Detection of Human’s Focus of Attention using Head Pose

Sadie Afroze
Dept. of Computer Science and Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
e-mail: sadiaafroze55@yahoo.com

Mohammed Moshiul Hoque
Dept. of Computer Science and Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
e-mail: moshiuhl@yahoo.com

Abstract—We address the problem of visual focus of attention (VFOA) based on human head pose. We observed the human head pose and the object that are in front of the human. We integrate the head pose with target object to monitor human attention where the human is actually looking at. The VFOA is a vital cue for drawing human attention and establishing human interaction. As it assists to understand what the person is doing and it notifies addressee-hood (who is seeing at whom). In Human Robot Interaction approach it is helpful as it attracts and control the attention of a target person depending on his or her current visual focus of attention. It is especially important to monitor human attention based on head pose as the person who is disabled and cannot speak but express their needs in terms of head pose. For recognizing a human VFOA, we noticed their head pose because head pose is an important hint in non-verbal communications. We evaluate our system in series of experiments and observed our system performance. We found a good result and got good exactness of the system performance.

Keywords—human computer interaction; head pose; visual focus of attention; object detection

I. INTRODUCTION

Human attention means the concentration of mind on a single object which is expressed by some physical expression and controlled by neural activities. Attention is also referred to as the allocation of limited processing resources. Eye movements, head movements or changes in body orientation are the observable behavioral responses for attracting a person’s attention. It is necessary for a system to know a user’s intention and focus of attention for building expert human interfaces. Detection of such information can be utilized to build natural and intuitive interfaces as the motion of a person’s head pose and gaze direction are closely related with his or her intention and attention.

One such characteristic of interest is the gaze, which indicates where and what a person is looking at, or, in other words, what the visual focus of attention (VFOA) of the person is. So, the first step in determining a person’s focus of attention and intention is to track his/her gaze. In robotics in order to provide assistance to people with disabilities it is helpful for a robot to recognize the person’s head pose. The robots have the potential to assist with a wide array of tasks and activities, assist people with diverse conditions, and assist people who are in bed, in a wheelchair, or are ambulating. If there is a camera in front of the disabled people and another camera in front of the objects of the disabled person, the robot can easily identify the person’s attention by integrating the person’s head with the target objects. Actually it is the common nature of human that he is look at the object that he needs. So, if we can implement the system it will be helpful for the disabled person who cannot speak but can express his needs or desire by using his head pose. Moreover, in absence of high definition image, we depend on human head pose to recognize the visual focus of attention. Thus, tracking the VFOA of people could have important applications for developing ambient intelligent systems. So, in our work, we consider head pose for recognizing visual focus of attention. Due to the physical placement of the VFOA targets, the identification of the VFOA can only be done using complete head pose representation. Here we use webcam for tracking the head of the person and from which we will estimate the head pose especially yaw pose of the person. Next, we will track the object which is in front of the person by another camera. Finally we will integrate this for getting visual focus of attention.

II. RELATED WORK

In HRI and HCI contexts, many conversational systems need VFOA information for analyzing and performing necessary interactions. Most works use either sensor-based or high definition image approaches which are not usually applicable or helpful for interaction with robots for recognizing VFOA. In the previous work [1] they proposed an intelligent robotic method of attracting a target person’s attention in away congruent to satisfying social requirements. The used HOG pattern features and an SVM classifier for recognize VFOA. In the previous work [2] they used 3D geometric modeling for head pose estimation. Based on two human eye-areas, they model a pivot point using distance measure devised by anthropometric statistic and MPEG-4 coding scheme. But this 3D approach is quite complicated and complex rather than 2D geometric modeling. In the previous work [3] they employ neural networks to estimate a person’s head pose from camera images, and a probabilistic model to recognize interesting targets in the scene based on the observed head pose. In the previous work [4], Hidden Markov Model (HMM) used to recognize VFOA and introduce the standard gaze model. They used both supervised and unsupervised learning approach. They mainly tried a
improved VFOA recognition. But in practice, to set such a value might be difficult and complicated. Because in the robot interaction application, the same strategy does not produce good result in all condition. Asteriadis et al. [5], information from head rotation and eye gaze are used. The authors use Bayesian modality fusion of both local and holistic information, in estimating head pose, as well as fuzzy fusion of head rotation and eye gaze estimates, in a fully automatic manner. However, in their proposed approach [7], the user needs to maintain a frontal pose to the camera at start-up and they do not address finding appropriate mappings between 2-D projections and head/eye gaze analysis to certain points on a target plane. In contrast to previous works we would like to focus on recognizing VFOA in real time.

