

ADAPTIVE VISUAL GESTURE RECOGNITION FOR HUMAN-ROBOT INTERACTION

*Mohammad Hasanuzzaman¹, Saifuddin Mohammad Tareeq¹, Tao Zhang²,
Vuthichai Ampornaramveth², Hironobu Gotoda², Yoshiaki Shirai³, Haruki Ueno²*

¹Department of Computer Science and Engineering,
University of Dhaka, Dhaka-1000, Bangladesh, Email: hzamancsdu@yahoo.com, smtareeq@univdhaka.edu

²National Institute of Informatics, Intelligent Systems Research Division, 2-1-2, Hitotsubashi,
Chiyoda-ku, Tokyo 101-8430, Japan, Email: gotoda@nii.ac.jp, ueno@nii.ac.jp

³Department of Human and Computer Intelligence, School of Information Science and Engineering,
Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan.

ABSTRACT

This paper presents an adaptive visual gesture recognition method for human-robot interaction using a knowledge-based software platform. The system is capable of recognizing users, static gestures comprised of the face and hand poses, and dynamic gestures of face in motion. The system learns new users, poses using multi-cluster approach, and combines computer vision and knowledge-based approaches in order to adapt to new users, gestures and robot behaviors. In the proposed method, a frame-based knowledge model is defined for the person-centric gesture interpretation and human-robot interaction. It is implemented using the Frame-based Software Platform for Agent and Knowledge Management (SPAK). The effectiveness of this method has been demonstrated by an experimental human-robot interaction system using a humanoid robot, namely, 'Robovie'.

Keywords: Adaptive visual gesture recognition, human-robot interaction, multi-cluster based learning, SPAK.

1.0 INTRODUCTION

Many researchers are working on a natural human robot interaction system due to the demand from the welfare services of many countries. Ueno proposed a concept of Symbiotic Information Systems (SIS) as well as symbiotic robotics system as one application of SIS, where humans and robots can communicate with each other in human ways using speech and gesture [1]. Multimodal user interfaces are a strong candidate for building natural user interfaces. In multimodal approaches, user can include simple keyboard and mouse with advance perception techniques like speech recognition and computer vision (gestures, gaze, etc.) as user machine interface tools. With this motivation automatic speech and gesture recognition are the topics of research for the last few decades and trying to reach human-human communication modalities into human-machine interaction.

Although there is no doubt that the fusion of gesture and speech allows more natural human-robot interaction, single modality gesture recognition can be considered more reliable than speech recognition systems as human voice varies from person to person and the system needs to take care of large number of data set when recognizing speech [2]. Gestures are expressive, meaningful body motions such as physical movements of head, face, fingers, hands or body with the intention to convey information or interact with the environment. Hand and face poses are more rigid, though it also varies little from person to person. Human hand and facial gestures (or emotions) are the means of non-verbal interaction among human. Though hands make most human gestures, same hand gestures may have different meaning in different culture. For example, the thumb and index finger are joined together to form an "O" to denote "Ok" in the United States, in the Latin America and France this gesture is a rude sign, in Brazil and Germany this gesture is obscene, and in Japan this gesture means money [3]. Thus, the interpretation of recognized gesture is user-dependent. As the skin colors of the hand region and hand shapes are also different for different persons, the recognition process itself is also user-dependent. As a result, person identification and adaptation is one of the prime factors in order to realize reliable gesture recognition and interpretation. This paper concentrates on adaptive visual face and gestures recognition in symbiotic robot system, which is an application of SIS. It is the capability of self-modification that some agents have, which allows them to maintain a level of performance in front of environmental changes, or to improve it when confronted repeatedly with the same situation [4]. Gesture-based human-robot natural interaction system could be designed so that it can understand different users, their gestures, meaning of the gestures and the robot behaviors.

There are significant amount of research on hand, arm and facial gesture recognition to control robot or intelligent machine in recent years. Pavlovic et al. [5] have made a good review on recent research on visual hand gesture

recognition systems, and Sturuman et al. [6] have summarized on gloved-based interface devices. Waldherr et al. have proposed gesture-based interface for human and service robot interaction [7]. 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 method but did not consider user or hand pose adaptation. Kortenkamp et al. have developed gesture-based human-mobile robot interface [8]. They have used static arm poses as gestures. Torras proposed robot adaptivity technique using neural learning algorithm [4]. This method is computationally inexpensive and there is no way to encode prior knowledge about the environment to gain the efficiency. Bhuiyan et al. detected and tracked face and eyes for human-robot interaction [9]. But only the largest skin-like region has been considered for detecting the probable face area, which may not be true when two hands are present in the image. However, all of the above papers focus primarily on visual processing and do not maintain knowledge of different users nor consider how to deal with them. In our previous research, we have combined computer vision and knowledge-based approaches for gesture-based human-robot interaction (HRI) [10]. In that research we have utilized pose specific Subspace (separate eigenspaces for each pose) method for face and hand poses classification, and segmented three larger skin like components from the images assuming that two hands and face may present in the image at the same time. However, in that system we did not consider new users, hand poses or gestures adaptation methods. It is essential for the system to cope with the different users. A new user should be included using on-line registration process. When a user is included the user may wants to perform new gesture that is ever been used by other persons or himself/herself. In that case, the system should include the new hand poses or gestures with minimum user interaction.

