The University of Birmingham
School of Computer Science
MSc in Advanced Computer Science

Second semester mini-project

# Controlling a Robot using Eye Gaze and Body Pose

**V Chauhan**

Supervisor: H J Chang

April 2018

## Abstract

Human-Robot interaction is an active area of computer vision research with many potential applications. Previously, HRI research focused on hand gestures and sign language. In order to operate HRI naturally understanding of gestures of the entire body is essential. Proper implementation of human-robot interactions is crucial due to the nature of interactivity and behavior of human.

This report documents a robotic system capable of human interaction and finds optimal path while avoiding obstacles. The aim of building this robot is to understand human gestures and perform task accordingly such as moving to a certain place where the user is pointing at such as (1) with the implementation of eye gaze estimation using a depth camera. The system consists of two components estimating human body poses and eye gaze and planning the optimal path to reach the destination where human is pointing.

***Keywords***— Human-Robot Interaction, Human Tracking, Body Pose Estimation, Eye Gaze Estimation, Path Planning

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Traditional robots were used only for specific purposes such as manufacturing and transportation. According to a recent survey of The United Nations (UN) robots can be divided into three categories: personal, industrial, and professional services. Industrial robots are commonly used everywhere. The primary purpose of professional and personal robots is to assist people to reach their goals (5). A human gesture is a form of non-verbal communication in which physical behavior is used to express information. Gestures can be divided into two categories(6):

(i) Communicative: A communicative gesture is a motion which consists a meaning to express goals.

(ii) Non-communicative gesture: A non-communicative gesture is a motion which connects communicative gesture to hidden goals.

To address the issue of recognizing human actions, first, the proper implementation of human pose estimation is required. The goal of applications of human pose estimation is to estimate the structural parameters that fit a human in single or multiple images. In recent years, gesture recognition benefits from many machine learning algorithms such as Hidden Markov Models, Support Vector Machines (7), and random forest classifiers (8), etc.

In this work, a Human-Robot Interaction system presented whose main purpose is to allow users to communicate with a robot in an intuitive gesture-based manner. The experimental setup is composed of a mobile robot ( Pioneer P3DX) and a Kinect sensor. In this system, the robot can understand human gestures and perform a task of moving to a specific location where human is pointing and looking. To validate the robustness of the system, it has been tested by three people multiple times to report the recognition rates as well as planning the optimal path.

The report is organized as follows: Chapter 2 illustrates the previous research that has been done in the area of human-robot interaction. Chapter 3 presents the experimental results including the performance of Body Pose estimation and local and global planner, and accuracy of pointing location. Finally, Chapter 4 concludes the paper.

# Chapter 2

# Background and Literature Review

The task aims to study the gestural communication for the human-robot interaction. This chapter classifies various robot interaction and path planning methods in different ways and gives general information about the traditional methods.

## 2.1 Feature Representation

The human body is a complex system which consists of many joints and limbs. In 3d, the representation of these parts is challenging task. (9) analyzed how human comprehend the 3D pose space and they created a massive dataset which is consist of 2d and 3d poses. In addition to that, they also recorded a variety of human body shapes using synchronized eye movement similar to humans and measured how accurately humans recreate 3D poses. In their experimentation, they found out that humans are not better than computer vision algorithms at acting out 3d poses in the lab environment.

Despite these challenges, Model-based approaches propose prior information of a human body/skeleton to overcome this difficulty. A typical stick figure representing skeleton has been shown in figure 2.1 (2) which defines kinematics and appearance characteristics.

According to (10), after the getting the structure of body model, constraints are usually applied to the parameters. For example, to calculate the pose parameters like the joint angle or limb length, specific rules have to follow. Furthermore, there are few constraints in which either some body parts are

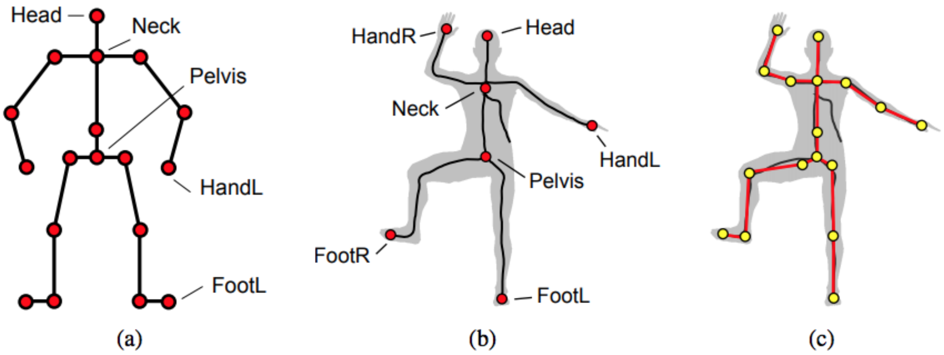


Figure 2.1: (a) Human skeleton with 18 joints. (b) skeleton with labeled end-points. (c) skeleton representing the body appearance. (2)

obstructed by other parts counted twice or because of symmetry in body parts appearance keeps changing.

After the model-based approach, feature extraction techniques were used to extract features like edges, color or silhouettes. To describe these features image descriptors were introduced to reduce the size of the feature space. Common feature representations are SIFT( Scale Invariant feature transform) (11), Appearance and Position Context (APC) (12), Shape context descriptor (13) appearance and position context descriptor is a local image descriptor which captures the context information of the local structure as well as their relative positions. Furthermore, (14) introduced posebits which represent geometrical connections between body parts. This technique can solve depth ambiguities and can also provide 3d poses without any 3d annotations.

## 2.2 Extracting 3D human pose estimation from a single image

A single monocular RGB image has three characteristics that affect its performance at the time of reconstruction of 3D points:

(i) There can be substantial issues in 3D space due to the small error in 2D body poses location

(ii) (15) explained that in higher dimensions the performance is terrible. Existing techniques illustrate different solution for this issue which is discussed in Section 2.2.1

(iii) Different 3D poses can produce similar image projections.

