Facial Pose Interpretation for Human-Robot Symbiosis

Bhuiyan, Md. Al-Amin
Daffodil International University

http://hdl.handle.net/20.500.11948/772

Downloaded from http://dspace.library.daffodilvarsity.edu.bd, Copyright Daffodil International University Library
Facial Pose Interpretation for Human-Robot Symbiosis

Md. Al-Amin Bhuiyan
Human-Robot Symbiosis Lab, Dept. of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka 1342, Bangladesh.
Email: alamin_bhuiyan@yahoo.com; alamin@juniv.edu

Abstract: This paper addresses the issues for the concepts of a vision based facial pose recognition system using knowledge based approach for human-robot symbiosis. The system is based on visual information of the human face and is commenced with the face recognition and facial pose classification scheme using pattern-matching strategies. With the knowledge of the known user’s profile, facial attributes are then classified and robots are instructed to perform some specific tasks by issuing corresponding commands. This paper discusses the concepts of facial attributes for human-robot interaction and SIS with its motivations and applications especially to symbiotic robots. The present state of the art in vision, recent computing models for pose interpretation and a few implementations of these novel computing concepts are presented.

1. Introduction
The 21st century is characterized as information technology (IT) that has been reforming immensely the system of the society. It was pointed out for a long time that digitization of information would bring a fusion in the field between communication and broadcast. The globalization of information system carried by the prevalence of Internet promotes every fusion of social informational activities – publication, news, production, distribution, administration, education, daily life, welfare, etc. A Symbiotic Information System (SIS) is an information system in which every element has its own autonomy and constitutes a collaborative community. The SIS aims at such environment as information system or information infrastructure is indispensable and everyone enjoys the benefit of it [1,2].

As robots increase in capabilities and are able to perform more human-like complicated tasks in an autonomous manner, we need to think about the interactions that human will have with robots. There are several ways to communicate with human beings and intelligent machines (e.g. robot, vehicle, etc.) such as text commands, speech commands, gesture commands, and so on. Text command based approaches are more rigid but it is not natural compared to visual perception of human-robot communications. Although the verbal command based human–robot interaction systems have been used based on few keywords (such as walk, turn left, turn right, stop walking, move right, etc.), but nevertheless, have been found with many difficulties to generalize human speeches. This paper proposes a facial expression recognition based human–robot nonverbal interaction system to communicate with robots more likely to human.

2. Facial Pose Interpretation
Face is the most distinctive and widely used key to a person’s identity. Facial pose interpretation, therefore, has attracted considerable attention in the advancement of human-machine interaction as it provides a natural and efficient way to communicate between humans and machines. A substantial amount of research works have been carried out in the last few decades, many of them have been described and compared in two interesting recent surveys by Zhao et al. [3]. Two approaches are commonly used to interpret the face pose for human machine interaction. One is gloved based approach [4] that requires wearing of cumbersome contact devices and generally carrying a load of cables that connect the device to a computer. Another approach is vision based technique that does not require wearing any of contact devices with human body part, but uses a set of video cameras and computer vision techniques to interpret poses [5].

Face pose recognition based on vision technology has been emerging with the rapid development of computer hardware of vision system in recent years and in future will dominate in Human-Computer and Human-Robot Interactions. The poses are being modeled by relating the appearance of any face to the appearance of the set of predefined, template poses. Waldherr et al. [6] proposed a gesture-based interface for human and
service robot interaction. They combined template-based approach and neural network based approach for tracking a person and recognizing gestures involving arm motion. Watanabe et al [7] 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 [8] proposed hand gesture recognition for human-machine interface of robot teleoperation using edge features matching. Rigoll et al [9] used HMM-based approach for real-time gesture recognition. In that work features are extracted from the differences between two consecutive images and considered that the target image is always in the center of the input images. But practically it is difficult to maintain such condition.

A face pose recognition system for person-specific human-robot interaction (HRI) through a knowledge-based approach is shown in Fig. 1. The image analysis and recognition system integrates the detection of human faces in complex backgrounds and recognition of the faces using PCA. The system is commenced with the extraction of largest connected skin like region from the input images by skin color segmentation. Segmented blocks are then normalized and passed through the PCA based face recognition system.

Changes in illumination usually causes large change in the appearance of a face. Therefore, face recognition under arbitrary illumination is employed by using rms scaling and histogram equalization representation.

The pose classification is achieved by employing a multilayer perceptron with back propagation algorithm. If the segmented block matches with the predefined pose at a particular image frame then corresponding pose command is issued. Pose commands and robot actions are interpreted in voice so that the human can hear which pose he/she made and which action is being accomplished by the robot. As an application of this method, the system has been implemented in real time mode using the Sony entertainment robot AIBO: such as to recognize human, greeting and mimic with human by poses, greeting, singing, dancing, and so on.