2.0 SYSTEM ARCHITECHTURE

Fig. 1 shows the overall architecture of our adaptive visual gesture-based HRI system using SPAK [11]. This system integrates various kinds of hardware and software components such as vision-based face and gesture recognizer and learner, text to speech converter, robots or robotic devices (robot arms, robot legs, robot neck and robot mouth, etc.) and a knowledge-based software platform. The system first detects human face using multiple features [10] and recognizes the face using eigenface method [12]. Then, using the knowledge of the identified person, face and hand poses are classified and gestures are recognized from the later image frames. The user profile consists of the threshold values for chrominance and luminance components of the skin colors of each known person. Images of face and hand poses are segmented using these color information and classified using the multi-cluster based pattern-matching approach. The system is capable of learning new users and new hand poses using multi-cluster based incremental learning method and registers new users and new poses in the knowledge base using minimal user interaction in online manner.

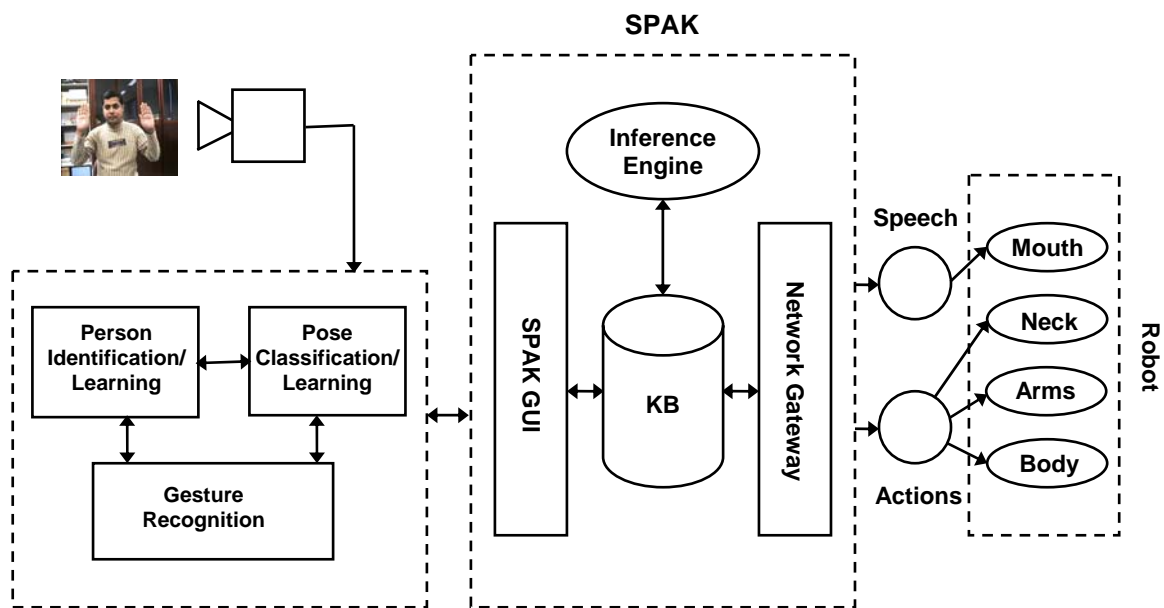


Fig. 1: Architecture of adaptive visual gesture-based HRI system using SPAK

In this system, face and gesture recognition method for person-centric human-robot interaction (HRI) is based on a knowledge-based software platform called SPAK. After hand and face poses are classified, the static gestures are recognized using frame-based approach [10, 13]. Known gestures are defined as frames in SPAK knowledge base. When the required combination of the pose components is found the corresponding gesture frame will be activated. Considering the transitions of the face poses in a sequence of time steps recognizes the dynamic gestures. After the gesture is recognized, the interaction between human and robot is determined by the knowledge also modeled as frame hierarchy in SPAK. Using the received gesture and user information, SPAK processes the facts and activates the corresponding behavior (action) frames to carry out predefined robot actions, which may include body movement and speech.

3.0 ADAPTIVE VISUAL GESTURE RECOGNITION

Before a system can recognize the face and hand poses, it must possess knowledge of the characteristic feature of these poses. This means that the system designer must either build the necessary discriminating rules into the system or the system must learn them. To adapt to new users and new hand poses the system must be able to perceive and extract relevant properties from the unknown faces and hand poses, find common patterns among them and formulate discrimination criteria consistent with the goals of the recognition process. This form of learning is known as clustering and it is the first steps in any recognition process where discriminating features of the objects are not known in advance [14]. Due to the orientation and illumination variation, same hand poses may be composed of multiple clusters and same person may include different face clusters. Considering these situations, we propose multi-cluster based learning approach. A pose P_i may include number of clusters and each cluster C_j may include number of images $\{X_1, X_2, \dots, X_o\}$ as a member of that cluster. This multi-clustering method is described using following steps:

Step 1: Generate eigenvectors [12] from training images that includes all the known poses.

Step 2: Select m -number of eigenvectors according to higher eigenvalues, these are known as principal components.