### 2.2.1 3D human pose estimation from a monocular image

Due to the nonlinear human motions, poses, different appearance, and ambiguity between 2D and 3D pose the extraction of 3D pose from single images is a very challenging task. (16) experimented with a HumanEva-I dataset to extract 3D poses by explicitly using a single monocular image. Steps were taken by author:

(i) Training:
(a) 3d poses are projected to 2D and regression model was created from 2D annotations

(ii) Tasting:
(a) 2D pose was estimated
(b) The nearest 3D poses were predicted
(c) Final 3D poses were obtained by minimizing the projection error

#### 2.2.1.1 Deep Learning Methods

Deep learning methods are composed of multiple non-linear transformations which are a representation-learning approach (17). Deep learning methods can be used in both supervised and unsupervised learning. This approach produces significantly better results in many computer vision tasks such as image classification and object recognition. Furthermore, New approaches which utilize deep-learning techniques have been proposed after addressing the 2D pose task with a great result (18), and after this, recently, 3D human pose estimation tasks were approached using deep learning techniques. (19) introduced a framework which takes as an input image and a 3D pose and gives a score value that represents similarity between the two inputs. In this framework, the author used a convolution network and two subnetworks for feature extraction and performed a non-linear transformation of the image and pose into a joint enclosure respectively. Moreover, also a maximum-margin cost function is used during the training which applies a re-scaling between the score value of the truth image poses and the rest image poses.

### 2.2.2 3D human pose estimation from a single image from a multi-camera view

The uncertainty in 3D space can be resolved in the presence of depth information which can be gathered from Microsoft Kinect (20). Kinect can only be used for a particular application due to the limited range of the sensor.

(21) developed a framework for multi-view articulated pose estimation. First of all, compute probability distribution with 2D part detectors which is based on HoG (Histogram of oriented gradients) features for the position of body parts. The parts and dependency structure of the variables of the model represented through tree graph and a Bayesian network respectively. The combination of prior pose and dynamic programming is used to approximate the state space using distinct quantities. For the translation, prior, max-product algorithm is used with two variations according to the limitation. At last, due to the tree structure, a two-step algorithm is applied to deal with the double counting.

(13) used a different approach of using the 3D pose over a set of 2D projections of the 3D pose in each camera view. This 2D pictorial structure model is extended with color features, flexible parts and mix pictorial structures. Appearance and spatial correspondence across views are applied to take advantage of the multi-view setting. The final 3D pose is created from the 2D projection by triangulation.

## 2.3 Extracting 3D human pose from a sequence of images

The most significant issue with locating the 3D position of the body parts from a progression of images is that the appearance of a body may change because of the number of things such as any changes in background, camera movement, illumination, and rotation in-depth of limbs.

### 2.3.1 3D human pose estimation from a sequence of monocular images

In general, most of the cameras use a single lens, so estimating human body pose from a monocular image are essential. Accurate pose estimation is an ill-posed problem, and these issues can be improved by utilizing the available information in a frame over a period (22).

#### 2.3.1.1 Discriminative approaches

To reduce depth uncertainties (23) exploited spatiotemporal data. In this work, authors developed two convolution networks to first align the bounding boxes of the human body in successive frames and then filter them to create a data. To reconstruct the 3D poses directly from the data authors used Kernel Ridge Regression (KRR), Kernel Dependency Estimation (KDE), and a 3D histogram of oriented gradients (HoG) descriptors. They also exhibited that (i) challenging poses where self-occlusion occurs can be estimated with more accuracy when information from different frames exploited. (ii) In early stages, the linking of detections in frames on which later enforcing temporal consistency which can improve the performance significantly. Furthermore, (24) convolutional networks were also employed in deep-learning regression architecture to encode dependencies between joint locations and an auto-encode were used during the training on existing human poses to learn a structured representation in 3D. After that, a convolutional network maps a regression framework that estimates 3D poses from an input image.

#### 2.3.1.2 Latent variable models

It is often difficult to determine the accurate estimates of the part labels because of possible occlusions, to solve this issue latent variables are used. (25) proposed an approach that uses latent variable models that successfully deal with the overfitting and poor generalization. They also introduced the canonical local preserving latent variable model which is a combination of canonical correlation analysis (CCA) and kernel canonical correlation analysis (KCCA) that preserves structure in the data. Latent spaces are learned for both features and 3D poses by maximizing the non-linear dependencies while preserving structure in the original space. (22) proposed a latent variable approach with the aim of estimating 3D human poses with multiple people in a real-world scenario where partial/full occlusions occur. The author proposed a three-step discriminative method using Bayesian formulation:

(i) enforced discriminative 2D part detection to determine the locations of joints in the image.

(ii) 2D tracking-by-detection approach applied to obtain the people tracklets which also exploits temporal coherency and improve the robustness of estimation.

(iii) A hierarchical Gaussian process latent variable model (hGPLVM) with

Hidden Markov Model (HMM) applied to retrieve 3D pose.

This method is useful to track multiple people in a real-world application.

#### 2.3.1.3 Particle filter algorithm

Particle filter algorithm is used to track 3D human motion. (26) proposed an exemplar-based conditional particle filter (EC-PF) to track the full human body motion in which system state is implemented which is conditional to image data and exemplars to improve the accuracy of prediction. At the time of construction of the model, a shape context matching is applied to estimate 3D pose from a monocular camera. (27) proposed a hybrid approach in which authors used Gaussian Process regression for the discriminative part and a motion model with an observation likelihood model to estimate the pose. Furthermore, They also used a discrete cosine transformation of the silhouette features to detect shapes. This approach shows quite promising results.

### 2.3.2 3D human pose from a sequence of multi-view images

(28) proposed an approach which employs 3D human models to estimate the pose from a progression of multi-view frames. This type of approach also increases some issues such as occlusions between individuals. Authors used triangulation of corresponding body joints sampled from the posteriors of 2D body part detectors in all pairs of views even though they wanted to solve the higher dimensional complex state space at the first place similar to (21). The authors introduced a 3D pictorial structures model which is based on Conditional Random Field (CRF) with multi-view potential functions and enforces rotation and collision constraints. This model concludes the multiple human's articulated poses while resolving uncertainties that arise from multiple views and human estimation. In the end, after sampling from marginal distributions, the assumption on the 3D pictorial structures model is applied with belief propagation algorithm.

This method was extended by creating the 3D pictorial structures model temporally consistent with (29). In this new method, they recover the identity of individuals using tracking and then infer the pose, this allows efficient conclusion because of the results are in a smaller state space. A solution approach is also introduced to penalize candidates who geometrically differ significantly from the temporal joint, and because of this, the inferred poses are consistent over time. In addition to that, authors built upon their previous work by applying a 3D pictorial structures model on a distinct body part parametrization (30). They did not define any parts in 3D positions and orientation instead they maintained only the position parameters and encoded the orientation in the factor graph.