The face pose recognition and robot interaction system includes three phases: (i) the feature extraction phase (for face detection), (ii) face recognition phase (for person identification), and (ii) pose classification and robot control phase (for human-robot interface). The feature extraction phase is commenced with an image processing technique, which involves an algorithm to detect and isolate the facial area in an image. For this, video color images are converted from RGB to HSV color model. After applying the skin color segmentation process, binary image is obtained from which largest connected component is being analyzed for face detection. On finding the facial area, the image is used for PCA analysis with those of the gray scale images kept into image database. Finally, in the classification stage, a

![Fig. 1 Overall architecture of the face pose based human-robot symbiosis system.](image-url)
multi-layer perceptron is used with back propagation algorithm to classify different poses occurring in the image sequences.

3. Illumination and Contrast Equalization
Contrast is a measure of the human visual system sensitivity. Although the role of contrast is significant in visual processing of computer displays, in almost all of the past literature address the face recognition process in different lighting conditions with different illumination and contrast. To achieve an efficient and psychologically-meaningful representation, all images are processed with same illumination and rms contrast.

The rms (root mean square) contrast metric, which is equivalent to the standard deviation of luminance, is given by:

$$C_{\text{rms}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$  \hspace{1cm} (1)

where $x_i$ is a normalized gray-level value such that $0 < x_i < 1$ and $\bar{x}$ is the mean normalized gray level:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$  \hspace{1cm} (2)

with this definition, images of different human faces have the same contrast if their rms contrast is equal. The rms contrast does not depend on spatial frequency contrast of the image or the spatial distribution of contrast in the image. All images are maintained with the same luminance and same rms contrast using the following equation:

$$g = \alpha f + \beta$$  \hspace{1cm} (3)

where $\alpha$ is the contrast and $\beta$ is the brightness to be increased or decreased from the original image $f$ to the new image $g$. The illumination and rms contrast equalization process is illustrated in Fig. 2.

4. Skin Color Segmentation & Face Detection
Skin color segmentation is based on visual information of the human skin colors from the image sequences. HSV color space is employed for skin color segmentation. The reason behind this choice is two folds: (i) Since the hue components of the HSV color space agrees better with human chromatic perception, the hue value has been employed to describe the colors of images, (ii) Application of the hue components, instead of all the $R,G,B$ colors, obviously reduces a large amount of computational cost. Therefore, the dominant and perceptually relevant skin colors extracted in RGB space from images, are converted into the HSV space. Images are being
searched in HSV 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 the HSV color model a color is described by three attributes hue, saturation and value. Hue is the attribute of visual sensation that corresponds to color perception associated with the dominant colors, saturation implies the relative purity of the color content and value measures the brightness of a color. The HSV space classifies similar colors under similar hue orientations. The conversion from RGB to HSV is given by the equations [2]:

\[
H_1 = \cos^{-1}\left( \frac{1}{2} \left( \frac{(R - G) + (R - B)}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \right) \right)
\]

ranging \([0, 2\pi]\), where \(H = H_1\) if \(B \leq G\); otherwise \(H = 360^\circ - H_1\);

\[
\begin{align*}
S &= \max(R, G, B) - \min(R, G, B) \\
V &= \frac{\max(R, G, B)}{255}
\end{align*}
\]

where \(R, G, B\) are the red, green, and blue component values which exist in the range \([0, 255]\).

The detection of face region boundaries by such a hue segmentation process is shown in Fig. 3. The exact location of the face is then determined from the image for pixels with largest connected region of skin-colored pixels. The connected components are being determined by applying a region-growing algorithm at a coarse resolution of the segmented image. In this experiment, 8-pixel neighborhood connectivity is employed. In order to remove the false regions from the isolated blocks, smaller connected regions are assigned by the values of the background pixels. After thresholding the image may be encountered by some holes in the face skin region. In order to remove the false regions, the face region is subjected to morphological dilation operation with a \(3 \times 3\) structuring element several times followed by the same number of erosion operations using the same structuring element. The dilation operation is used to fill the holes and the erosion operations are performed on the dilation results to restore the shape of the face.

Once the face is found in an image (assuming the largest connected area), it then justified by using ellipse-fitting [2]. The gravity center of the eyes are then localized.