Step 3: Read the initial cluster image database (initialize with the known cluster images) and cluster information table that's hold the starting pointer of each cluster. Project each image onto the eigenspaces and form feature vectors using equation (1) and (2).

$$\omega_i^j = (u_m)^T (X_j) \quad (1)$$

$$\Omega_j = [\omega_1^j, \omega_2^j, \dots, \omega_k^j] \quad (2)$$

Where, (u_m) is the m^{th} eigenvector, X_j is the j^{th} image (60×60) in the cluster database.

Step 4: Read the unlabeled images those should be clustered or labeled.

a) Project each unlabeled image onto the eigenimages and form feature vectors (Ω) using equation (1) and (2).

b) Calculate Euclidean distance to each eigenimage in the known dataset (cluster) using equation (3) and (4),

$$\varepsilon_j = \|\Omega - \Omega_j\| \quad (3)$$

$$\varepsilon = \arg \min \{\varepsilon_j\} \quad (4)$$

Step 5: Find the nearest class,

a) If $(T_i \leq \varepsilon \leq T_c)$ then add the image in the neighbor cluster; increment the insertion parameter where T_i is the threshold for identification and T_c is the threshold for new cluster.

b) If $(\varepsilon < T_i)$, then the image is recognizable and no need to include it in the cluster database.

Step 6: If the insertion rate in the known cluster is greater than zero, then update the cluster information table that's holds the starting pointer of all clusters.

Step 7: Repeat the step 3 to 6 until the insertion rate (α) in the known cluster dataset is zero (0).

Step 8: If insertion rate is zero, then check the unlabeled dataset, which follows the condition $(T_c < \varepsilon \leq T_f)$. Where, T_f is the threshold that defines for discarding the image.

Step 9: If maximum number of unlabeled data (for a class) $> N$ (predefined), then select one image (based on minimum Euclidean distance) as a member of the new cluster. Then update the cluster information table.

Step 10: Repeat from step 3 to 9 until the numbers of unlabeled data is less than N .

Step 11: If height of the cluster (number of member images in the cluster) is $>L$, then add it as a permanent cluster.

Step 12: After clustering the user defines the associations of the clusters in the knowledge base. Each pose may be associated with multiple clusters. For undefined cluster there is no association link.

3.1 User Identification and Learning

We have already argued that robot should be able to recognize and remember the users and learn about them. A number of techniques have been developed to detect and recognize faces [15]. This gesture-based HRI system is person centric, therefore, person-identification and adaptation is one of the attraction of this system. If the new user comes in front of the robot eye's camera or system camera, the system identifies the user as unknown and asks for registration.

The face is first detected from the cluttered background using multiple feature-based approaches [10]. 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 [16]. 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 [12], whether it belongs to known person or unknown person. The eigenvectors are calculated from the known persons face images for all face classes and m-number of eigenvectors corresponding to the highest eigenvalues are chosen to form principal components. 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. If the minimal Euclidian is less than the predefined threshold value then person is known, otherwise unknown [17]. For unknown person, based on judge function learning process will be activated and the system will learn new user using multi-clustering approach. The judge function is based on the ratio of the number of unknown faces to total number of detected faces for a specific time slot. The learning function develops new clusters corresponding to new person. The user defines the person name and skin color information in the user profile knowledge base and associates with the corresponding cluster. For known user, person-centric skin color information (Y, I, Q components) is used to reduce the computational cost.

3.2 Pose Classification and Learning Method

In real life, we can understand several hand and facial gestures on the spot by using other modalities or context or scene analysis without having prior knowledge of the new gestures. For the machine it is difficult to understand the new poses without prior knowledge. It is essential to learn new poses based on judge function or predefined knowledge. The judge function determines the user intention, i.e., intention to create new gesture. The judge function is based on the ratio of the number of unknown poses to total number poses for a specific time slots. For example, the user shows same hand pose for 10 image frame times that are unknown to the system, that means he wants to use it as a new gesture. In this situation, based on judge function learning function activates and the system learns new user using multi-clustering approach. The learning function develops new cluster/clusters corresponding to new pose. The user defines the pose name in the knowledge base and associates with the corresponding clusters. If the pose is identified then corresponding pose frame will be activated [13].

3.3 Recognizing and Learning Gesture

Gesture recognition is the process by which gestures made by the user are identified in the system. There are static gesture and dynamic gesture. The recognition of gesture is carried out in two phases. In the first phase, face and hand poses are classified from the each captured image frame using the method described in previous section. Then sequence of poses and combination of poses are analyzed to identify the occurrence of gesture. Interpretation of identified gesture is user-dependent since the meaning of the gesture may differ from person to person based on their culture. For example, when user 'Hasan' comes in front of 'Robovie' eyes, 'Robovie' recognizes the person as 'Hasan' and says "Hi Hasan! How are you?", then 'Hasan' raises his 'Thumb up' and 'Robovie' replies to 'Hasan' "Oh! You are not fine today". In the similar situation, for another user 'Cho', 'Robovie' says, "Hi, You are fine today". That means 'Robovie' should understand the person-centric meaning of gesture. To accommodate different user's desires, our person-centric gesture interpretation is achieved using frame-based knowledge representation approach [10, 15]. The user predefines these frames into the knowledge base. The system maintains frames with necessary attributes (gesture components, gesture name) for all predefined gestures. Our current system recognizes 13 gestures: 11 static gestures and 2 dynamic facial gestures. These are: 'TwoHand' (raise left hand and right hand palms), 'LeftHand' (raise left hand palm), 'RightHand' (raise right hand palm), 'One' (raise index finger), 'Two'

