國立臺灣大學工學院機械工程學研究所

碩士論文

Department of Mechanical Engineering

College of Engineering

National Taiwan University

Master Thesis

多感測模態機器人之人物辨識及跟隨

Human Recognition and Following of a Robot with Multi-Sensing Models

郭俊傑

Chon-Kit Kuok

指導教授：黃漢邦 博士

Advisor: Han-Pang Huang, Ph.D.

中華民國 105 年 7 月

July 2016

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

國立臺灣大學碩士學位論文
口試委員會審定書

多感測模態機器人之人物辨識及跟隨

Human Recognition and Following of a Robot with Multi-Sensing Models

本論文係郭俊傑君（R03522838）在國立臺灣大學機械工程學系完成之碩士學位論文，於民國 105 年 7 月 15 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

黃漢邦（簽名）

（指導教授）

趙儀珊　林顯達

林沛群　黃心健

系主任　楊燿州（簽名）

doi:10.6342/NTU201602427

Department of Mechanical Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.

Date: July 15, 2016

We have carefully read the thesis entitled

"*Human Recognition and Following of a Robot with Multi-Sensing Models*"

submitted by Chon-Kit Kuok in partial fulfillment of the requirement of the degree of **MASTER OF SCIENCE** and recommend its acceptance.

Yee-San Teoh

Advisor:

Chairperson of
Department of Mechanical Engineering:

doi:10.6342/NTU201602427

誌謝

能夠順利地從碩士畢業，首先感謝父母對我的支持及諒解，能讓我沒有後顧之憂地專心學業。很感謝指導教授 黃漢邦 教授，對學生在課業上的指導與生活上的關心，給予我機會參加計畫，使我在這兩年的研究生活中學到了很多未來在研究上以至工作上的寶貴經驗。感謝 趙儀珊 教授、林沛群 教授、黃心健 教授以及 林顯達 教授撥冗擔任我的口試委員並給予我許多寶貴的意見。

感謝大威志學長在研究上提供我很多的想法。感謝冠禎學長提供我各種研究上所需的知識。感謝良健學弟在實驗上的各種協助。感謝實驗室的各位學長及學弟們在大小事上的關照。最後感謝尚庭、貽仁、建興、劍航、世全、鼎元、逸祥在我兩年研究所生活的參與及關照，有了你們的參與，使我在研究上以至生活上，除了令我獲益良多，還令我留下了美好的回憶以及經驗。

感謝所有曾經幫助我的人並獻上祝福。

郭俊傑

臺大機器人實驗室

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

摘要

近年來，機器人已經融入人類生活中，而服務型機器人的發展是為了讓人類的生活更舒適便利。與此同時，機器人也不只是單純的機器，機器人應該具備與人互動以及在環境順暢移動的能力。本篇論文旨在發展一個跟隨機器人，此機器人能夠跟隨特定使用者，並能自動搜尋此使用者。我們使用 SLAMMOT 使機器人能了解自身週遭環境位置，以及使用者的位置。當機器人丟失跟隨目標時，機器人能尋找使用者，令使用者能重新進入機器人的視線後重新跟隨。

本篇論文也提出使用者的辨認方法，讓機器人能夠辨識特定使用者，使機器人不會認錯要跟踪的使用者。所建構的跟隨機器人系統，經實驗驗證具有不錯的效果。

關鍵字: 使用者辨識、人機互動、跟隨、行動式機器人、多模態融合

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

Abstract

In recent years, robots are assimilated into humans' life. Service robots are built for making people more convenient and more comfortable. A robot is not just a machine. It should be capable of interacting with people appropriately and moving in the surroundings smoothly. This thesis attempts to develop a shopping assistant robot system. It can recognize a specific user, automatically follow him and search for him. SLAMMOT is used in the robot for understanding environments, localizing itself and locating the user. If the robot loses the tracking user, it searches for the user and causes the user can be found in the view of the robot. In addition, the human recognition system is constructed to recognize the specific user. The proposed system has been justified in experiments. The results are promising.

Keywords: User Recognition, Human-Robot Interaction, Following, Mobile Robot, Multimodal Fusion

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

Contents

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

List of Tables

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

List of Figures

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

Chapter 1 Introduction

The motivation of this thesis is explained in this chapter. The related works of the service robots and the integrated system of service robots will be introduced.

1.1 Motivation

Industrial robots have been widely used in factories. Through controller and language development, sensing, and drive systems of the robot system have been greatly enhanced. Various robots have been developed to work in indoor or outdoor environments for more than decades [25, 43, 46]. The research of personal robots or social robots keep progress. Building an intelligent service robot is one of the main tasks in the field of robotics. Service robots act like assistants. They can take care of the elders or people with reduced mobility.

The human following robot is the main issue of this thesis. This type of robots can be widely used as a shopping assistant [11, 22], self-propelled baggage [3], etc. Sometimes, the robot will lose the tracking target. It should have the ability of re-finding the specific target. The user can do some interaction with the following robot to help the robot re-find the user.

doi:10.6342/NTU201602427

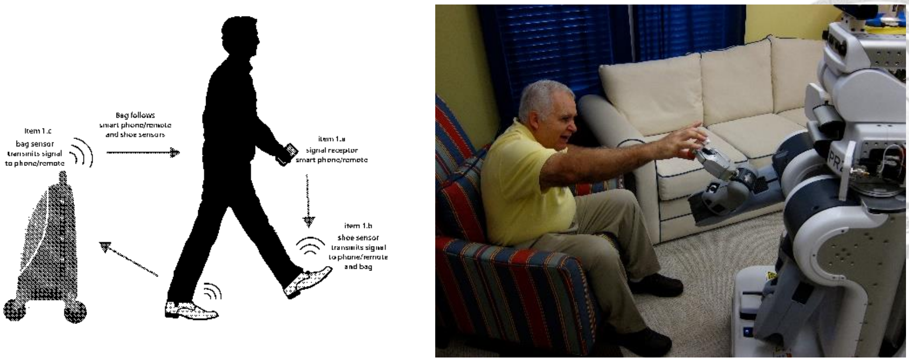


Figure 1-1 Applications of the assistant robots [3] [6]

This thesis aims to develop a following robot so that it can recognize a specific user, following him, and re-find him after losing him.