5. Face Recognition using Subspace PCA

The Principal Component Analysis (PCA) based approach is used to recognize gesture from an unknown input image. In the traditional PCA eigenvectors are calculated from training images that include all the poses or classes. This approach of face pose recognition includes the following operations:

1. Generate a set of face poses corresponding to training images \(T^j_i (M \times M)\), where \(i = 1, 2, \ldots, N\) is the number of training images of \(j\)-th class.
2. Compute eigenvectors \(u^j_i\) for each group and choose \(k\) number of eigenvectors \((u^j_k)\) corresponding to the highest eigen values to form PCA.
3. Compute corresponding distribution in \(k\)-dimensional weight space for the known images by projecting them onto the corresponding group and determine weight vectors \((\Omega^j_c)\):
\[ w_k^j = u_k^j (1_c - \Psi_j), \]  
\[ \Omega_c^j = \left[ w_1^j, w_2^j, ..., w_k^j \right]. \]  

where average image of \( j \)-th class \( \Psi_j = \frac{1}{M} \sum_i T_i \) and \( 1_c (N \times N) \) is the known image of \( j \)-th sub-class.

4. Treat segmented region as individual input image and transform each into eigen-image component and calculate a set of weight vectors (\( \Omega_c^j \)) by projecting the input image onto each of the eigen images as Eqs. (6) and (7).

5. Determine if the pose of the image based on the minimum Euclidean distance among weight vectors,

\[ e_c^j = \left\| \Omega_c^j - \Omega_c^j \right\|, \]  
\[ e = [e_1, e_2, ..., e_c^j]. \]

If \( \min(e_c) \) is lower than predefined threshold and closer to any sub class, then the corresponding sub-class is identified.

6. If any segment is face then classify the pattern whether it is known or unknown person. For exact matching \( e_c \) should be zero but for practical purposes we should use a threshold value through experiment.

6. Pose Classification Phase

The pose classification phase includes neural network training for the recognition of pose patterns of the face image. The system is based on supervised learning in which the learning rule is provided with the set of example face patterns (the training set). The multi-layer perceptron, as shown in Fig. 4, with backpropagation algorithm is employed for this purpose. The number of nodes in the input layer is equal to the dimension of the feature vector and the number of nodes in the output layer equals the number of classes the neural network is required to recognize. The network recognizes a feature vector as belonging to class \( m \) if the output of the \( m \)-th node is greater than that of the other nodes in the output layer. During training, the weight and biases of the network are iteratively adjusted to minimize the network performance. The default performance function, for back propagation neural networks, is the mean square error (MSE), which is defined as the average squared error between the networks outputs and target outputs.

7. Knowledge-Based Interface

The system first detects human face using skin color segmentation and then face poses are classified depending on the knowledge stored in the knowledgebase. On face pose classification, the static gestures are recognized using frame-based approach. Known poses are defined as frames in SPAK knowledge base. When the required combination of the pose components is found the corresponding pose frames are activated. Dynamic gestures are recognized by considering the transitions of the face poses in a sequence of time steps. After the pose is recognized, the interaction between human and robot is determined by the knowledge modeled as frame hierarchy in SPAK. SPAK works as a knowledge and data management system, communication channel, intelligent scheduler, and so on [10].

7.1 Knowledge Model for Gesture-based HRI

Knowledge is the perception about and understanding of a subject or a domain. Knowledge representation using frame-based approach is employed for pose classification. A frame is a data-structure for representing a stereotype situation. Attached to each frame there are several kinds of information about a particular object or concept it describes, such as frame name, a set of attributes called slots with values. Table I illustrates the instance frame “Alamin” of the class.
frame “User” and he has semantic link from the robot Aibo and to Behavior “Welcome”. In this frame-based model, several robots can be used and according to user selection the corresponding robot will be activated. Table II shows an example instance frame “Aibo” of the class frame “Robot” and that one is semantically linked to the user “Alamin”. Table III shows the example components of an instance frame “Welcome” of the class frame “Pose”. This frame has three slots for the gesture components. If NormalAlignedRight-sidedFace90 is found in the particular image frame then Pose frame “Welcome” will be activated.

7.2 Human-Robot Interface using SPAK

SPAK has occupied a platform on which various software components for different robotic tasks are integrated over a networked environment. It coordinates the operation of these components by means of a frame-based knowledge modeling. This research combines vision and knowledge-based approaches for human-robot interaction so that user can define or edit robot behavior according to his/her desire. User can also define or edit the rules for pose recognition/interpretation in the knowledge database, which also contains information regarding user (profile) to improve reliability and robustness of image classification.