(form V sign using index and middle fingers), ‘Three’ (raise index, middle and ring fingers), ‘ThumbUp’ (thumb up), ‘Ok’ (make circle using thumb and index finger), ‘FistUp’ (fist up), ‘PointLeft’ (point left by index finger), ‘PointRight’ (point right by index finger), ‘YES’ (nods face up and down or down and up), ‘NO’ (shakes face left and right or right and left). It is possible to recognize more gestures including new poses and new rules for the gestures using this system. New poses can be included in the training image database using learning method and corresponding frames can be defined in the knowledge base to interpret the gesture. To teach the robot a new poses, the user should perform the poses several times (e.g. 10 image frame times). Then the learning method detects it as a new pose and creates cluster/clusters for that pose. Sequentially, it updates the knowledge base for the cluster information.

3.3.1 Static Gesture Recognition

Static gestures are recognized using frame-based approach with the combination of the pose classification results of three skin-like regions at a particular time. For example, if left hand palm, right hand palm and one face are present in the input image then it recognizes as “TwoHand” gesture and corresponding gesture frame will be activated. The user predefines these frames into knowledge base using SPAK knowledge editor. The system maintains frames with necessary attributes (gesture components, gesture name) for all predefined gestures. Gesture components are the face and hand poses. If the pose is identified then pose name is fed to SPAK and corresponding instance frame of the pose-frame will be activated. The gesture frame is defined using three slots corresponding to pose recognition results of three skin-regions. Fig. 2 shows the contents of “TwoHand” gesture frame in SAPK. If image analysis and recognition module classifies the ‘FACE’, ‘LEFTHAND’ and ‘RIGHTHAND’ poses at an image frame it sends the pose names (face, lefthand, righthand, etc.) to the SPAK knowledge module. According to pose names corresponding pose frames (‘FACE’, ‘LEFTHAND’ and ‘RIGHTHAND’) will be activated. In this combination “TwoHand” gesture is recognized and corresponding frame will be activated.

Name	Type	Value	Condition	Argument	Required	Shared	Unique
Name	String	TwoHand	ANY		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
mFace	Instance		ANY	FACE	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mRightHand	Instance		ANY	RIHGTHAND	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mLeftHand	Instance		ANY	LEFTHAND	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Fig. 2: Example frame for the gesture “TwoHand”

3.3.2 Dynamic Gesture Recognition

Dynamic gesture is the gesture that uses motion of hand or body to emphasize or help to express a thought or feelings such as “NO” (shake face left-right), “YES” (shakes face up-down). It is difficult to visually recognize dynamic gesture due to large variations in the speed of position change of the physical objects that describe the gesture. This system recognizes two dynamic facial gestures considering the transition of the face poses in a sequence of time steps. If user face shakes left to right or right to left, then it is recognized as “NO” gesture. If user face shakes up and down or down and up, then it is recognized as “YES” gesture.



Fig. 3: Example of face sequences for dynamic gesture ‘YES’ and ‘NO’.

This method uses a 3-layers queue (FIFO) that holds the last three results (different poses) of the detected face poses. This method defines five specific face poses: frontal face (NF), right-rotated face (RF), left-rotated face (LF), up position face (UF) and down position face (DF). For every image frame, face pose is classified using pose classification method. If pose is classified as predefined face pose then it is added to the 3-layer queue. If the classified pose value is same as previous frame, then queue values will remain be unchanged. From the combination

of 3-layers queue values this method determines the gesture. For example, if the queue values are {UF, NF, DF} or {DF, NF, UF} pose sets then it is recognized as “YES” gesture. Similarly, if the queues values are {RF, NF, LF} or {LF, NF, RF} pose sets then it is recognized as “NO” gesture. After a specific time period the queue values are refreshed. Fig. 3 shows the example pose sequences for dynamic gestures “YES” and “NO”. If gesture is recognized then corresponding gesture frame will be activated.

4.0 EXPERIMENTAL RESULT AND DISCUSSION

This system uses a standard video camera or ‘Robovie’ eye’s camera for data acquisition. Each captured image is digitized into a matrix of 320×240 pixels with 24-bit color. The recognition approach has been tested with real world human-robot interaction system using a humanoid robot, namely, ‘Robovie’ developed by ATR [18]. This recognition and learning approach is verified using real-time input images as well as static images.