III. PROPOSED FRAMEWORK

In our work, we recognize the visual focus of attention of human in real time using head poses of human. To detect the visual focus of attention, first of all we track the human face and estimate its pose. The system detects the object from multiple objects corresponding to human head pose. That means the person where he is looking is detect by the system through integrating head pose and target object. Thus the system shows the visual focus of attention of human. This task is very important for a disabled person.

We also all detect the objects that are in front of the person. We use multiple object detection algorithms using shape of the object. We also number the object so that it can easily recognized by the user and it is helpful to count the number of object. Fig. 1 Illustrates the schematic diagram of propose system methodology.

A. Tracking the Human Head

The first step of our work is to track human face. Here we track human face using an open source named Stasm (standard active shape model). We use the Stasm 4.1.0 in our work. Stasm employ the OpenCV frontal face detector. The face should be at least a quarter of the image wide. Stasm does not use color information, i.e., it internally converts the image to monochrome before searching for landmarks. An unused or uninitialized landmark has a position in a shape matrix with both x and y equal to 0. The x position of a valid landmark that happens to be at [0,0] is thus jittered to 0.1 (a one tenth of a pixel offset). Unused landmarks are for more esoteric applications of Tasm where landmarks are synthesized or models are built with different sized shapes. It is unlikely that you will need unused landmarks when creating your own shape files. Fig. 2 represents the overview of the face tracking algorithm.

Suppose now we have s sets of points x, which are aligned into a common coordinate frame. These vectors form a distribution in the 2n dimensional space in which they live. If we can model this distribution, we can generate new examples, similar to those in the original training set, and we can examine new shapes to decide whether they are plausible examples. Approximate any of the original points using a model with less than 2n parameters. If we apply a PCA to the data, we can then approximate any of the training set, x using

\[ x = \bar{x} + b \]  

(1)

Where \( \bar{x} \) is average of all training sets, b is n dimensional vector. The vector b defines a set of parameters of a deformable model. By varying the elements of b we can vary the shape, x.

B. Head Pose Estimation of Human

After tracking the human face, we estimate human head pose. We have performed point base geometrical analysis and expected pose value measurement to estimate the head pose of the human. After performing these two measurements,
we estimate the head pose of the human head to know where the visual focus of attention of him/her.

- **Expected Pose Value Measurement**: The expected value of a random variable is intuitively the long-run average value of repetitions of the experiment it represents. Suppose random variable X can take value x₁ with probability p₁, value x₂ with probability p₂, and so on, up to value xₖ with probability pₖ. Then the expectation of this random variable X is defined as:

\[
E[X] = x_1p_1 + x_2p_2 + \ldots + x_kp_k
\]  

(2)

To find the expected value of the head pose, we train our system by establishing three models. The first model is for left head pose, the second model is for front head pose and the third model is for right head pose. We observe the three models and find the threshold value for each model which was a floating point value. We observed the facial landmarks that are obtained from the face tracking module. We separate out the left side points from the right side points. We found 13 significant points on left side of the head and 13 significant points on right side of the head. Then we calculate the expected value by the equation 2. The value of current head pose value is compared with each head pose model and if it is nearest to the any model, it is said to the corresponding head pose. Thus we calculate the expected head pose of the head.