## 2.4 Eye Gaze Estimation

The eye gaze is a process of calculating the gaze point and the eye's motion concerning the head position. After evaluating the movement of eyes, the gaze point can be estimated (31). The tracking algorithms of eye gaze are based on corneal reflection methods in which Near-infrared illuminations (NIR) are used for estimating the gaze direction using a geometrical model or polynomial function of an eye. There are different methods available to estimate gaze direction which uses information such as shape, texture, and local features and visible light.

### 2.4.1 2D regression based methods

Regression-based methods use a polynomial transformation to map gaze coordinates of the vector between pupil and corneal glint. According to (32), the mapping function can be presented as:

$$f : (X_e, Y_e) \rightarrow (X_s, Y_s)$$

where $X_e, Y_e, X_s$ and $Y_e$ are equipment.

$$X_s = a_0 + \Sigma_{p=1}^{n} * \Sigma_{i=0}^{p} a_{i,p} X_e^{p-i} Y_e^{i} \tag{2.1}$$

$$Y_s = b_0 + \Sigma_{p=1}^{n} * \Sigma_{i=0}^{p} b_{i,p} X_e^{p-i} Y_e^{i} \tag{2.2}$$

Where n describes polynomial order, $a_i$ and $b_i$ are the coefficients. The polynomial is optimized by calibration. To reduce mean squared difference ($\epsilon$) between original and estimated coordinates the coefficients and order are taken, which states:

$$\epsilon = (X_s - Ma)^T (X_s - Ma) + (Y_s - Mb)^T (Y_s - Mb) \tag{2.3}$$

Where M is transformation matrix and a and b are the coefficient vectors given by:

$$a^T = [a_0a_1...a_m], b^T = [b_0b_1...b_m] \tag{2.4}$$

$$M = \begin{bmatrix} 1 & X_{e1} & Y_{e1} & ... & X_{e1}^n & ... & X_{e1}^{n-i}Y_{e1}^i & ... & Y_{e1}^n \\ 1 & X_{e2} & Y_{e2} & ... & X_{e2}^n & ... & X_{e2}^{n-i}Y_{e2}^i & ... & Y_{e2}^n \\ : & : & : & ... & : & ... & : & ... & ... \\ 1 & X_{eL} & Y_{eL} & ... & X_{eL}^n & ... & X_{eL}^{n-i}Y_{eL}^i & ... & Y_{eL}^n \end{bmatrix} \tag{2.5}$$

Where M is the transformation matrix, m the no. of coefficients, and L is the calibration point (33). The coefficients can be gained by inverting the matrix as (34):

$$A = M^{-1}X_s, b = M^{-1}Y_s \tag{2.6}$$

(32) compared multiple calibration configurations and mapping functions with 15x9 point grid and concluded that to determine the accuracy of an eye tracker calibration arrangements and components of mapping functions are essential. (35) achieved robust gaze estimation by using neural networks. In addition to that, a 2D mapping algorithm was introduced by (36) which can handle unconstrained head movements and distorted corneal reflections. This algorithm shows high-reliability measures for loss of corneal reflections which is caused by head or eye movement. Furthermore, the method of (37) shown better accuracy than simple regression. In this method, the author used a 3-layer neural network to estimate the mapping function between pupil vector and gaze coordinates.

### 2.4.2 3D model based methods

3d model-based methods estimate the center of the cornea, optical, and visual axes of an eye by using a geometrical model of the human eye. This method can be separated regarding single and multi-cameras. There is no moving parts or fast re-gaining capabilities in single camera system because of its simple geometry system. (38) used a single camera with LED to get an accuracy of 0.5 degrees after user calibration. (39) presented a mathematical model to estimate the point of gaze from light sources and single and multiple cameras. Authors used a captured video frames to reconstruct visual and optical axes from the pupil and glint. The model demonstrates the performance of gaze tracking system with an accuracy of 0.9 degrees which is implemented by one camera and two Near-infrared illuminations light sources. After this, (40) proposed a new system in which they achieved 3D gaze tracking with free head motion using multiple LEDs and a single camera.

Multi-camera methods proved their robustness in case of head motion, the issue with this method is that the cameras should be calibrated with the estimated position of LEDs. (41) introduced a 3D gaze tracking technique which includes a gaze detection and eye position unit with a free head motion by using two camera system with simple two-point calibration. To achieve free head motion tracking, eye positioning unit uses a narrow field of stereo cameras and also controls gaze positioning unit direction.

(42) estimated the optical axis of an eye by solving linear equations. In this simplified model author used two LEDs and two cameras and a single point calibration. Another method introduced by (43) uses two camera system which is calibrated once by the user and a dynamic head compensation model with gaze mapping function which updates the mapping function whenever the head moves to achieve tracking.

A new type of 3D gaze tracking with the usage of a depth camera (Kinect) has been proposed by (37). These sensors include an infra-red and RGB camera with the resolution of 640x480 pixels and 45 and 58 degrees vertical and horizontal field of view respectively. In this method, eye parameters are derived by a calibration process, and Kinect is used for getting 3D coordinates of an eye. This is useful to track 3D gaze coordinates in real time. Another work, (44) used Kinect data is used to obtain 3D positions of head pose, iris, and center of the eyeball of a person. They use a convolution-based means of gradients iris center localization and a geometric constraints-based method to derive the 3D model parameters and estimate the center of the eyeball.

### 2.4.3 Appearance based methods

In these methods, a model is trained with a set of features extracted from eye images to represent eye region. (3) used a statistical model to represent texture and shape variations and trained eye region annotated with landmarks as shown in the figure 2.2(3).
The shape vector is concatenated coordinates of all points

$$s = (x_1, x_2, ...., X_L, y_1, y_2, ..., y_L)^T \tag{2.7}$$

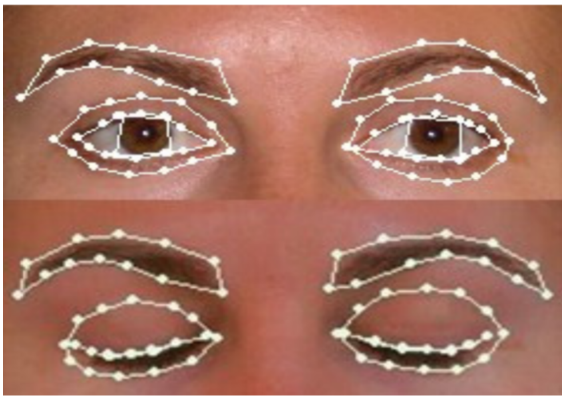


Figure 2.2: Image annotations for open and closed eyes. (3)

Where L is the total no. of landmark points. The model is acquired by PCA (Principal Component Analysis) on a set of aligned shapes.

$$\begin{aligned} s &= \bar{S} + \phi_s b_s \\ \bar{S} &= 1/N_s \Sigma_{i=1}^{N_s} s_i \end{aligned} \tag{2.8}$$

This equation is derived from (3). Here $N_s$ is the no. of shape observation; $\bar{S}$ is the mean of shape vector; $b_s$ is the set of shape parameters.

