer Systems: Theories, Techniques and Technologies*, (ed. By W. Buxton), 2002. [Billinghurst, 2002]
- L. Brethes, P. Menezes, F. Lerasle and J. Hayet, Face Tracking and Hand Gesture Recognition for Human-Robot Interaction, in *Proceeding of International Conference on Robotics and Automation (ICRA 2004)*, pp. 1901-1906, 2004. [Brethes, 2004]
- H. Birk, T. B. Moeslund, and C. B. Madsen, Real-time Recognition of Hand Alphabet Gesture Using Principal Component Analysis, in *Proceeding of 10th Scandinavian Conference on Image Analysis*, Finland, 1997. [Birk, 1997]
- R. Chellappa, C. L. Wilson, and S. Sirohey, Human and Machine Recognition of faces: A survey, in *Proceeding of IEEE*, Vol. 83, No. 5, pp. 705-740, 1995. [Chellappa, 1995]
- D. Chetverikov and A. Lerch, Multiresolution Face Detection, *Theoretical Foundation of Computer Vision*, Vol. 69, pp. 131-140, 1993. [Chetverikov, 1993]
- K. Chung, S. C. Kee, and S. R. Kim, Face Recognition using Principal Component Analysis of Gabor Filter Responses, in *Proceedings of International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pp. 53-57, 1999. [Chung, 1999]
- C. J. Cohen, G. Beach, G. Foulk, A Basic Hand Gesture Control System for PC Applications, in *Proceedings of Applied Imagery Pattern Recognition Workshop (AIPR'01)*, pp. 74-79, 2001. [Cohen, 2001]
- J. L. Crowley and F. Berard, Multi Modal Tracking of Faces for Video Communications, In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pp. 640-645, 1997. [Crowley, 1997]
- R. Cutler, M. Turk, View-based Interpretation of Real-time Optical Flow for Gesture Recognition, in *Proceedings of 3rd International Conference on Automatic Face and Gesture Recognition (AFGR'98)*, pp. 416-421, 1998. [Cutler, 1998]
- Y. Dai and Y. Nakano, Face-Texture Model Based on SGLD and Its Application in Face Detection in a Color Scene, *Pattern Recognition*, Vol. 29, No. 6, pp.1007-1017, 1996. [Dai, 1996]
- T. Darrel, G. Gordon, M. Harville and J. J Woodfill, Integrated Person Tracking Using Stereo, Color, and Pattern Detection, *International Journal of Computer Vision*, Vol. 37, No. 2, pp. 175-185, 2000. [Darrel, 2000]
- T. Darrel and A. Pentland, Space-time Gestures, in *Proceedings of IEEE International Conference on Computer Vision and Pattern recognition (CVPR'93)*, pp. 335-340, 1993. [Darrel, 1993]
- J. W. Davis, Hierarchical Motion History Images for Recognizing Human Motion, in *Proceeding of IEEE Workshop on Detection and Recognition of Events in Video (EVENT'01)*, pp.39-46, 2001. [Davis, 2001]
- S. S. Fels, and G. E. Hinton, Glove-Talk: A neural Network Interface Between a Data-Glove and Speech Synthesizer, *IEEE Transactions on Neural Networks*, Vol. 4, pp. 2-8, 1993. [Fels, 1993]
- The Festival Speech Synthesis System* developed by CSTR, University of Edinburgh, <http://www.cstr.ed.ac.uk/project/festival>. [Festival, 1999]

- T. Fong, I. Nourbakhsh and K. Dautenhahn, A Survey of Socially Interactive Robots, *Robotics and Autonomous System*, Vol. 42(3-4), pp.143-166, 2003. [Fong, 2003]
- W.T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, Computer Vision for Computer Games, in *Proceedings of International Conference on Automatic Face and Gesture Recognition (AFGR'96)*, pp. 100-105, 1996. [Freeman, 1996]
- M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, and H. Ueno: Gesture-Based Human-Robot Interaction Using a Knowledge-Based Software Platform, *International Journal of Industrial Robot*, Vol. 33(1), 2006. [Hasanuzzaman, 2005a]