4.1 Results of User Recognition and Learning

Seven individuals were asked to act for the predefined face poses in front of the camera and the sequence of face images were saved as individual image frame. All the training and test faces are 60×60 pixels gray images. The learning algorithm is verified for 7 persons frontal face or normal face (NF) images and five directional face images (normal face, left directed face, right directed face, up directed face, down directed face). Fig. 4 shows the sample outputs of the learning method for normal faces. In the first step, the system is trained using 60 normal face images of three persons and developed three clusters (top 3 rows in Fig. 4) corresponding to three persons. The cluster information table (that’s store the starting pointer of each cluster) contents are {1, 11, 23, 30}. If any input face image matches with the known member between 1 and 10 then the person is identified as person_1 (‘Hasan’). In the second step, 20 face image sequences of another person are fed to the system as input. The minimum Euclidian distances (ED) from three known persons face images are shown using upper line graph (B_adap) in Fig. 6. The system identifies these faces as unknown person based on threshold values (Euclidian distance for the known/unknown classification) and activates the user learning function. The learning function develops new clusters (4th row of Fig. 4) and updates the cluster information table {1, 11, 23, 30, 38}. After adaptation, the minimum Euclidian distance distribution line (A_adap line in Fig. 6) shows that for 8 images minimum ED is zero and those are included in the new cluster so that the system can recognize the person. This method is tested for 7 persons including 2 females and formed different clusters with different length (number of images per cluster) for different persons as shown in Fig. 4.


Cluster	Members of the cluster	Associations
NFC ₁ (1)		Person_1 ‘Hasan’
NFC ₂ (11)		Person_2 ‘Huda’
NFC ₃ (23)		Person_3 ‘Yumiko’
NFC ₄ (30)		Person_4 ‘Cho’
NFC ₅ (38)		Person_5 ‘Osani’
NFC ₆ (46)		Person_6 ‘Yosida’
NFC ₆ (51-62)		Person_7 ‘Satomi’

Fig. 4: Sample outputs of the learning method for frontal faces

Cluster Table	Members of the clusters	Associations
FC ₁ (1)		Person_1 'Hasan'
FC ₂ (12)		
FC ₃ (17)		
FC ₄ (27)		
FC ₅ (44-53)		
FC ₆ (53)		Person 2 'Cho'
FC ₇ (68)		
FC ₈ (82)		
FC ₉ (86)		
FC ₁₀ (96)		
FC ₁₁ (106-117)		

Fig. 5: Sample outputs of learning method for five directional faces

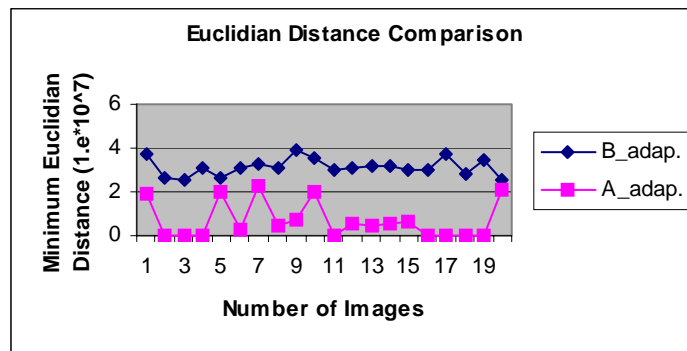


Fig. 6: Euclidian distances: before and after adaptation

The user learning system is also tested for five directional face images of 7 persons for 700 test face images. Fig. 5 shows five clusters for person_1 'Hasan' and 6 clusters for person_2 'Cho'. Similarly we get 7 Clusters for person_3 'Osani', 7 clusters for person_4 'Satomi' (female), 7 clusters for person_5 'Huda', 4 clusters for person_6 'Yosida', 5 Clusters for person_7 'Yomiko' (female). Fig. 7 shows the sample errors in clustering process. In cluster 26, up directed faces of person_6 and person_5 are overlapped (7 (a)). In cluster 31, up directed faces of person_5 and normal faces of person_6 are overlapped (7 (b)). This problem can be solved by using narrow threshold, but in this case number of iteration as well as discarding rates of the images will be increased. In our previous research, we have found that the accuracy of frontal face recognition is better than up, down and left-right directed faces [15]. This system prefers frontal and a little left-right rotated face for person identification. We have tested this face recognition method with 680 faces of 7 persons, where two of them are female. The average precision for face recognition is about 93% and recall rate is about 94.08%.