Similarly, a texture vector is defined to train images: $t = (t_1, t_2, ..., t_p)^T$. The model is derived by using means of PCA:

$$\begin{aligned} t &= \bar{T} + \phi_t b_t \\ \bar{T} &= 1/N_t \Sigma_{i=1}^{N_t} t_i \end{aligned} \tag{2.9}$$

where $N_t$ is no. of texture observations; $\bar{T}$ is mean texture vector.

The set of texture and shape parameters describe the appearance variability of the model:

$$c = \begin{pmatrix} W_s b_s \\ b_t \end{pmatrix} \tag{2.10}$$

where $w_s$ is vector of weights.
Active Appearance Model algorithm uses this statical model to fit the best

model to an eye image.

In (45) the face and eye region fitted from training local and global appearance models. Eye gaze has been classified in 6 directions by adopting two different approaches.

#### 2.4.3.1 Gaussian Mixture Models

In this model, two methods are used to determine changes in gaze angles (i) HoG based method is trained for small changes (ii) GMM is trained for significant changes. (46) proposed a method for tracking 3D gaze with no active illumination. In this author introduces a synthetic iris appearance fitting method to compute 3D gaze direction from the shape of iris. After that, it captures the eye image which fits the best solution. This method claims that it removes the detection problem of unreliable iris contour.

#### 2.4.3.2 Support Vector Machine

(47) used support vector machine with local features to locate eye region using 36 feature points which represent the iris size, position of pupils, and contour of eyes. Gaze direction is classified using support vector machine (SVM) and estimated from 2D coordinates of feature points. In addition to that, In (48) used a method based on Local Binary Pattern Histogram (LBPH) with Principal Component Analysis to extract eye features. To test the accuracy of gaze estimation, they used several classification methods based on Neural Networks, k-Nearest Neighbor, and Support Vector Machine. They claimed to obtain the best accuracy by using LBPH with SVM.

In recent year, methods like convolutional neural networks and deep learning methods have been proposed for eye gaze estimation. (49) created a smartphone app to collect eye images from 1450 participants. This image collection is used to train a convolutional neural network based tracker that can run without any calibration in real time. The CNN is trained with cropped images of the face and eye region. After this, (50) created an extended version of eye gaze dataset and tested a method based on the multi-modal convolutional neural network. This dataset contains different eye appearances and illumination levels of more than 200,000 images. This network learns the mapping between input and output parameters.

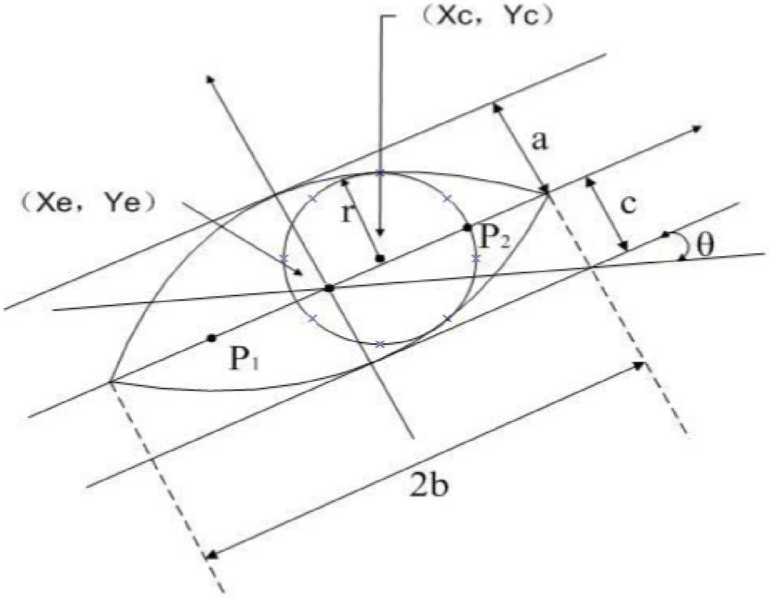


Figure 2.3: deformable template for a human eye. (4)

### 2.4.4 Shape based methods

(4; 51; **?**) used two parabolas for eye contours and a circle to employ deformable templates of the eye region and iris and then fit them to an eye image shown in figure 2.3. In this method, The idea is to find the similarities among images of the chosen region with the template of that region.

The process can be performed by mean square error or modified or normalized (51).

$$S(i,j) =< TxI_T > - < T >< I_T > \div \sigma(T)\sigma(I_T) \tag{2.11}$$

where S(i,j) is similarity measure and I(i,j) and T(u,v) represents the pixel intensity of the image. ¡¿ is an average operator which is a product of pixel obtained by:

$$\begin{aligned} < T > &= 1/n\Sigma T(u,v) \\ < TxI_T > &= 1/n\Sigma T(u,v)I(i+u,j+v) \end{aligned} \tag{2.12}$$

$$\sigma^2(T) = 1/n - 1\Sigma(T(u,v))^2 - < T >^2 \tag{2.13}$$

Here $\sigma$ is the standard deviation of a matched area.

$$S(i,j) = 1/n\sigma(T(u,v) - I(i+u, j+v))^2 \tag{2.14}$$

This provides the similarity measure of mean squared error because cross-correlation is computationally costly.

## 2.5 Path Planning

Nowadays, mobile robots are widely used in many places such as offices and factories, etc. The task of path planning is to find a route to reach the desired destination while avoiding all the obstacles.

### 2.5.1 Path Planning in Static Environment

Path planning in a static environment is to move a mobile robot from start to end position while getting an optimal solution with minimum cost where all the obstacles are stationary.