- M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, H. Gotoda, Y. Shirai, and H. Ueno, Knowledge-Based Person-Centric Human-Robot Interaction by Means of Gestures, *International Journal of Information Technology*, Vol. 4(4), pp. 496-507, 2005. [Hasanuzzaman, 2005b]
- M. Hasanuzzaman, V. Ampornaramveth, T. Zhang, M. A. Bhuiyan, Y. Shirai, H. Ueno, Real-time Vision-based Gesture Recognition for Human-Robot Interaction, in *Proceeding of IEEE International Conference on Robotics and Biomimetics (ROBIO'2004)*, China, pp. 379-384, 2004. [Hasanuzzaman, 2004a]
- M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, M.A. Bhuiyan, Y. Shirai, H. Ueno, Gesture Recognition for Human-Robot Interaction Through a Knowledge Based Software Platform, in *Proceeding of IEEE International Conference on Image Analysis and Recognition (ICIAR 2004)*, LNCS 3211 (Springer-Verlag Berlin Heidelberg), Vol. 1, pp. 5300-537, Portugal, 2004. [Hasanuzzaman, 2004b]
- M. Hasanuzzaman, V. Ampornaramveth, T. Zhang, M.A. Bhuiyan, Y. Shirai, H. Ueno, Face and Gesture Recognition Using Subspace Method for Human-Robot Interaction, *Advances in Multimedia Information Processing - PCM 2004: in Proceeding of 5th Pacific Rim Conference on Multimedia*, LNCS 3331 (Springer-Verlag Berlin Heidelberg) Vol. 1, pp. 369-376, Tokyo, Japan, 2004. [Hasanuzzaman, 2004c]
- M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, P. Kiatisevi, Y. Shirai, H. Ueno, Gesture-based Human-Robot Interaction Using a Frame-based Software Platform, in *Proceeding of IEEE International Conference on Systems Man and Cybernetics (IEEE SMC'2004)*, Netherland, 2004. [Hasanuzzaman, 2004d]
- [M. Hasanuzzaman, M.A. Bhuiyan, V. Ampornaramveth, T. Zhang, Y. Shirai, H. Ueno, Hand Gesture Interpretation for Human-Robot Interaction, in *Proceeding of International Conference on Computer and Information Technology (ICCIT'2004)*, Bangladesh, pp. 149-154, 2004. Hasanuzzaman, 2004e]
- C. Hu, Gesture Recognition for Human-Machine Interface of Robot Teleoperation, in *Proceeding of International Conference on Intelligent Robots and Systems*, pp. 1560-1565, 2003. [Hu, 2003]
- [Huang, 1994] G. Yang, and T. S. Huang, Human Face Detection in Complex Background *Pattern Recognition*, Vol. 27, No. 1, pp. 53-63, 1994.
- Gary Imai Gestures: *Body Language and Nonverbal Communication*, <http://www.csupomona.edu/~tassi/gestures.htm>, visited on June 2004. [Imai, 2004]
- Robovie*, <http://www.mic.atr.co.jp/~michita/everyday-e/> [Imai, 2000]
- A. K. Jain, *Fundamental of Digital Image Processing*, Prentice-Hall of India Private Limited, New Delhi, 1995. [Jain, 1995]

- T. Kanade, *Computer Recognition of Human Faces*, Birkhauser Verlag, Basel and Stuttgart, ISR-47, pp. 1-106, 1977. [Kanade, 1977]
- S. Kawato and J. Ohya, Real-time Detection of Nodding and Head-Shaking by Directly Detecting and Tracking the 'Between-Eyes', in *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition (AFGR'2000)*, pp.40-45, 2000. [Kawato, 2000]
- M. D. Kelly, Visual Identification of People by Computer, *Technical report*, AI-130, Stanford AI projects, Stanford, CA, 1970. [Kelly, 1970]
- R. Kjeldsen, and K. Kender, Finding Skin in Color Images, in *Proceedings of 2nd International Conference on Automatic Face and Gesture Recognition (AFGR'96)*, pp. 312-317, 1996. [Kjeldsen, 1996]
- C. Kotropoulos and I. Pitas, Rule-based Face Detection in Frontal Views, in *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, Vol. 4, pp. 2537-2540, 1997. [Kotropoulos, 1997]
- J. Kramer, L. Larry Leifer, The Talking Glove: A Speaking Aid for Non-vocal Deaf and Deaf-blind Individuals, in *Proceedings of 12th Annual Conference, RESNA (Rehabilitation Engineering & Assistive Technology)*, pp. 471-472, 1989. [Kramer, 1989]