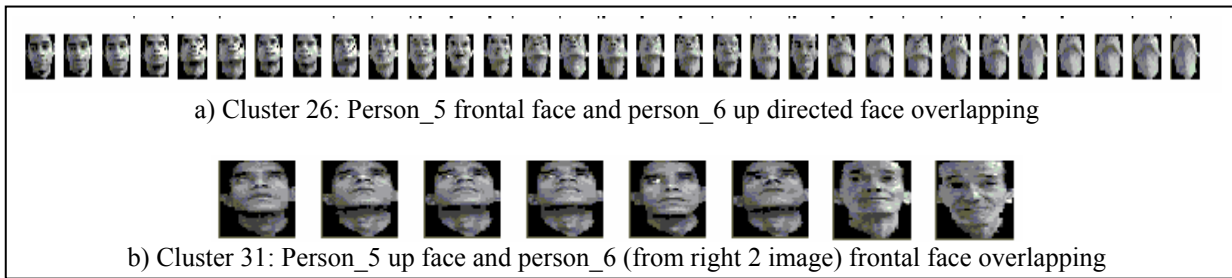


Fig. 7: Example of errors in clustering process

4.2 Results of Pose Classification and Learning

The system uses 10 hand poses of 7 persons for evaluating pose classification and learning method. All the training and test images are 60 x 60 pixels gray images. This method can automatically cluster the training images. This system is first trained using 200 images of 10-hand poses of person_1 (20 images of each pose). It automatically clusters the images into 13 clusters. Fig. 8 shows the sample outputs of hand poses learning method for person_1 ('Hasan'). If the user uses two hands to make the same pose then it forms two different clusters for the same pose. Different clusters can also be formed for the variation of orientation even the pose is same. If the person is changed, then it may form different clusters for the same hand poses (gestures) due to the variation of hand shapes and colors. After trained with 10-hand poses of person_1, 200 images of 10-hand poses of person_2 are feed to the system. The system develops 8 more clusters for the person_2 corresponding to 8 hand poses. For, the 'LEFTHAND' and 'RIGHTHAND' palms it did not develop new clusters, rather inserted new members in those clusters.

Cluster	Cluster Members	Associated Pose
PC1		ONE
PC2		FIST UP
PC3		FIST UP
PC4		OK
PC5		TWO
PC6		TWO
PC7		THREE
PC8		LEFT HAND
PC9		RIGHT HAND
PC10		THUMB UP
PC11		THUMB UP
PC12		POINT LEFT
PC13		POINT RIGHT

Fig. 8: Sample outputs of multi-clustering approach for pose classification

The pose learning system is also tested using 14-American sign language (ASL) characters (A-G, I, K, L, P, V, W, Y) [19]. Fig. 9 depicts the graphical representations of 14-ASL character classification accuracy using learning approach (after adaptation) and without learning approach (before adaptation). The comparison curves show that the system performs better with adaptation.

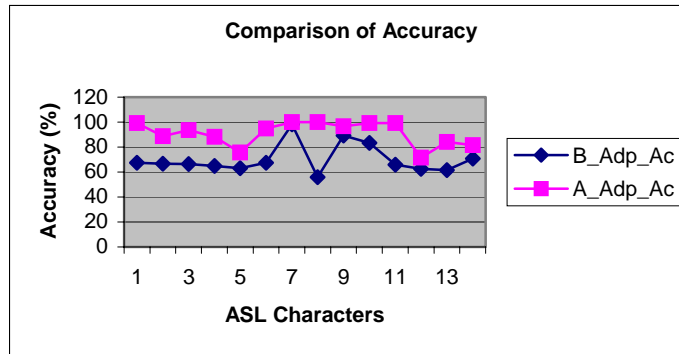


Fig. 9: Comparison of pose classification accuracy

4.3 Comparative study

It is very difficult to compare the performance of this system with other systems because the training and test images as well as scenarios are different. In the case of face and pose classification, accuracy of this method is better than general PCA method as shown in Fig. 9. Multi-clustered based learning feature provide better accuracy than general PCA method but in the same time it makes the system slower due to the inclusion of new training images. This system is capable of online learning with minor user acknowledgement. Integration of knowledge and vision provide user specific gesture and robot action mapping facilities which is not introduced in other related researches.

5.0 IMPLEMENTATION OF HUMAN-ROBOT INTERACTION

The real-time gesture based human-robot interaction is implemented as an application of this system. This approach has been implemented on a humanoid robot, namely, 'Robovie'. Since the same gestures can mean different tasks for different persons, we need to maintain the gesture with person-to-task knowledge. The robot and the gesture recognition PC are connected to SPAK knowledge server [10, 15]. From the image analysis and recognition PC, person identity and pose names (or gesture name for dynamic gesture) are sent to the SPAK for decisions making and the robot activation. According to gesture and user identity, the knowledge module generates executable codes for robot actions. The robot then follows speech and body action commands. This method has been implemented on a humanoid robot, namely, 'Robovie' for the following scenario:

<p>User: "Hasan" comes in front of Robovie eyes camera and robot recognizes the user as "Hasan".</p> <p>Robot: "Hi Hasan, How are you?" (speech)</p> <p>Hasan: uses the gesture "Ok"</p> <p>Robot: " Oh Good! Do you want to play now?" (speech)</p> <p>Hasan: uses the gesture "YES" (nods face)</p> <p>Robot: "Oh Thanks" (speech)</p> <p>Hasan: uses the gesture "TwoHand"</p> <p>Robot: imitates user's gesture "Raise Two Arms" as shown in Fig. 10.</p> <p>Hasan: uses the gesture "FistUp" (stop the interaction)</p> <p>Robot: Bye-bye (speech).</p>	<p>User: "Cho" comes in front of Robovie eyes camera, robot detect the face as unknown,</p> <p>Robot: "Hi, What is your Name?" (speech)</p> <p>Cho: Types his name "Cho"</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 "RightHand"</p> <p>Robot: imitate user's gesture ("Raise Right Arm")</p> <p>Cho: uses the gesture "Three"</p> <p>Robot: This is three (speech)</p> <p>Cho: uses the gesture "TwoHand"</p> <p>Robot: Bye-bye (speech)</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