#### 2.5.1.1 Voronoi Diagram and Fast Marching

Voronoi Diagram and Fast Marching method use two steps for path planning(52). First, it extracts the safest regions and creates a Voronoi diagram, and then Fast Marching method is applied. This method uses parameters for path planning as a sensor frequency.

#### 2.5.1.2 Tube Skeletons Structure and Fast Marching

This method consists of local obstacle avoidance and global motion planning which is based on a nonholonomic path planning(53). In the first step, it models the safest area similar to a Voronoi diagram using a tube skeleton. After that, fast marching method is applied to achieve the best path regarding safety and smoothness. This method uses both non-homonymic constraint and sensor frequency.

#### 2.5.1.3 Voronoi Diagrams and Genetic Algorithms

(54) proposed Voronoi Diagrams and Genetic Algorithms for static path planning. In this method, the Voronoi diagram is used to generate obstacles as points, and genetic algorithm is used to find the best path without collisions. In the genetic algorithm fitness function considers the length, smoothness, and safety of the path.

### 2.5.2 Path Planning in Dynamic Environments

In a dynamic environment, robot navigation is more difficult than a static environment. This is due to the uncertainty of the location of the moving obstacle. The path planning algorithm can be categorized in two cases. At first, when the movement of obstacles is known, the path planning is much easier due to the prior information of the movement. On the other hand, when the movement of obstacles is unknown, optimal path planning is a critical task to perform. (55) used Ant Colony Optimization with Dynamic Voronoi Diagram method. The method uses Dynamic Voronoi Diagram for modeling dynamic environment and then Ant Colony Optimization is applied for finding the shortest path between the start and end position.

In (56) proposed Roadmap-Based Path Planning which creates a roadmap from the Voronoi Diagram. Minimum clearance criteria are used to obtain optimal path. The method finds the overall length, smoothness, and the quality path from obstacles which based on clearance.

Another method was proposed by (57) Multi-agent Navigation Graph data structure by using the Voronoi diagram for Path Planning for multiple robots. In this method, the author used a new data structure made of the first and second order Voronoi Diagrams called Multi-agent Navigation Graph. Multi-agent Navigation Graph performs proximity computations and path planning for each agent in real time.

# Chapter 3

# Experimentation

## 3.1 Performance of Body Pose estimation

### 3.1.1 Aim

The aim of this experiment is to check the how long it takes to localize the human body poses.

### 3.1.2 Results

In order to evaluate the system, multiple tests were conducted by different users. The data was gathered once the system was able to recognize the user as shown in the figure 3.1. The test took place in a known environment, and the same task was done 15 times by three users.

According to the figure 3.3, we can say that the performance of the system varies from user to user but on average it takes around 3.59 seconds to recognize the user and estimate the poses. Out of 15 test, the failure rate of the system is 0.06%.

## 3.2 Accuracy of Pointing Location

### 3.2.1 Aim

The aim of this experiment is to check the accuracy of the location where user's hand and the eye is pointing at.

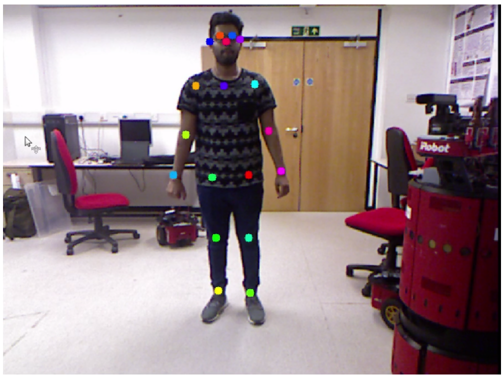


Figure 3.1: Body Pose Estimation

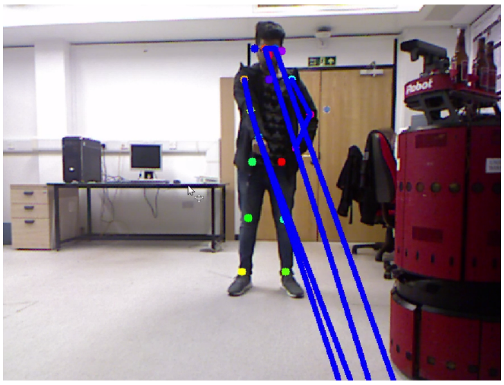


Figure 3.2: Pointing direction of hand and eye gaze

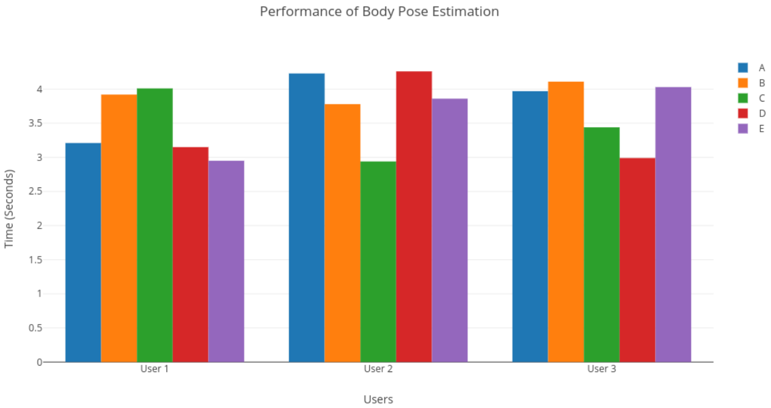


Figure 3.3: Performance of Body Pose Estimation

### 3.2.2 Results

In this experiment, The readings were taken once the system was able to find the intersection of the range of eye gaze and hand direction where the user is pointing as shown in the figure 3.2 and this data was obtained from different users and to evaluate the performance same task was done 15 times. Table 3.1 & 3.2 show the obtained result.

As we can see, the response time from table 3.1 is 9 and 15 seconds for recognizing the gesture and finding the location respectively where the user is pointing. The system was able to locate the gesture and the position 12 out of 15 times. On the other hand, table 3.2 illustrates the error between the actual and user gesture location is 13.33%.

Table 3.1: Response time and recognition rates for different task in 15 tests

| Task | Time (Seconds) | Recognition Rate (%) |
|---|---|---|
| Point at gesture recognition | 9.103 | 80 |
| pointing location estimation | 15.525 | 80 |

Table 3.2: Error rates of wrong pointing location in 15 tests

| Task | Error |
|---|---|
| Wrong pointing location estimation | 13.33% |

### 3.2.3 Performance of global planner

### 3.2.4 Aim

The aim of the experiment was to compare the performance among three global planners. To find the optimal global planner for path planning.