1.2 Related Works

This section summarizes previous research about the field of service robots and Human-Robot-Interaction. Service robots are developed for various usage in shopping malls [24, 46, 51], museums [10, 26, 47], households [54-56], hospital [15, 19, 21], etc. They are capable of perceiving and reacting to their surroundings, avoiding obstacles near them, knowing their positions in their working spaces, and going to the target place. In order to interact with a pedestrian, some service robots are designed with human-like appearance or communication abilities, such as voice system and expression system. It makes them more friendly when the robots interact with humans.

Human following robots have been developed with many sensors and effective

doi:10.6342/NTU201602427

algorithms. They are built in vision-based, generally. Various Cameras are used on the following robot [34, 41], and laser range finder [20, 57] is a must.

In order to increase the accuracy of the sensor data, the sensor fusion algorithm proposed by Luo et al. [31] and implemented by Chen [11] are applied.

Impedance Control implemented by Chen [11] is applied to this thesis for solving the interaction problems between the user and the robot. The robot needs a higher speed to follow the user when the user walks faster. In order to recognize the specific user precisely, multimodal biometric recognition technology [29, 39, 40] is implemented to identify the specific person.

Moreover, human following robots sometimes lose the following target when the target turns into other directions. OTA et al. [37, 38] proposed a recovery function that the human following robot can re-find the user after the user makes a turned at the corner. The robot can predict the path where the user has walked. Lai [28] proposed a re-tracking scenario that the robot turns left or right according to the disappeared direction of the user.

1.3 Objectives and Contributions

Human following robots can transport goods so that the master (viz. human) may

doi:10.6342/NTU201602427

not take the goods. This function can make more convenience to humans. Our robot, Bunny, can achieve it. However, to recognize the specific user by using the color-based particle filter is not robust. Multimodal biometric recognition [23, 29] is provided for more robust recognition.

Combined with the techniques of user recognition, our robot can re-track the specific user accurately after the robot loses the tracked user. If the user is not found finally, the user can do something to interact with the robot, such as gesture, body posture, sound, etc. Many methods of interaction can be implemented in Human-Robot Interaction tasks.

The contributions of this thesis are summarized as follows:

- This thesis develops an amiable human following robot, to make the user feel secure and familiar, and a basket for carrying.
- In order to recognize the specific user accurately, multimodal user recognition algorithm is implemented.
- The robot can re-find the specific user after the user is lost from the vision of the robot. The behaviors of the robot are provided in this thesis.

1.4 Thesis Organization

The thesis organization is given in **Figure 1-2**.

doi:10.6342/NTU201602427

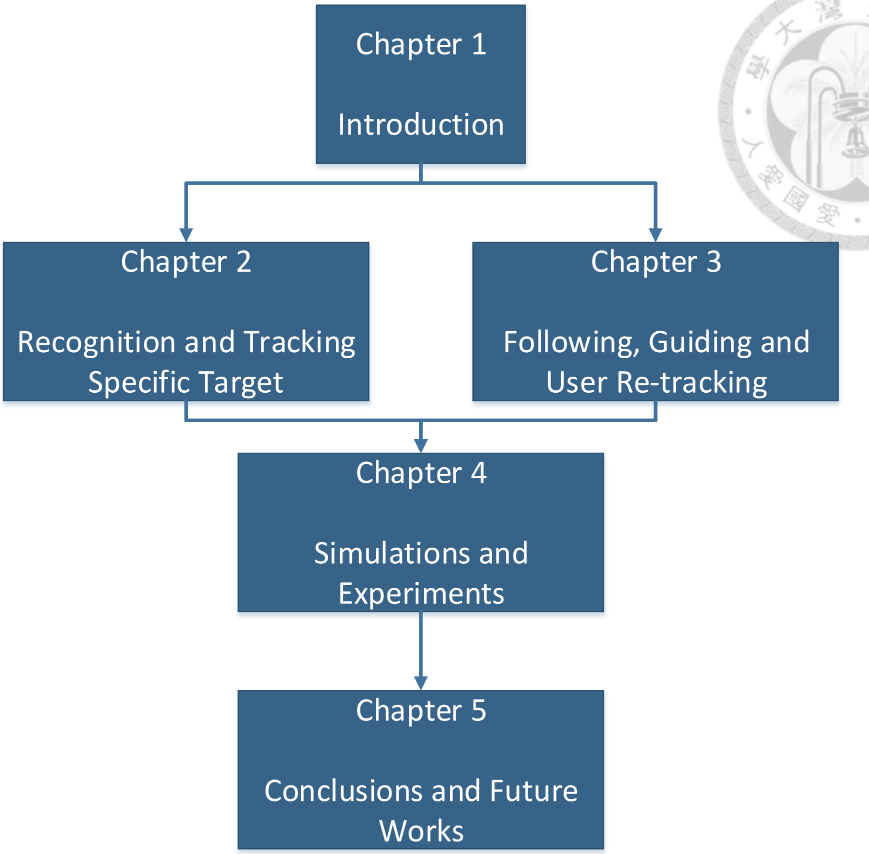

Chapter 1
Introduction
Chapter 2
Recognition and Tracking Specific Target
Chapter 3
Following, Guiding and User Re-tracking
Chapter 4
Simulations and Experiments
Chapter 5
Conclusions and Future Works

Figure 1-2 The organization of thesis

Chapter 2 briefly introduces the existing human following algorithm developed by Chen. A structure of the sensor fusion method is applied for tracking and calculating the relative position of the user based on the robot. Chapter 2 will also introduce the method of the specific user recognition by using external biometric features such as face, body size, and dress color. A Multimodal fusion algorithm is applied for combining the different modalities to improve the accuracy.

Chapter 3 discusses the human following system briefly. It introduces the behavior

doi:10.6342/NTU201602427

of the user re-tracking. If the user disappears from the sight of the robot, the robot does something to re-track the user and follow the user again.

Chapter 4 conducts a series of experiments and shows the results of the interactions between the user and the robot. The experiment scenario can be easily observed in our daily life.

Chapter 5 addresses conclusions and future works.

doi:10.6342/NTU201602427

Chapter 2 Recognition and Tracking a Specific Target

There are great amounts of research about human tracking [30-32]. However, following a specific moving target is a difficult task. Several important factors considered are presented as follows.