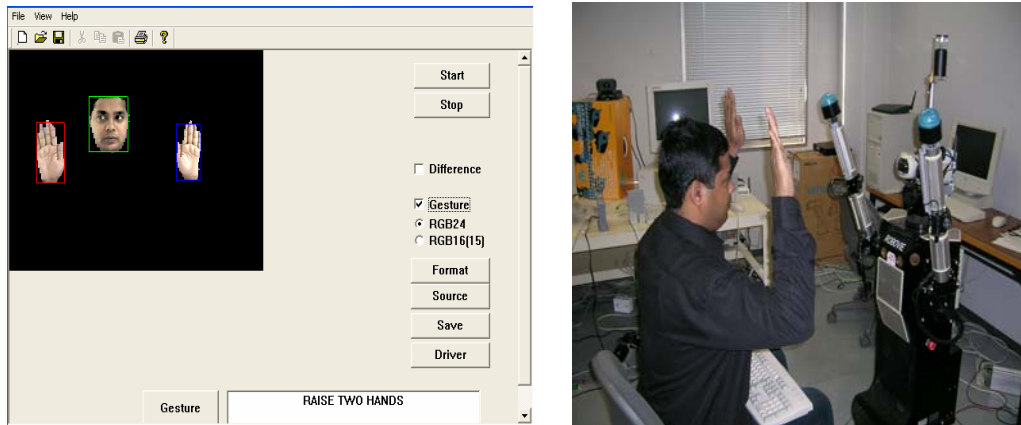


Fig. 10: Example human-robot (Robovie) interaction

The above scenario shows that the same gesture can be used to represent different meanings and several gestures can be used to denote the same meaning for different persons. A user can design new actions according to his/her desires using 'Robovie' and can design corresponding knowledge frames using SPAK to implement their desired actions.

6.0 CONCLUSIONS

This paper describes adaptive visual gesture recognition system for human-robot interaction using a knowledge-based software platform. This system is able to identify and learn new users, poses, gestures and robot behaviors. In this system the user can define or update the rules or conditions for gesture recognition/interpretation, and the robot behaviors corresponding to his/her gestures. This paper presents a multi-cluster based interactive learning approach for adapting new user and pose. However, if a large number of users use a large number of hand poses it is impossible to run this systems in real time. To overcome this problem, in future we should maintain person-specific subspaces (individual PCA for each person of all hand poses) for pose classification and learning. By integrating with knowledge-based software platform, gesture-based person-centric human-robot interaction system has also been successfully implemented using 'Robovie'. The future aim is to make the system more robust, dynamically adaptable to new users and new gestures for interaction with different robots such as 'Aibo', 'Robovie', 'Scout', etc. Ultimate goal of this research is to establish a human-robot symbiotic society so that they can share their resources and work cooperatively with human beings.

REFERENCES

- [1] H. Ueno, "A Knowledge-Based Information Modeling for Autonomous Humanoid Service Robot", *IEICE Transactions on Information & Systems*, Vol. E85-D No. 4, 2002, pp. 657-665.
- [2] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A Survey of Socially Interactive Robots", *Robotics and Autonomous Systems*, Vol. 42(3-4), 2003, pp. 143-166.
- [3] R. E. Axtell, "The meaning of hand gestures", http://www.allsands.com/Religious/NewAge/handgestures_wca_gn.htm, 1990.
- [4] C. Torras, "Robot Adaptivity", *Robotics and Autonomous Systems*, Vol. 15, 1995, pp.11-23.
- [5] 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, 1997, pp. 677-695.
- [6] D. J. Sturman and D. Zetler, "A Survey of Glove-Based Input", *IEEE Computer Graphics and Applications*, Vol. 14, 1994, pp-30-39.
- [7] S. Waldherr, R. Romero, S. Thrun, "A gesture Based Interface for Human-Robot Interaction", *Journal of Autonomous Robots*, Kluwer Academic Publishers, 2000, pp. 151-173.

- [8] D. Kortenkamp, E. Hubber, and P. Bonasso, "Recognizing and Interpreting Gestures on a Mobile robot", in: Proceedings of AAAI'96, 1996, pp. 915-921.
- [9] 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, 2004, pp. 42-54.
- [10] M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, H. Gotoda, Y. Shirai, H. Ueno, "Knowledge-based Person-centric Human-Robot Interaction by Means of Gestures", *Information Technology Journal*, Vol. 4, No. 4, 2005, pp. 496-507.
- [11] V. Ampornaramveth, H. Ueno, "SPAK: Software Platform for Agents and Knowledge Systems in Symbiotic Robots", *IEICE Transactions on Information and systems*, Vol. E86-D, No.3, 2004, pp 1-10.
- [12] M. Turk and A. Pentland, "Eigenface for Recognition", *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp. 71-86.
- [13] 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*, 2005.
- [14] D. W. Patterson, "*Introduction to Artificial Intelligence and Expert Systems*", Prentice-Hall Inc., Englewood Cliffs, N.J, USA, 1990.
- [15] 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, 2002, pp. 34-58.
- [16] M. Hasanuzzaman, V. Ampornaramveth, T. Zhang, M.A. Bhuiyan, Y. Shirai, H. Ueno, "Real-time Vision-based Gesture Recognition for Human-Robot Interaction", in: *Proceedings of the IEEE International conference on Robotics and Biomimetics (ROBIO'2004)*, China, 2004, pp. 379-384.
- [17] M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, M.A. Bhuiyan, Y. Shirai, H. Ueno, "Face and Gesture Recognition Using Subspace Method for Human-Robot Interaction", in: *Proceedings of the PCM'2004 (5th Pacific Rim Conference on Multimedia)*, Tokyo, Japan, LNCS, Vol. 3331, No. 1, Springer-Verlag Berlin Heidelberg, 2004, pp. 369-376.
- [18] T. Kanda, H. Ishiguro, M. Imai, T. Ono and K. Mase, "A Constructive Approach for Developing Interactive Humanoid Robots", in: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002, pp. 1265-1270.
- [19] "American Sign Language", <http://commtechlab.msu.edu/sites/aslweb/browser.htm> visited on April 2004.