### 3.2.5 Results

In this experiment, three algorithms based global planner were used: A*, Dijkstra, and Navfn. The readings were taken at a target point in a predefined area. The experiment was conducted with multiple static obstacles shown in figure 3.5. The starting and ending position of the robot is marked with the initials S and E respectively and a blue circle. The robot was then instructed to move towards the target position while avoiding obstacles. At the destination, the time was recorded which the robot took to perform this task. The same task was performed ten times for each global planner to obtain the optimal results.

As it can be seen from figure 3.4 and table 3.3 Navfn and Dijkstra based global planner performed way better than A*. On average Dijkstra, Navfn,

Table 3.3: Error rates of different global planners

| Global Planner | Error |
|---|---|
| A* | 33.33% |
| Dijkstra | 0.06% |
| Navfn | 13.33% |

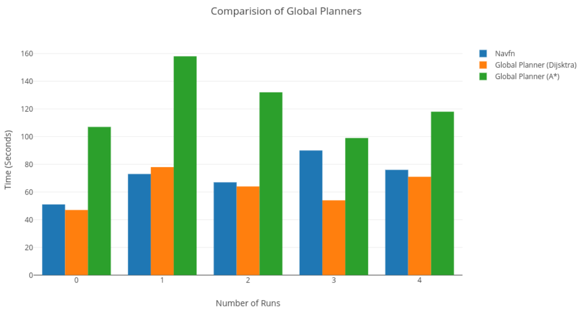


Figure 3.4: Comparison of global planner

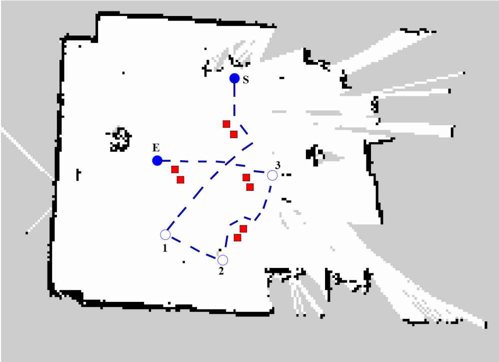


Figure 3.5: Obstacle course for global planner

and A* took 62.8, 71.4, and 112.8 respectively. This route was selected to evaluate the robustness of the system by placing many obstacles in the path of the robot where it had to avoid obstacles and plan the best path. Furthermore, the system failed 33.33% while using A* based global planner whereas Dijkstra and Navfn failed 0.06% and 13.33%.

# Chapter 4

# Conclusion

In this work, a mobile robot system is proposed which is designed to interact with humans in a real-time gesture-based manner. A gesture recognition method is used which is based on the Kinect sensor to understand human body features. The robot is shown an effective way of combining eye gaze and body pose to understand humans.

Furthermore, an extensive set of user tests were conducted to check the robustness of the system. The performance of the system was also improved through different experimentations.

Nevertheless, different elements of the system such as eye gaze estimation and detection of the pointing direction can be improved in future. For example, instead of hand direction, the direction of user's finger can be used to obtain precise pointing location.

In conclusion, the achieved robot state has fulfilled the project outline.

# Bibliography

[1] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

[2] Matthias Straka, Stefan Hauswiesner, Matthias Ruther, and Horst Bischof. Skeletal graph based human pose estimation in real-time. pages 1–12, 2011.

[3] Ioana Bacivarov, Mircea Ionita, and Peter Corcoran. Statistical models of appearance for eye tracking and eye-blink detection and measurement. *IEEE Transactions on Consumer Electronics*, 54(3):1312–1320, 2008.

[4] Wei Wang, Yingping Huang, and Renjie Zhang. Driver gaze tracker using deformable template matching. *Proceedings of 2011 IEEE International Conference on Vehicular Electronics and Safety*, 2011.

[5] Sebastian Thrun. Toward a framework for human-robot interaction. *Human-Computer Interaction*, 19(1):9–24, 2004.

[6] Kanav Kahol, Priyamvada Tripathi, Sethuraman Panchanathan, and Thanassis Rikakis. Gesture segmentation in complex motion sequences. volume 2, pages II–105. IEEE, 2003.

[7] Bin Liang and Lihong Zheng. Multi-modal gesture recognition using skeletal joints and motion trail model. *Computer Vision - ECCV 2014 Workshops*, pages 623–638, 2015.

[8] Necati Cihan Camgoz, Ahmet Alp Kindiroglu, and Lale Akarun. Gesture recognition using template based random forest classifiers. pages 579–594. Springer, 2014.

[9] Elisabeta Marinoiu, Dragos Papava, and Cristian Sminchisescu. Pictorial human spaces: How well do humans perceive a 3d articulated pose? *2013 IEEE International Conference on Computer Vision*, 2013.

[10] A. Gupta, A. Mittal, and L.S. Davis. Constraint integration for efficient multiview pose estimation with self-occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):493–506, 2008.

[11] Jurgen Muller and Michael Arens. Human pose estimation with implicit shape models. pages 9–14. ACM, 2010.

[12] Huazhong Ning, Wei Xu, Yihong Gong, and Thomas Huang. Discriminative learning of visual words for 3d human pose estimation. pages 1–8. IEEE, 2008.

[13] Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3d human pose estimation. *Procedings of the British Machine Vision Conference 2013*, 2013.

[14] Gerard Pons-Moll, David J. Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[15] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.

[16] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016.

[17] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[18] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Personalizing human video pose estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. *International Journal of Computer Vision*, 122(1):149–168, 2016.

[20] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, and Alex Kipman. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840

[21] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[22] Mykhaylo Andriluka and Leonid Sigal. Human context: Modeling human-human interactions for monocular 3d pose estimation. *Articulated Motion and Deformable Objects*, pages 260–272, 2012.

[23] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *Procedings of the British Machine Vision Conference 2016*, 2016.

[25] Yan Tian, Leonid Sigal, Fernando De la Torre, and Yonghua Jia. Canonical locality preserving latent variable model for discriminative pose inference. *Image and Vision Computing*, 31(3):223–230, 2013.

[26] Jigang Liu, Dongquan Liu, Justin Dauwels, and Hock Soon Seah. 3d human motion tracking by exemplar-based conditional particle filter. *Signal Processing*, 110:164–177, 2015.

[27] Suman Sedai, Mohammed Bennamoun, and Du Q. Huynh. A gaussian process guided particle filter for tracking 3d human pose in video. *IEEE Transactions on Image Processing*, 22(11):4286–4300, 2013.