First, since the person and the robot are both moving in the real world, we need not only to recognize the relative position between the target and the robot but also to perceive the environmental information. This thesis utilizes SLAMMOT algorithm proposed by Wang et al. [53] and implemented by Chung [12] to acquire the information of the environment and find out the humans in the map.

Second, in view of the inevitable uncertainty caused by any kind of sensors, using sensor fusion to improve the sensor accuracy is reasonable. For example, laser range finder is very useful in building a map and recognizing the moving object in the map, But, it does not contain the visual information. Kinect has depth and image information, but it has narrow sensing range compared to laser range finder.

Third, in the general case, servant robots recognize a specific target by face recognition [7], or track dress color histogram by particle filter [36]. However, it may recognize the wrong target when features are similar. The multimodal fusion [23, 29]

doi:10.6342/NTU201602427

will be used to improve the modality accuracy. For example, face recognition can find out a specific person but it does not contain the body information. The classifier will make a mistake when it recognizes the similar face such as human twins. Dress color histogram contains the color information, but it does not contain the face information. The classifier makes a mistake if someone wears dresses with the similar color.

Biometric identification is a convenient way to accomplish the task since they do not require individuals to carry authentication tokens (e.g. keys, cards) or remember usernames and passwords. They can be divided into two approaches: active and passive. An active biometric system requires the user to interact in some way with an interface. This is the case of fingerprint recognition. In passive biometrics systems, the user is not required to directly interact with the system. Facial and voice recognition fall into this category and are becoming increasingly common in daily tasks.

In summary, the use of multiple sensors can improve the accuracy of the multimodal system for a robust specific user recognition. It allows robots to have a robust user recognition for Human-Robot Interaction.

2.1 Human Following Algorithm

This section shows the entire structure of the human tracking system. This thesis

doi:10.6342/NTU201602427

utilizes a similar system implemented by Luo et al. [31] and Chen [11]. After modifying the part of the vision-based tracking system, the structure of human following is shown in **Figure 2-1**.

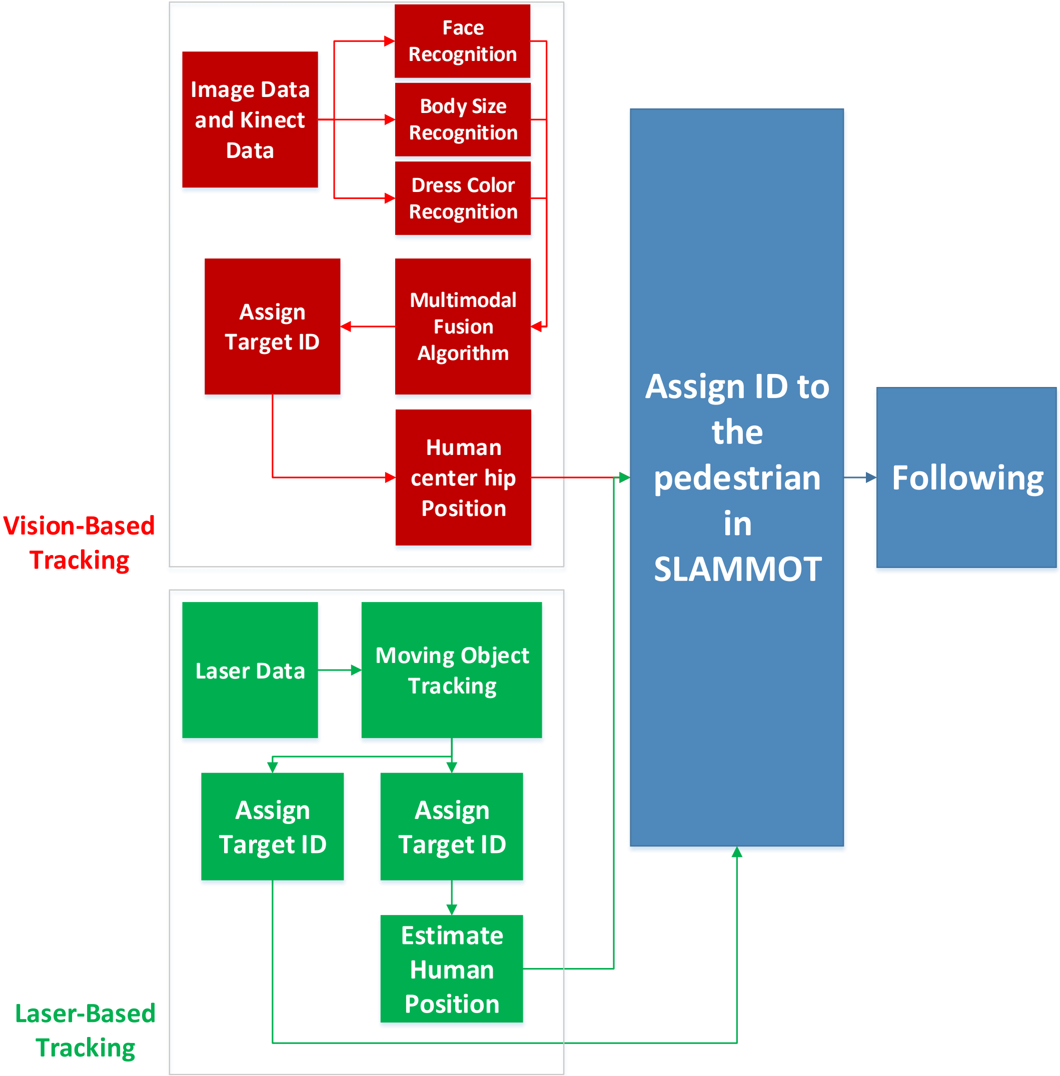


Figure 2-1 Structure of human following

There are three parts in the human following system (see **Figure 2-1**). The red part denotes the vision-based tracking system. The green part denotes the laser-based

doi:10.6342/NTU201602427

tracking system. The blue part denotes the part of the sensor fusion algorithm and motion planning. The vision-based tracking system is used to track and recognize the specific user. The laser-based tracking system is used to track the pedestrians in the map, which is built by SLAMMOT. After the specific user is recognized, the corresponding ID is assigned in SLAMMOT. The target position data tracked by Kinect and laser range finder is sent to the part of the sensor fusion algorithm, which is proposed by Luo et al. [31] and implemented by Chen [11]. After fusing the information and assigning ID to the pedestrian in SLAMMOT, the robot operates the motion planning mission.