- **Point Based Geometric Analysis**: After calculating the expected value of the head pose we calculate the point bases geometric analysis of the head pose. In the learning part we first build up the point relationship model and we got some significant change in the points of the head poses. We observed that the distance between the point of eye center and nose tip is always change in the different head pose. We found out these points and calculate the Euclidian distance between them. The formula of Euclidian distance is:

\[
D_1(p, q) = \sum_{i=1}^{n} |p_i - q_i|
\]  

(3)

After combining these two approaches we measure the head pose of the human.

**C. Detecting the Objects in Front of the Human**

In this work we detect the objects based on shape and edge. After detecting the edge we need to focus the detected object. We represent the detected objects by a rectangle. After detecting all the objects we only detect the target object. But then the bounding rectangle is on the only target objects and it represents the target object where the visual focus of attention of the person was. Fig. 3 shows the object detection algorithm.

Then we detect the edge of the object from image. Here we use Sobel edge detector for detecting edge. After detecting edges we perform morphological image processing. We use this as it process images based on shapes and apply a structuring element to an input image and generate an output image. We use erosion and dilation as structuring element. After performing morphological operation we apply thresholding and binarization of input frame.

After this we find out the contour (boundary of shape) of the object it is like finding out the white object from black background. Each contour is stored as a vector of points. After finding the contour we estimate the bounding box of the target objects. We estimate it by analyzing the contour area of the object. We estimate a fixed value from various shapes of the objects and compare the contour area with this value. The contours that are satisfy this constraint is considered as target objects. We represent the object by a bounding box. We use rectangles to focus the target object. The bounding box only contains the topmost coordinate, width and height of the rectangle. Thus we detect multiple objects in front of the human.

**D. Integrating Head Pose with Detected Object to Recognize the VFOA**

After tracking object and face, the system integrate the tracked object on the corresponding head pose. For mapping the head pose with the corresponding object the system first tracked all the objects in front of the human. Then on the basis of the head pose it recognizes the object that the person is looking at. Such as if a human moves head on the left side; the system tracked only the object left side of the person and shows that the person’s attention is on that object. On the other hand if a human moves head on the right side; the system tracked only the object right side of the person and shows that the person’s attention is on that object. Besides this, the system also shows that if the human face is frontal face, the system integrates frontal pose with middle object and indicate that his attention is in front of the middle object. Actually the integration is done on the basis of head pose and its corresponding tracked object. So we can see that we have two parameters for mapping the head pose with target object and we can represent it as follows:
\[
\text{VFOA} = f(H, O)
\]
Where, VFOA= Visual focus of attention, H= Head pose of the person and, O= Object where the person is looking at.

The head pose of a person is intersects with the corresponding object. We represent the target object in a rectangle. But when we integrate face tracking module with the object detection module, only the rectangle is placed only in the object with corresponding head pose. We draw a straight line from the nose tip of the person. We consider is an initial head rotation which we represent by another straight line. When the angle between initial head rotation and original head rotation zero we map this head pose with the middle object. Again there is a horizontal line from the boundary box of the object. If the two lines intersect perpendicularly then we can say that the attention of the person is on that object. If we consider the line from nose tip is \(y_1\) and the line on the boundary box of the object is \(y_2\), then we find out perpendicular condition using the following formula:

\[
tan(\theta) = \frac{m_1 - m_2}{1 + m_1 \times m_2}
\]  

Where \(m_1\) is the slope of line that is drawn from the nose tip and \(m_2\) is the slope of the line that is drawn from bounding box. Fig. 4 shows the mapping of head pose with target object.

\[\text{IV. EXPERIMENTAL RESULTS}\]

To analyze the face tracking module accuracy we emphasize on quantitative approach. We worked with 10 people in the real environment to run this experiment. The average age of participants are 23 years (SD= 4.50). To evaluate the system, we told the participants about our experiment and also described them what they should do. Participants were interacted with the system one by one. Each participant attended in a two steps experiment.

\[\text{A. Experimental Setup}\]

The first phase of the implementation for this system is to establish a setup environment with two webcam. One camera can detect the human face and another can track the objects in front of the human. Fig. 5 shows the experimental setup of our system.