[28] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[29] Vasileios Belagiannis, Xinchao Wang, Bernt Schiele, Pascal Fua, Slobodan Ilic, and Nassir Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. *Computer Vision - ECCV 2014 Workshops*, pages 742–754, 2015.

[30] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1929–1942, 2016.

[31] Kyung Nam Kim and RS Ramakrishna. Vision-based eye-gaze tracking for human computer interface. 2:324–329

[32] Pieter Blignaut. Mapping the pupil-glint vector to gaze coordinates in a simple video-based eye tracker. *Journal of Eye Movement Research*, 7(1

[33] Z.R. Cherif, A. Nait-Ali, J.F. Motsch, and M.O. Krebs. An adaptive calibration of an infrared light device used for gaze tracking. *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.00CH37276).*

[34] X.L.C. Brolly and J.B. Mulligan. Implicit calibration of a remote gaze tracker. *2004 Conference on Computer Vision and Pattern Recognition Workshop.*

[35] Chi Jian-nan, Zhang Chuang, Yan Yan-tao, Liu Yang, and Zhang Han. Eye gaze calculation based on nonlinear polynomial and generalized regression neural network. *2009 Fifth International Conference on Natural Computation*, 2009.

[36] Chunfei Ma, Kang-A Choi, Byeong-Doo Choi, and Sung-Jea Ko. Robust remote gaze estimation method based on multiple geometric transforms. *Optical Engineering*, 54(8):083103, 2015.

[37] Jianzhong Wang, Guangyue Zhang, and Jiadong Shi. 2d gaze estimation based on pupil-glint vector using an artificial neural network. *Applied Sciences*, 6(6):174, 2016.

[38] Andre Meyer, Martin Bohme, Thomas Martinetz, and Erhardt Barth. *A single-camera remote eye tracker.* Springer, 2006.

[39] E.D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006.

[40] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. *Proceedings of the 2006 symposium on Eye tracking research and applications - ETRA '06*, 2006.

[41] Takehiko Ohno and Naoki Mukawa. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. *Proceedings of the Eye tracking research and applications symposium on Eye tracking research and applications - ETRA 2004*, 2004.

[42] S.-W. Shih and J. Liu. A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 34(1):234–245, 2004.

[43] Zhiwei Zhu and Qiang Ji. Novel eye gaze tracking techniques under natural head movement. *IEEE Transactions on biomedical engineering*, 54(12):2246–2260

[44] Xiaolong Zhou, Haibin Cai, Zhanpeng Shao, Hui Yu, and Honghai Liu. 3d eye model-based gaze estimation from a depth sensor. *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2016.

[45] Petros Koutras and Petros Maragos. Estimation of eye gaze direction angles based on active appearance models. *2015 IEEE International Conference on Image Processing (ICIP)*, 2015.

[46] Feng Lu, Yue Gao, and Xiaowu Chen. Estimating 3d gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia*, 18(9):1772–1782, 2016.

[47] Yi-Leh Wu, Chun-Tsai Yeh, Wei-Chih Hung, and Cheng-Yuan Tang. Gaze direction estimation using support vector machine with active appearance model. *Multimedia Tools and Applications*, 70(3):2037–2062, 2012.

[48] Cagatay Murat Yilmaz and Cemal Kose. Local binary pattern histogram features for on-screen eye-gaze direction estimation and a comparison of appearance based methods. *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, 2016.

[49] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[50] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[51] MJT Reinders. Eye tracking by template matching using an automatic codebook generation scheme. pages 85–91, 1997.

[52] Santiago Garrido, Luis Moreno, Mohamed Abderrahim, and Fernando Martin. Path planning for mobile robot navigation using voronoi diagram and fast marching. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[53] Santiago Garrido, Luis Moreno, M Abderrahim, and D Blanco. Robot navigation using tube skeletons and fast marching. pages 1–7. IEEE, 2009.

[54] Facundo Benavides, Gonzalo Tejera, Martin Pedemonte, and Serrana Casella. Real path planning based on genetic algorithm and voronoi diagrams. *IX Latin American Robotics Symposium and IEEE Colombian Conference on Automatic Control, 2011 IEEE*, 2011.

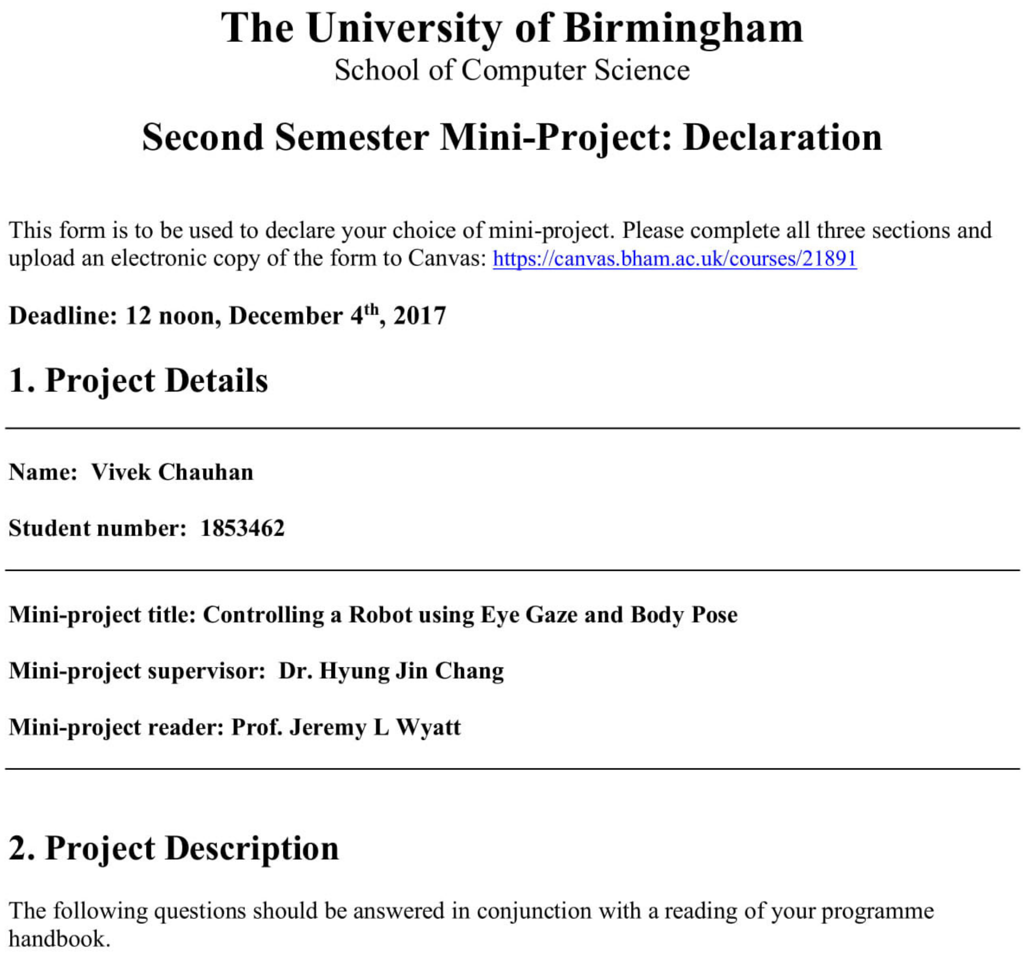
[55] Yaohang Li, Tao Dong, Marwan Bikdash, and Yong-Duan Song. Path planning for unmanned vehicles using ant colony optimization on a dynamic voronoi diagram. pages 716–721, 2005.