The sensors used for tracking the user in this thesis are Kinect and laser range finder. The information captured from the sensors are fused and calculate a goal behind the user.

Kinect is a universal and cheap RGB-D camera (see **Figure 2-2**). It combines image with depth data. We can develop some applications by using official SDK. After recognizing the target user, the Kinect is used for tracking the center of hip of the target user.

doi:10.6342/NTU201602427

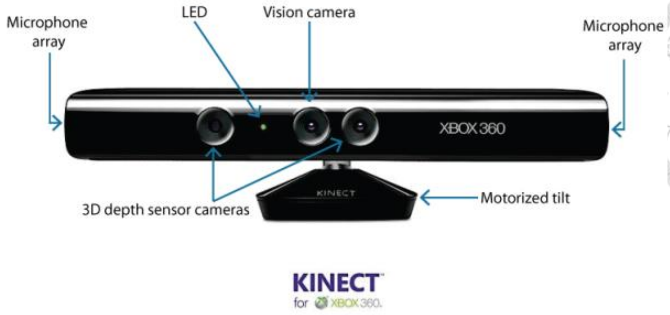


Figure 2-2 Kinect v1.0 sensor

There are total 20 skeleton joints, which can be tracked by Kinect. The skeleton picture is shown in **Figure 2-3**. We can define the hip center of the skeleton data as a tracking joint.

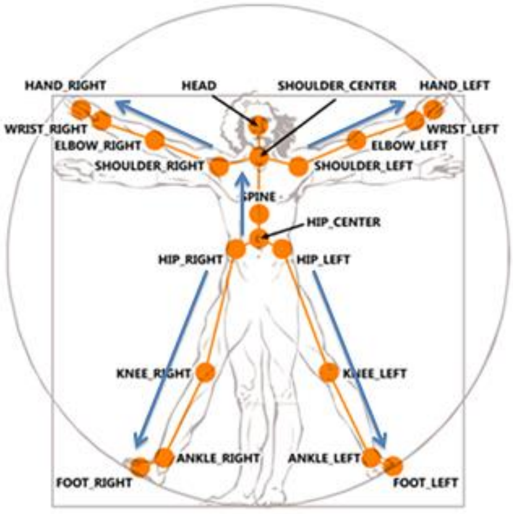
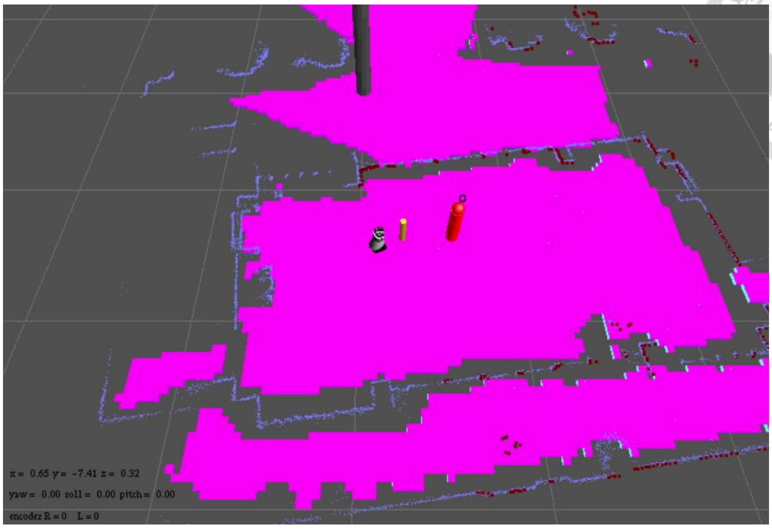


Figure 2-3 Skeleton Joints tracked by Kinect [5]

Laser range finder can capture the obstacle information around the robot. SLAMMOT proposed by Wang [53] and implemented by Chung [12, 13] can build the map and locate the robot simultaneously, and track the moving object. The real

doi:10.6342/NTU201602427

operation of SLAMMOT is shown in **Figure 2-4**.



```
x =  0.65 y =  -7.41 z =  0.32
yaw =  0.00 roll =  0.00 pitch =  0.00
encoder R = 0   L = 0
```

Figure 2-4 Operating process of SLAMMOT [12, 13]

In **Figure 2-4**, the robot can locate itself on a map. Simultaneously, it can track the pedestrian in the map by recognizing the tracked specific shape of laser points. A red cylinder shown in **Figure 2-4** denotes the tracked pedestrian. A yellow cylinder denotes the goal of the robot.

2.2 Multimodal User Recognition

In order to ensure the accuracy of the user identification system, there are three modalities are implemented in the system. The three modalities are facial recognition, body size recognition, and dress color recognition.

Two variables will be discussed in user identification: False Acceptance Rate and False Rejection Rate. False Acceptance Rate denotes the acceptance rate of a non-

doi:10.6342/NTU201602427

Target. False Rejection Rate denotes the rejection rate of a True-Target. The mathematical function of False Acceptance Rate and False Rejection Rate are shown in **Table 2-1**.

Table 2-1 The relations between truth/falseness of the hypothesis and outcomes

Condition Predicted	Accept	Reject		
Accept	True Positive	False Positive (Type I error)	True Positive Rate = $\frac{TP}{TP+FP}$	False Positive Rate = $\frac{FP}{FP+TN}$
Reject	False Negative (Type II error)	True Negative	False Negative Rate = $\frac{FN}{FN+TP}$	True Negative Rate = $\frac{FN}{FN+TN}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}\times 100\%$			

where TP denotes true positive, TN denotes true negative, FP denotes false positive and FN denotes false negative.

The definitions in Biometric identification [58], False Acceptance Rate is defined as Type I error and False Rejection Rate is defined as Type II error. It denotes the False Acceptance Rate is equals to False Positive Rate, and False Rejection Rate is equals to False Negative Rate.

2.2.1 Face Modality

Face Recognition is an easy task for humans. In order to recognize a specific user, face recognition should be included. In this part, OpenCV function [1, 27] is implemented. There are two processes for this function: Face Detection and Face Recognition.