- S. J. Lee, S. B. Jung, J. W. Kwon, S. H. Hong, Face Detection and Recognition Using PCA, in *Proceedings of IEEE Region 10th Conference (TENCON'99)* pp. 84-87, 1999. [Lee, 1999]
- C. Lee, and Y. Xu, Online, Interactive Learning of Gestures for Human/Robot Interfaces, in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA'96)*, Vol. 4, pp. 2982-2987, 1996. [Lee, 1996]
- J. Lin, Y. Wu, and T. S Huang, Capturing Human Hand Motion in Image Sequences, in *Proceeding of Workshop on Motion and Video Computing*, Orlando, Florida, December, 2002. [Lin, 2002]
- X. Lu, Image Analysis for Face Recognition-A Brief Survey, *Personal notes*, pp. 1-37, 2003. [Lu, 2003]
- J. Miao, B. Yin, K. Wang, L. Shen, and X. Chen, A Hierarchical Multiscale and Multiangle System for Human Face Detection in a Complex Background Using Gravity-Centre Template, *Pattern Recognition*, Vol. 32, No. 7, pp. 1237-1248, 1999. [Miao, 1999]
- Application Wizard: Microsoft Foundation Class, VideoIn*, Microsoft Corp. [Microsoft]
- B. Moghaddam and A. Pentland, Probabilistic Visual Learning for Object Detection, in *Proceeding of 5th International Conference on Computer Vision*, pp. 786-793, 1995. [Moghaddam, 1995]
- Y. Nam and K. Y. Wohn, Recognition of Space-Time Hand-Gestures Using Hidden Markov Model, in *Proceedings of ACM Symposium on Virtual Reality Software and Technology*, pp. 51-58, 1996. [Nam, 1996]
- J. L. Nespoulous, P. Perron, and A. Roch Lecours, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986. [Nespoulous, 1986]
- R. O'Hagan, Finger Track-A Robust and Real-Time Gesture Interface, in *Proceeding of 10th Australian Joint Conference on Artificial Intelligence: Advance Topics in Artificial Intelligence*, LNCS, Vol. 1342, pp. 475-484, 1997. [O'Hagan, 1997]
- K. Oka, Y. Sato, and H. Koike, Real-Time Tracking of Multiple Finger-trips and Gesture Recognition for Augmented Desk Interface Systems, in *Proceeding of International*

- Conference in Automatic Face and Gesture Recognition (AFGR'02)*, pp. 423-428, Washington D.C, USA, 2002. [Oka, 2002]
- D. W. Patterson, *Introduction to Artificial Intelligence and Expert Systems*, Prentice-Hall Inc., Englewood Cliffs, N.J, USA, 1990. [Patterson, 1990]
- A. Pentland, B. Moghaddam, and T. Starner, View-based and Modular Eigenspaces for Face Recognition, in *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pp. 84-91, 1994. [Pentland, 1994]
- V. I. Pavlovic, R. Sharma and T. S. Huang, Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 19, No. 7, pp. 677-695, 1997. [Pavlovic, 1997]
- J. M. Rehg and T. Kanade, Digiteyes: Vision-based Hand Tracking for Human-Computer Interaction, in *Proceeding of Workshop on Motion of Non-Rigid and Articulated Bodies*, pp. 16-94, 1994. [Rehg, 1994]
- G. Rigoll, A. Kosmala, S. Eickeler, High Performance Real-Time Gesture Recognition Using Hidden Markov Models, in *Proceeding of International Gesture Workshop on Gesture and Sign Language in Human Computer Interaction*, pp. 69-80, Germany, 1997. [Rigoll, 1997]
- H. A. Rowley, S. Baluja and T. Kanade, Neural Network-Based Face Detection *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 23, No. 1, pp. 23-38, 1998. [Rowley, 1998]
- E. Saber and A. M. Tekalp, Frontal-view Face Detection and Facial Feature Extraction Using Color, Shape and Symmetry Based Cost Functions, *Pattern Recognition Letters*, Vol. 17(8) pp.669-680, 1998. [Saber, 1998]
- T. Sakai, M. Nagao and S. Fujibayashi, Line Extraction and Pattern Detection in a Photograph, *Pattern Recognition*, Vol. 1, pp.233-248, 1996. [Sakai, 1996]