[56] Priyadarshi Bhattacharya and Marina Gavrilova. Roadmap-based path planning - using the voronoi diagram for a clearance-based shortest path. *IEEE Robotics and Automation Magazine*, 15(2):58–66, 2008.

[57] A K Pandey and R Alami. A framework towards a socially aware mobile robot motion in human-centered dynamic environment. *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

# Appendix A

# Mini Project Declaration



## The University of Birmingham

School of Computer Science

## Second Semester Mini-Project: Declaration

This form is to be used to declare your choice of mini-project. Please complete all three sections and upload an electronic copy of the form to Canvas: https://canvas.bham.ac.uk/courses/21891

**Deadline: 12 noon, December 4th, 2017**

### 1. Project Details

**Name: Vivek Chauhan**

**Student number: 1853462**

**Mini-project title: Controlling a Robot using Eye Gaze and Body Pose**

**Mini-project supervisor: Dr. Hyung Jin Chang**

**Mini-project reader: Prof. Jeremy L Wyatt**

### 2. Project Description

The following questions should be answered in conjunction with a reading of your programme handbook.

| | |
|---|---|
| **Aim of mini-project** | Control a mobile robot using two algorithms: eye gaze and body pose. |

| | |
|---|---|
| **Objectives to be achieved** | 1. Learn the dynamics of the mobile robot.<br>2. Learn how to control depth camera.<br>3. Understand the navigation in mobile robots.<br>4. Implement the eye gaze algorithm<br>5. Implement the body pose algorithm<br>6. Control robot according to these techniques. |

| | |
|---|---|
| **Project management skills**<br><br>**Briefly explain how you will devise a management plan to allow your supervisor to evaluate your progress** | *1. Gather all the necessary information before starting the project*<br>*2. Start using the depth camera till the end of the month.*<br>*3. Implementation of gaze estimation algorithm (till February)*<br>*4. Implementation of body pose algorithm (till March)*<br>*5. Send a weekly report* |

| | |
|---|---|
| **Systematic literature skills**<br><br>**Briefly explain how you will find previous relevant work** | *1. I will start searching about the relevant work from different journal publishers such as: CVPR, IROS, TRO, and ICRA etc.*<br><br>*2. I will also use search tools like Google Scholar and Medeley for paper reviews.* |

| | |
|---|---|
| **Communication skills**<br><br>**What communication skills will you practise during this mini-project?** | *1. From this mini project, I'm expecting to improve my communication skills for scientific research collaboration.*<br>*2. I have to prepare meeting topics based on priorities because of the meeting time with the supervisor is limited.*<br>*3. I also have to learn about how to report my progress efficiently.* |

## 3. Project Ethics Self-Assessment Form

**Please answer YES/NO to the following questions:**

- **Does the research involve contact with NHS staff or patients?**
  **NO**

- **Does the research involve animals?**
  **NO**

- **Will any of the research be conducted overseas?**
  **NO**

- **Will any of the data cross international boarders?**
  **NO**

- **Are the results of the research project likely to expose any person to physical or psychological harm?**
  **NO**

- **Will you have access to personal information that allows you to identify individuals, or to corporate or company confidential information (that is not covered by confidentiality terms within an agreement or by a separate confidentiality agreement)?**
  **NO**

- **Does the research project present a significant risk to the environment or society?**
  **NO**

- **Are there any ethical issues raised by this research project that in the opinion of the PI or student require further ethical review? If you are unsure, consider whether the project has the potential to cause stress or anxiety in the people you are involving.**
  **NO**

- **Human subjects can be involved as users, providers of system requirements, testers, for evaluation, or similar such activities. Does the experiment involve the use of human subjects in any other capacity? If you are unsure, answer YES.**
  **NO**

- **Answer YES if ANY of the following are true**

  *** the project has the potential to cause stress or anxiety in the people you are involving, e.g. it addresses potentially sensitive issues of health, death, religion, self-worth, financial security or other such issues**
  **NO**

  *** the project involves people under 18**
  **NO**

*** the project involves a lack of consent or uninformed consent**
**NO**

*** the project involves misleading the subjects in any way**
**NO**

**If the project's involvement of people relates only to straightforward information gathering, requirements specification, or simple usability testing, then you can indicate NO.**

- **If any of the above questions is answered YES, or you are unsure if further review is needed (the first point is usually a good indicator - may cause stress or anxiety) then you should refer it for review.**

- **Further review will involve the School Ethics Officer meeting with the supervisor and ideally the student, reviewing the project, and suggesting any procedures necessary to ensure ethical compliance.**

## DECLARATION

By submitting this form, I declare that the questions above have been answered truthfully and to the best of my knowledge and belief, and that I take full responsibility for these responses. I undertake to observe ethical principles throughout the research project and to report any changes that affect the ethics of the project to the University Ethical Review Committee for review.

# Appendix B

# Statement of information search strategy

| | |
|---|---|
| Databases searched | IEEE Xplore<br>Mendeley<br>Google Scholar |
| Part of journals searched | Title<br>Abstract<br>Chapter Headings<br>Conclusion |
| Language | English |
| Journal Publishers | CVPR<br>ICRA<br>TRO<br>IROS |