Haar feature-based cascade classifiers proposed by Viola et al. [52] is applied for face detection. It is a machine learning based approach where a cascade function is trained from many positive and negative images. Each feature is a single value obtained by the subtracting sum of pixels under white rectangle from the sum of pixels under black rectangle (see **Figure 2-5**).

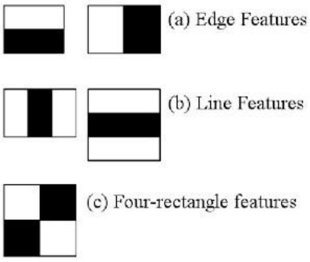


Figure 2-5 Haar Features

After the faces are detected from images, face recognition function can be initiated. The recognition is performed by calculating the Euclidean distance between feature vectors of a probe and reference images. Such a method has a huge drawback. The accurate registration of the marker points is complicated. We should decrease the number of dimensions of the image. In general, eigenfaces method, fisherfaces method and Local Binary Patterns Histograms method are used in face recognition.

doi:10.6342/NTU201602427

Fisherfaces method is implemented in this modality. The Principal Component Analysis (PCA) [42], which is the core of the eigenfaces method, finds a linear combination of features that maximizes the total variance in data. The Linear Discriminant Analysis performs a class-specific dimensionality reduction.

A dataset with 10000 facial images with size 150×150 pixels is created. This dataset becomes the template for face recognition. Part of facial images in the dataset is shown in **Figure 2-6**. Each facial image contains a label: Accept and Reject.

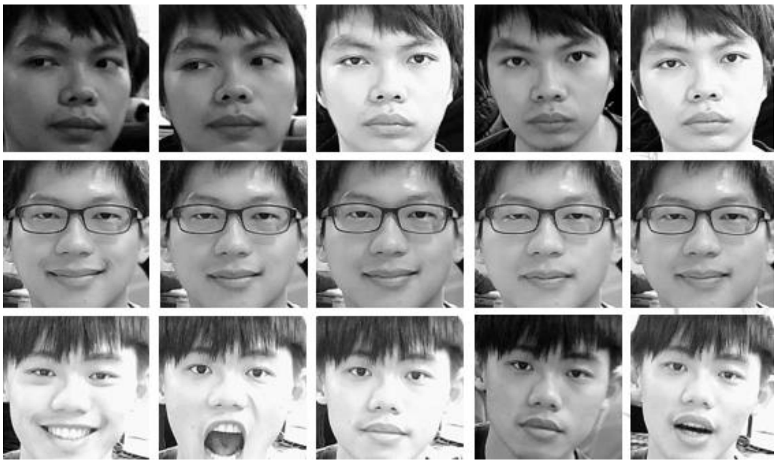


Figure 2-6 Face dataset

In order to find the combination of features that separates best between classes, the Linear Discriminant Analysis maximizes the ratio of between-classes to within-classes scatter, instead of maximizing the overall scatter. The idea is simple. Same classes should cluster tightly together, while different classes are as far away as possible from each other in the lower-dimensional representation.

doi:10.6342/NTU201602427

The Fisherfaces method learns a class-specific transformation matrix, so they do not capture illumination as obviously as that in the eigenfaces method. The Discriminant Analysis instead finds the facial features to discriminate between people. The algorithms are shown below.

Let X be a random vector with samples drawn from c classes.

$$X = \{X_1, X_2, \ldots, X_c\} \quad \textbf{(2.1)}$$

The scatter matrices S_B and S_w are calculated as.

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \quad \textbf{(2.2)}$$

$$S_w = \sum_{i=1}^{c} \sum_{x_j \in X^i} (x_j - \mu_i)(x_j - \mu_i)^T \quad \textbf{(2.3)}$$

where μ is the total mean.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \textbf{(2.4)}$$

And μ_i is the mean of class $i \in \{1, \ldots, c\}$.

$$\mu_i = \frac{1}{|X_i|}\sum_{x_j \in X_i} x_j \quad \textbf{(2.5)}$$

Fisher’s classical algorithm now looks for a projection W, that maximizes the class separability criterion. We have

doi:10.6342/NTU201602427

$$W_{opt} = \arg\max_{w} \frac{|W^T S_B W|}{|W^T S_W W|} \tag{2.6}$$

A solution for this optimization problem is given by solving the following general eigenvalue problem:

$$S_B v_i = \lambda_i S_w v_i \tag{2.7}$$

$$S_W^{-1} S_B v_i = \lambda_i v_i \tag{2.8}$$

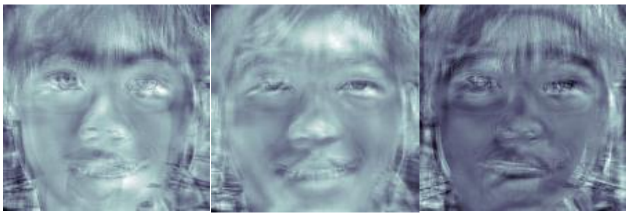


Figure 2-7 Dimension Regression by FisherFaces Method

The results of lower-dimensional representation is shown in **Figure 2-7**. The important features are kept after the dimension regression, such as the shape of eyes, nose, etc. The dimension regression can help to keep the speed of computing.

Face Recognizer function of OpenCV is implemented in this modality. The dataset which is shown above is obtained from the members of our laboratory. The matching approach of the face modality is nearest neighbor algorithm. The number of training images affects the accuracy of the modality. The recognition rate is shown in **Figure 2-8**.