\[\text{Fig. 5. (a) Participant interacting with the proposed system} \quad \text{(b) Experimental setup of the system.}\]

We test the system performance using different types of the object and also detect the visual focus of attention by different person. We asked the person to take sit before the webcam and set all the objects in front of the human. Then we evaluate the system performance. We can also set this webcams into a tripod to test our system performance.
B. Evaluation Method for Face Tracking Module

We positioned the participant in front of the camera within the viable range and more than 10 trials have been given by each of the participants. We build 640×480 size image of the different person of three head poses which is used as training set of the face tracking module performance. To evaluate the system, we told the participants about our experiment and also described them what they should do. The qualitative evaluation of the module, which is performed with the experimental data, is given below.

<table>
<thead>
<tr>
<th>No of Trials</th>
<th>Tacked Face</th>
<th>Deviation</th>
<th>Accuracy (%)</th>
<th>Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>15</td>
<td>14</td>
<td>1</td>
<td>93.33</td>
</tr>
<tr>
<td>P2</td>
<td>13</td>
<td>11</td>
<td>2</td>
<td>84.61</td>
</tr>
<tr>
<td>P3</td>
<td>17</td>
<td>15</td>
<td>2</td>
<td>88.23</td>
</tr>
<tr>
<td>P4</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>91.67</td>
</tr>
<tr>
<td>P5</td>
<td>19</td>
<td>18</td>
<td>1</td>
<td>94.73</td>
</tr>
<tr>
<td>P6</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td>P7</td>
<td>18</td>
<td>15</td>
<td>3</td>
<td>83.33</td>
</tr>
<tr>
<td>P8</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>P9</td>
<td>17</td>
<td>16</td>
<td>1</td>
<td>94.11</td>
</tr>
<tr>
<td>P10</td>
<td>14</td>
<td>13</td>
<td>1</td>
<td>93</td>
</tr>
</tbody>
</table>

Graphical representation of the experiments is given below in Fig. 6.

![Fig. 6: Performance analysis of face angle estimation module](image)

C. Evaluation Method of Object Tracking Module

In Fig. 7, we evaluate the object tracking module based on ROC (Receiver Operating Characteristic curve) analysis shown in.

![Fig. 7: ROC curve of object detection module](image)

Fig. 7 represents a graphical plot that illustrates the performance of object tracking module. The curve is created by plotting true positive rate and false positive rate of the detected object and ground truth object.

D. Overall System Performance Analysis

We observed the performance of face tracking module and object tracking module. Then each of modules is combined to detect the visual focus of attention detection. Prior to that some important factors (i.e. face tracking of human, head pose estimation, object detection) are checked to evaluate the overall system performance. Also these values are compared to the physical values to find the standard deviation. Table II shows the data table to determine system accuracy.

<table>
<thead>
<tr>
<th>Head Pose</th>
<th>No. of Trials</th>
<th>No. of Head Pose</th>
<th>Detected object with corresponding Head Pose</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>10</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Front</td>
<td>16</td>
<td>30</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Right</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

The graphical representation of the system accuracy is shown in Fig. 8. We observe the correct detected object of corresponding head pose and represent the accuracy by a graphical representation where we plot the number of head pose and tracked object in front of the human. We tested this system by several persons and the quantitative evaluation proved that the system working quite satisfactory.
Our main concern was to develop a system that recognizes the visual focus of attention of a person. We consider head pose to recognize a person attention as head pose plays an important cue for developing communication channel. In our work we addressed the visual focus of attention based on their head pose. First of all we track the human face and estimate his/her pose and finally mapped pose with corresponding VFOA targets. We use a setup of multiple camera views in order to achieve unobtrusive captures of participant’s visual attention. From the estimated field of view, we deduce the most likely focus target by using an adaptive scheme of mapping head orientation to its most likely gaze angle counterpart and hence the target this person is looking at. Integration of multiple cameras will increase the field of view of the camera and can tracks the multiple human’s focus of attention. This issues are left for future research.

REFERENCES