BIOGRAPHY

Mohammad Hasanuzzaman, graduated (with Honors) in 1993 from the department of Applied Physics and Electronics, University of Dhaka, Bangladesh. He received Masters of Science (M.Sc.) in Computer Science in 1994 from the same University. He received Ph.D from the Department of Informatics, National Institute of Informatics (NII), The Graduate University for Advanced Studies, Tokyo, Japan in 2006. He has joined as a lecturer in the Department of Computer Science and Engineering, University of Dhaka in 2000. Since April 2006 he has been serving as an Assistant Professor in the Department of Computer Science and Engineering, University of Dhaka. His research interests include human-robot interaction, computer vision, image processing and artificial intelligence.

Saifuddin Mohammad Tareeq, graduated (with Honors) in 1993 from the department of Applied Physics and Electronics, University of Dhaka, Bangladesh. He received Masters of Science (M.Sc.) in Computer Science in 1994 from the same University. He received M. Sc. from the School of Computer Science and Software Engineering of the University of Western Australia, Australia in 2004. He has joined as a lecturer in the Department of Computer Science and Engineering, University of Dhaka in 1999. Since July 2004 he has been serving as an Assistant Professor in the Department of Computer Science and Engineering, University of Dhaka. His research interests include computer vision, image processing, artificial intelligence and bioinformatics.

Tao Zhang, received B.S, M.S and Ph.D. degrees in Electrical Engineering from Tsinghua University of China, Beijing, in 1993, 1995 and 1999 respectively and the Ph.D. degree in Electrical Engineering from the Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, Saga, Japan, in 2002. From 2002 to 2003 he was an Associate Professor of Electrical Engineering in the Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, Saga, Japan. Since 2003, he has been a Research Scientist in the Intelligent Systems Research Division, National Institute

of Informatics, Tokyo, Japan. His current research interest includes robotics, nonlinear system control, neural networks and biomedical engineering. He is a member of IEEJ, IEEE, and RSJ.

Vuthichai Ampornaramveth, received his B.Eng degree (with honors) in Electrical Engineering from Chulalongkorn University, Thailand in 1992, M.Eng and D.Eng in Control and Systems Engineering from Tokyo Institute of Technology in 1995, and 1999 respectively. His research interests include knowledge management for symbiotic robots, web and mobile applications, and distance learning systems. He is a member of IEEE, and RSJ.

Hironobu Gotoda, received the B.Sc., M.Sc. and D.Sc. from the Faculty of Science and Graduate School of Science of the University of Tokyo, Japan in 1989, 1991 and 1994 respectively. Currently he is an Associate Professor, Information Use Research, Human and Social Information Research Division, National Institute of Informatics, The Graduate University for Advanced Studies (Sokendai), Japan. His main area of research is representation and recognition of 3D objects. His current research topics include image-based modeling of deformable objects, and similarity of 3D geometries.

Yoshiaki Shirai, received the bachelor's degree from Nagoya University in 1964, the master's and the doctor's degree from the University of Tokyo in 1966 and 1969, respectively, all in mechanical engineering. In 1969 he joined the Electrotechnical Laboratory. He was a visiting researcher at the M.I.T. Artificial Intelligence Laboratory from 1971 to 1972. He was a Professor of the Dept. of Computer-Controlled Mechanical Systems, Graduate School of Engineering, Osaka University, Japan. Currently he is a professor at the Department of Human and Computer Intelligence, School of Information Science and Engineering, Ritsumeikan University, Japan. His research area has been computer vision, robotics and artificial intelligence. He is a member of the IEEE Computer Society, the Japanese Society of Robotics, and the president of the Japanese Society of Artificial Intelligence.

Haruki Ueno, received the B.E. in Electrical Engineering from National Defense Academy in 1964, and M.E. and Dr. Eng. in Electrical Engineering from Tokyo Denki University in 1968 and 1977 respectively. He is a professor of National Institute of Informatics, Japan (NII). He is also a professor and chair of Department of Informatics, The Graduate University for Advanced Studies (Sokendai). His interests include knowledge-based systems, symbiotic robotics and distance learning for higher education. Dr. Ueno is the originator of International Joint Conference on Knowledge-Based Software Engineering (JCKBSE), SIG-KBSE, SIG-AI, etc. He is a member of Engineering Academy of Japan (EAJ).