doi:10.6342/NTU201602427

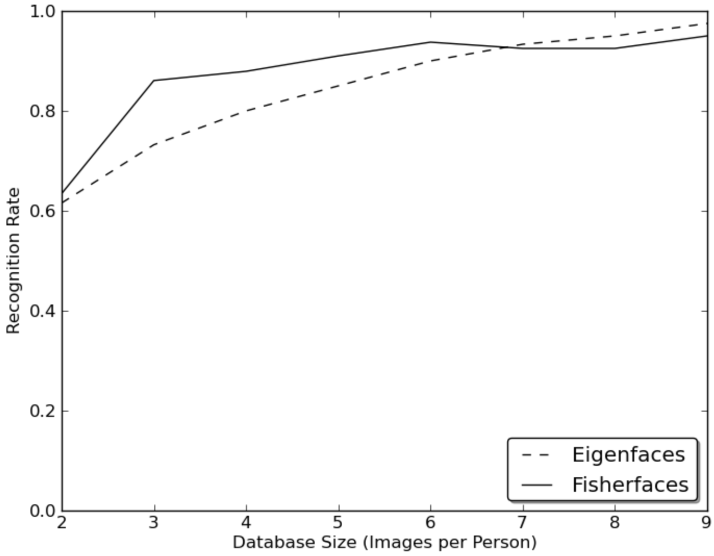
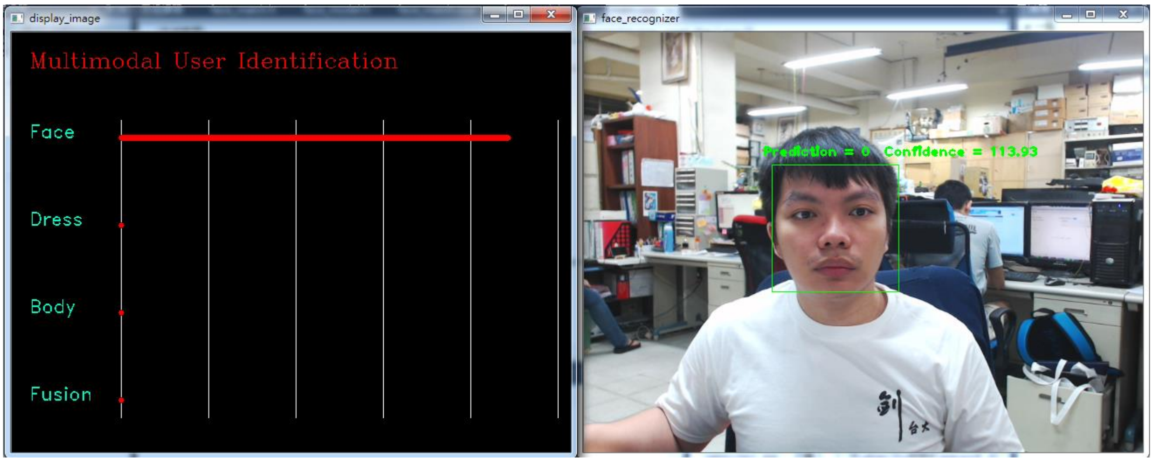


Figure 2-8 Recognition rate of facial modality

According to the **Figure 2-8**, the recognition rate increases when the number of training images per person increases. Therefore, we collect 15 training images per person and there are 10 people in our dataset.



Prediction = 0 Confidence = 113.93

Figure 2-9 Operating interface in face modality

Two information can be obtained: predicted label and confidence. In **Figure 2-9**, a target user is identified in label 0, which denotes Accept. The other label denotes

Reject. A Euclidean distance of measurement is obtained for the multimodal fusion modality. In the function of OpenCV, six people can be detected simultaneously in a frame of the image.

The threshold is the maximum value of the Euclidean distance between the database image and the input image. The input face image is recognizes in a label if the measured distance is lower than the threshold. The subjects were asked to show their faces 10 times in the camera. The online test results are shown in **Table 2-2** and **Table 2-3**.

Table 2-2 Threshold testing for Target

Class Threshold	Correct Recognition	False Recognition
1250	10	0
1000	10	0
750	8	2
500	5	5

It is observed that for the known target in **Table 2-2**, the threshold value in the range 1000~1250 gives us a 100% accuracy. The threshold value in the range 1000~750 gives us an 80% accuracy. When the threshold value decreases, the false recognition of the person increases.

Table 2-3 Threshold testing for Non-Target

Class Threshold	Correct Recognition	False Recognition
1250	10	0
1000	10	0
750	10	0
500	9	1

For non-Target case in **Table 2-3**, the threshold range 1250~1000 gives us 100% accuracy, but the accuracy decreases when the threshold range decreases to 750.

According to the results in **Table 2-2** and **Table 2-3**, the accuracy decreases when the threshold is set at 1000. Therefore, we set the threshold in 1000.

2.2.2 Body Size Modality

The sizes of body parts are unique for a human. They can be the unique features to recognize the specific person. Sinha et al. [48] proposed an approach, that captured the skeletal size and gait cycle data to be the features. Artificial Neural Network is one of the classifiers in their approach. Barbosa et al. [9] proposed an approach that the size of skeleton and contour of the user are the features for the classifier.

In the approach of this thesis, the selected features are referred by the approach of Araujo et al. [8]. This data was captured while subjects walked in front of the sensor.

doi:10.6342/NTU201602427

The Kinect provides frames that contains depth information from its sensors, at a rate of 30 frames per second. In this section, skeleton information is implemented. The skeleton information is shown in **Figure 2-10**.

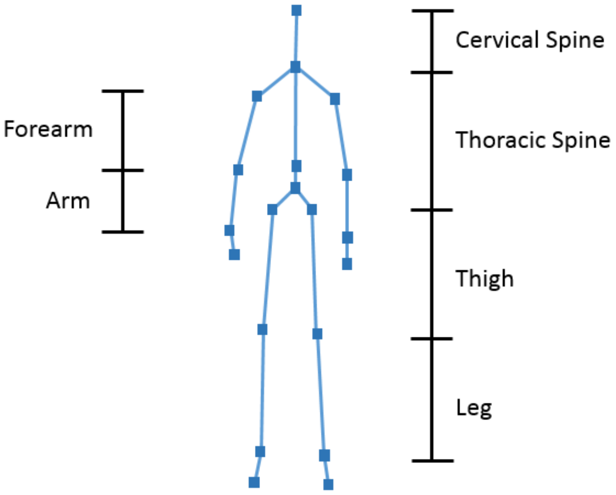


Figure 2-10 Skeleton joints returned by Kinect SDK and defined body parts

For each frame, the length of different parts of the body based on the coordinates provided by the sensors are calculated.

The distance between two skeleton joints are calculated as

$$d_{i,i+1} = \sqrt{(P_{x_i} - P_{x_{i+1}})^2 + (P_{y_i} - P_{y_{i+1}})^2 + (P_{z_i} - P_{z_{i+1}})^2} \quad \textbf{(2.9)}$$

where P denotes the joint of the skeleton in the real world and i denotes the id of the skeleton joints defined by Kinect SDK. It contains 20 joints. They can be found by the captured data of the Kinect.