Using information received from these agents, and based on the predefined frame knowledge hierarchy, SPAK inference engine determines the actions to be taken and submit corresponding commands to the target robot control agents. SPAK consists of a frame-based knowledge management system and a set of extensible autonomous software agents representing objects inside the environment and supporting human-robot interaction and collaborative operation with distributed working environment. SPAK consists of the following major components: GUI interface, Knowledge Base (KB), Network Gateway and Inference engine as shown in Fig. 5. SPAK allows TCP/IP based communication with other software components in the network and provides knowledge access and manipulation via TCP socket. Frame-based knowledge is entered into the SPAK system with full slot information: attributes, conditions and actions. Based on information from remote software components (e.g. pose classification, face recognition, etc.), SPAK inference engine processes facts, instantiate frame instances and carries out the users predefined actions [11].

---

Table I: Instance frame “Alamin” of class frame “User”

<table>
<thead>
<tr>
<th>Frame</th>
<th>Alamin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Instance</td>
</tr>
<tr>
<td>A-kind-of</td>
<td>User</td>
</tr>
<tr>
<td>Has-part</td>
<td>NULL</td>
</tr>
<tr>
<td>Semantic-link-from</td>
<td>Aibo</td>
</tr>
<tr>
<td>Semantic-link-to</td>
<td>Behavior (Welcome)</td>
</tr>
</tbody>
</table>

Table II: Instance frame “Aibo” of class frame “Robot”

<table>
<thead>
<tr>
<th>Frame</th>
<th>Aibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Instance</td>
</tr>
<tr>
<td>A-kind-of</td>
<td>Robot</td>
</tr>
<tr>
<td>Has-part</td>
<td>NULL</td>
</tr>
<tr>
<td>Semantic-link-from</td>
<td>Alamin</td>
</tr>
<tr>
<td>Semantic-link-to</td>
<td>Behavior (Welcome)</td>
</tr>
</tbody>
</table>

Table III: Instance-frame “Welcome” of class-frame “Pose”

<table>
<thead>
<tr>
<th>Frame</th>
<th>Welcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Instance</td>
</tr>
<tr>
<td>A-kind-of</td>
<td>Pose</td>
</tr>
<tr>
<td>Has-part</td>
<td>NULL</td>
</tr>
<tr>
<td>Semantic-link-from</td>
<td>Alamin</td>
</tr>
<tr>
<td>Semantic-link-to</td>
<td>Behavior</td>
</tr>
</tbody>
</table>

Fig. 5 SPAK knowledge management system.
8. Implementation of Human-Robot Interaction

The system has been implemented by gesture recognition program running in server PC. The PC has been considered as the server and the robot AIBO as a client. The communication link has been established through TCP/IP protocol. The human-robot interaction results are: (i) start walking, (ii) turn left, (iii) turn right, (iv) stop walking, (v) dance, and so on, according to the gestures of FrontalFace, LeftDirected45, RightDirected45, FrontalEyeClose, ShakeFaceUpDown etc. One of the actions of the robot as a result of gesture recognition process is shown in Fig. 6. The gesture recognition commands employed for controlling the robot are listed in Table 1.

9. Conclusions

Recognition of facial pose using machine vision techniques has many useful applications. Though human beings accomplish these tasks countless times a day, they are still very challenging for machine vision. Most of the researchers attack this kind of problem with face localization and feature selection with frontal view faces and without facial expression and normal lighting conditions although the variation between the images of the same face is too large due to facial expression, hair style, pose variation, lighting conditions, make-up, etc. We have investigated the facial pose interpretation and its application for symbiotic robots. While present robots are mechatronic sensor and action systems without intelligence. Research and development on intelligence features are indispensable for a next generation symbiotic robot. Our research aims at providing of the basic concept for a next generation symbiotic robot. Our main target is to establish a symbiotic society so that the robots can exchange their ideas and thoughts with human beings for the benefit of each other.

Table 1 Commands to control AIBO

<table>
<thead>
<tr>
<th>Gesture Commands</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal face</td>
<td>Start walking</td>
</tr>
<tr>
<td>Frontal face with closed eye</td>
<td>Stop walking</td>
</tr>
<tr>
<td>Right directed face (45°)</td>
<td>Turn Right</td>
</tr>
<tr>
<td>Left directed face (45°)</td>
<td>Turn Left</td>
</tr>
<tr>
<td>Up state face</td>
<td>Forward</td>
</tr>
<tr>
<td>Down state face</td>
<td>Backward</td>
</tr>
<tr>
<td>Right directed face (90°)</td>
<td>Turn light on</td>
</tr>
<tr>
<td>Left directed face (90°)</td>
<td>Turn light off</td>
</tr>
<tr>
<td>Shake face left/right</td>
<td>Sing a song</td>
</tr>
<tr>
<td>Shake face up/down</td>
<td>Dance</td>
</tr>
</tbody>
</table>

References