- N. Shimada, and Y. Shirai, 3-D Hand Pose Estimation and Shape Model Refinement from a Monocular Image Sequence, in *Proceedings of VSMM'96 in GIFU*, pp.23-428, 1996. [Shimada, 1996]
- S. A. Sirohey, Human Face Segmentation and Identification, *Technical Report CS-TR-3176*, University of Maryland, pp. 1-33, 1993. [Sirohey, 1993]
- L. I. Smith, *A Tutorial on Principal Components Analysis*, February 26, 2002. [Smith, 2002]
- T. Starner, J. Weaver, and Alex Pentland, Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 20, No.12, pp. 1371-1375, 1998. [Starner, 1998]
- D.J. Sturman and D. Zetler, A Survey of Glove-Based Input, *IEEE Computer Graphics and Applications*, Vol. 14, pp-30-39, 1994. [Sturman, 1994]
- J. Triesch and C. V. Malsburg, Classification of Hand Postures Against Complex Backgrounds Using Elastic Graph Matching, *Image and Vision Computing*, Vol. 20, pp. 937-943, 2002. [Triesch, 2002]
- A. Tsukamoto, C.W. Lee, and S. Tsuji, Detection and Pose Estimation of Human Face with Synthesized Image Models, in *Proceeding of International Conference of Pattern Recognition*, pp. 754-757,1994. [Tsukamoto, 1994]
- C. Torras, Robot Adaptivity, *Robotics and Automation Systems*, Vol. 15, pp.11-23, 1995. [Torras, 1995]

- M. Turk and G. Robertson, Perceptual user Interfaces, *Communication of the ACM*, Vol. 43, No. 3, pp.32-34, 2000. [Turk, 2000]
- M. Turk and A. Pentland, Eigenface for Recognition, *Journal of Cognitive Neuroscience*, Vol. 3, No.1, pp. 71-86, 1991. [Turk, 1991]
- H. Ueno, Symbiotic Information System: Towards an Ideal Relationship of Human Beings and Information Systems, *Technical Report of IEICE*, KBSE2001-15: pp.27-34, 2001. [Ueno, 2001]
- A. Utsumi, N. Tetsutani and S. Igi, Hand Detection and Tracking Using Pixel Value Distribution Model for Multiple-Camera-Based Gesture Interactions, in *Proceeding of IEEE Workshop on Knowledge Media Networking (KMN'02)*, pp. 31-36, 2002. [Utsumi, 2002]
- S. Waldherr, R. Romero, S. Thrun, A Gesture Based Interface for Human-Robot Interaction, *Journal of Autonomous Robots*, Kluwer Academic Publishers, pp. 151-173, 2000. [Waldherr, 2000]
- T. Watanabe, M. Yachida, Real-time Gesture Recognition Using Eigenspace from Multi-Input Image Sequences, *System and Computers in Japan*, Vol. J81-D-II, pp. 810-821, 1998. [Watanabe, 1998]
- L. Wiskott, J. M. Fellous, N. Kruger, and C. V. Malsburg, Face Recognition by Elastic Bunch Graph Matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 19, No.7, pp. 775-779, 1997. [Wiskott, 1997]
- M. H. Yang, D. J. Kriegman and N. Ahuja, Detection Faces in Images: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 24, No. 1, pp. 34-58, 2002. [Yang, 2002]
- M. H. Yang, Hand Gesture Recognition and Face Detection in Images, *Ph.D Thesis*, University of Illinois, Urbana-Champaign, 2000. [Yang, 2000]
- J. Yang, R. Stiefelagen, U. Meier and A. Waibel, Visual Tracking for Multimodal Human Computer Interaction, in *Proceedings of ACM CHI'98 Human Factors in Computing Systems*, pp. 140-147, 1998. [Yang, 1998]
- G. Yang and T. S. Huang, Human Face Detection in Complex Background, *Pattern Recognition*, Vol. 27, No.1, pp.53-63, 1994. [Yang, 1994]
- A. Yuille, P. Hallinan and D. Cohen, Feature Extraction from Faces Using Deformable Templates, *International Journal of Computer Vision*, Vol. 8, No. 2, pp 99-111, 1992. [Yuille, 1992]
- W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, Face Recognition: A Literature Survey, *ACM Computing Surveys*, Vol. 35, No. 4, pp. 399-458, 2003. [Zhao, 2003]
- T. Zimmerman, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvil, A Hand Gesture Interface Device, in *Proceedings of (CHI+GI)'87*, pp. 189-192, 1987. [Zimmerman, 1987]