Height can be calculated by the following equation.

doi:10.6342/NTU201602427

$$d_{height} = d_{cervical\_spine} + d_{thoracic\_spine} + \frac{(d_{left\_leg} + d_{right\_leg})}{2} + \frac{(d_{left\_thigh} + d_{right\_thigh})}{2} \quad \textbf{(2.10)}$$

Namely, height is calculated by summing up the Cervical Spine, Thoracic Spine, the mean between left and right legs, and the mean between left and right thighs. It must be noted that, these attributes are not independent of each other.

However, the measured distance between the joints are different when the distance between the tracked user and robot is different. The skeleton size information cannot total implemented in real application. The ratio of limbs become the features. The 12 features implemented in this modality are shown as follows.

- Thoracic spine
- The length of left arm and right arm (the length of forearm + the length of upper arm)
- The length of left leg and right leg (the length of thigh + the length of lower leg)
- The ratio of forearm and forearm (included left and right)
- The ratio of thigh and lower leg (included left and right)
- The ratio of leg and thoracic spine (included left and right)
- Height

The output of the person identification system includes: Accept and Reject. The labels for supervised learning system are defined as accept and reject. Two learning classifiers are tested in this modality: Multi-layer Perceptrons (using Backpropagation,

doi:10.6342/NTU201602427

with 10 hidden units) and Support Vector Machine.

Multi-layer Perceptron [44] is one of the methods of deep learning. It can separate a set of data that is linearly inseparable. It is widely used in voice recognition, image recognition, etc. However, it spends a lot of time in training. Support Vector Machine [14, 16] is implemented for a classifier in this modality. SVM is good at two-class recognition and training datasets with less time. It can process the small dataset. The advantages of support vector machines are:

- Effective in high dimensional spaces
- Still effective in cases where the number of dimensions is greater than the number of samples
- Uses a subset of training points in the decision function, to make memory efficient.
- It is versatile and different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

A set of the offline testing result is shown in **Table 2-4**. Libsvm-3.16 library, which is provided by Chang et al. [2], is implemented in this modality. The type of SVM in this modality is C-Support Vector Classification (C-SVC). The kernel of SVM in this

doi:10.6342/NTU201602427

modality is RBF kernel. After cross-validation, the training variables of SVM are: cost=32, gamma=8.

In order to make the differences between the features clearly, the normalization of feature vectors is implemented. The equation of normalization implemented in this modality is shown as

$$X'_i = 2 \times \frac{(X_i - X_{i\_\min})}{(X_{i\_\max} - X_{i\_\min})} - 1 \qquad \textbf{(2.11)}$$

where X'_i denotes the normalized feature value, X_i denotes the raw feature value, $X_{i\_\min}$ and $X_{i\_\max}$ denote the minimum value and maximum value of feature, respectively. i denotes the index of feature vectors. The feature values are mapped into the range -1 to 1.

There are 27000 pieces of data for training and 8000 pieces of data for testing. The testing confusion matrix is shown in **Table 2-4**.

Table 2-4 Confusion matrix in offline test of body modality

Condition Predicted	Accept	Reject
Accept	**1324**	234
Reject	165	**6295**
Accuracy	**95.0237%**	

Table 2-4 shows the confusion matrix and accuracy in the offline test. The

accuracy of the body size modality is 95.024%. The False Acceptance Rate (False Positive Rate) is 0.1108 and the False Rejection Rate (False Negative Rate) is 0.026.

In the online test, the user is asked to stand forward of the Kinect sensor. The robot recognizes whether the user is a target or not. If the target is recognized, a high grade value is returned. If a wrong target is recognized, a low grade value is returned.

2.2.3 Dress Color Modality

There are many types of research for the human following robot to recognize the dress color of the specific user. In general, color-based particle filter [18, 36], SIFT [45] or SURF [17] are used for tracking the color of users' dresses. However, the algorithms spend a lot computing power. We should save the computing power for operating the multi-modal system.

Sugiyama et al. [49, 50] proposed a simple personal identification method using Dress Color Information for a guide robot. Unlike the general approaches such as particle filter or SURF, this approach is only calculated at narrow areas around a user's joint position obtained by Kinect. This method can identify the specific user in real time.

There are total 20 skeleton points and 20 narrow areas for matching the color of dresses (see **Figure 2-11**).

doi:10.6342/NTU201602427

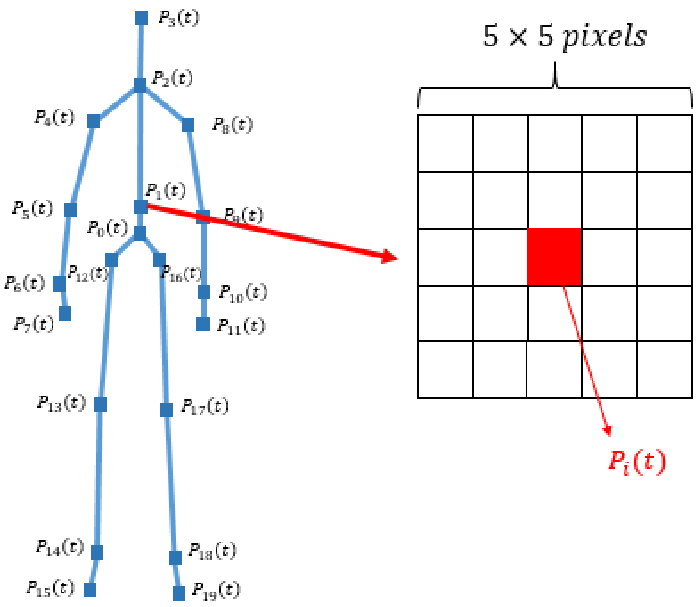


Figure 2-11 Definition of Personal Information in the current time t without height information and shoulder width information

There are 25 pixels in each narrow area of skeleton joints. The center of narrow area (shown in red square) denotes the joint of skeleton. The image of narrow area is extracted for the dresses matching. There are 20 skeleton joints in a human, as there are 20 dress color images for the dresses matching.

The matching scores of dresses matching $Similarity(t)$ is calculated by cosine similarity method, and is given as

$$Similarity(t) = \frac{Z(t)^T Z(0)}{|Z(t)||Z(0)|} \tag{2.12}$$

where $Z(t)$ denotes the feature vector of the current time t, and $Z(0)$ denotes that of the initial time (t=0).

doi:10.6342/NTU201602427

The color histogram information, height and shoulder length are elements in the feature vector. The feature vector is represented as

$$Z(t)=\left[P_0(t),P_1(t),...,P_{19}(t),H(t),W(t)\right]^T \tag{2.13}$$

where $Z(t)$ denotes the feature vector of the current time. $P_i(t)$ $(i=0,1,...,19)$ denotes the color feature vectors in $n\times n$ area of RGBD camera image around the skeleton joint. Position $P_i(t)$ represented by

$$P_i(t)=[R_{i1}(t),R_{i2}(t),R_{i3}(t),...,G_{i1}(t),G_{i2}(t),G_{i3}(t),...,B_{i1}(t),B_{i2}(t),B_{i3}(t),...]^T \tag{2.14}$$

where $R_{il}(t),G_{il}(t),B_{il}(t)$ denote each RGB color histogram value in $n\times n$ pixels of the current time t, respectively, l denotes the divided number in the histogram, $H(t)$ denotes the height, and $W(t)$ denotes the shoulder length.

However, the influence of height and shoulder width feature is not large enough to find out the difference for calculating the personal conformity. They are separated and these two features are included in the body size modality described in section 2.2.2. Eq. (2.12) becomes

$$Z(t)=\left[P_0(t),P_1(t),...,P_{19}(t)\right]^T \tag{2.15}$$

RGB is not a good color space. Instead, we use HSV space without V (intensity) channel. Then we can handle with large variations of illumination.

Considering the contrast of the environment light, the color space is changed to

HSV space without the V channel. Eq. (2.15) becomes

$$P_i(t) = [H_{i1}(t), H_{i2}(t), H_{i3}(t), ..., \mathrm{S}_{i1}(t), \mathrm{S}_{i2}(t), \mathrm{S}_{i3}(t), ...]^T \quad \mathbf{(2.16)}$$

where $H_{il}(t)$ denotes the H channel histogram value in HS space and $S_{il}(t)$ denotes the S channel histogram value in HS space.

The Kinect sensor is setup with a distance 2.5m from a user. The divided numbers of the color histogram are 8. The 2 divided and 4 divided are not good because they are high values after changing dress. The 16 divided and 32 divided are also not good because they are low values before changing dress. Therefore, the histogram is divided in 8, and the feature vector contains $2 \times 20 \times 8 = 320$ features. There are 8 divided histogram levels, 2 HSV colors, and 20 skeletal joints.

The personal conformity in this modality is calculated by Eq. (2.15) by using the Eq. (2.12) and Eq. (2.13). The result is shown in **Figure 2-12**.

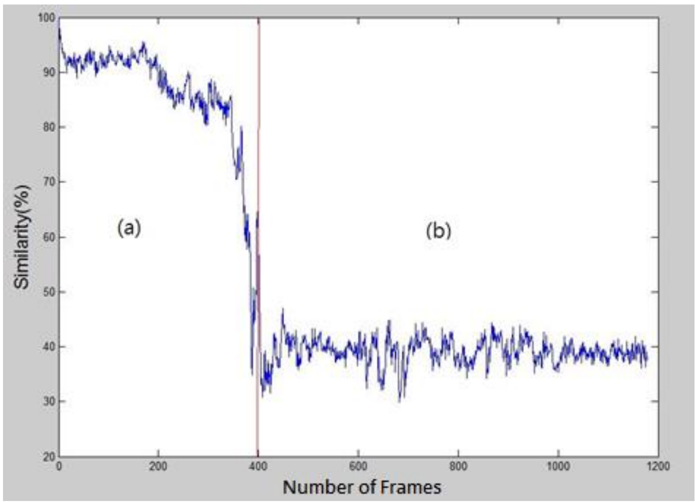


Figure 2-12 Online test of Dress color similarity diagram in time t

doi:10.6342/NTU201602427

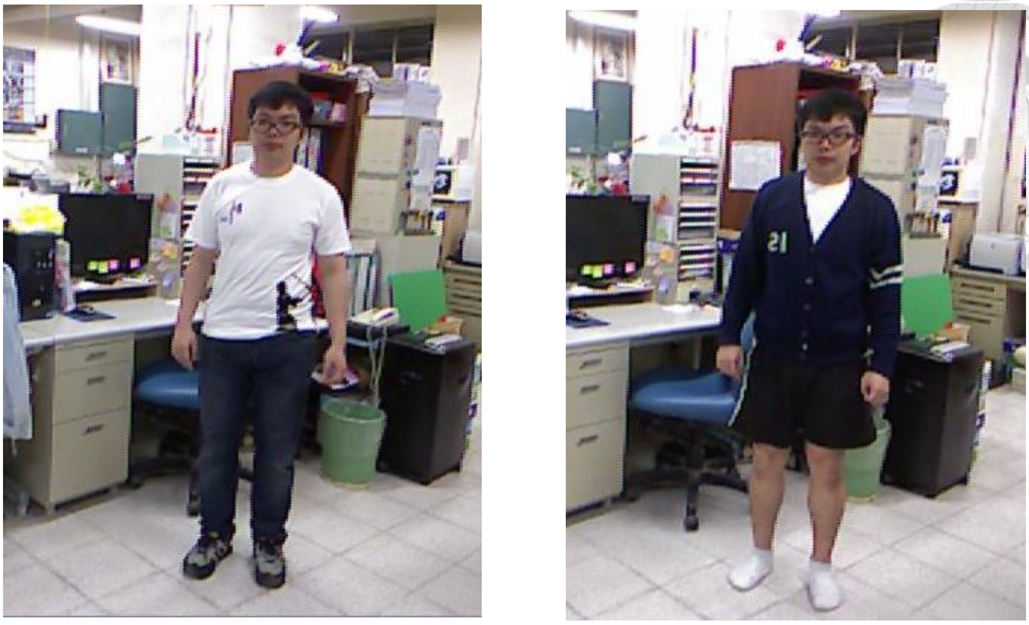


Figure 2-13 (a) Original dress (b) other dress

An online testing experiment result is shown in **Figure 2-13**. The original dress data Z(0) is shown in **Figure 2-13 (a).**

The corresponding similarity diagram is shown in region (a) of **Figure 2-12**. The similarity maintains about 90%. After changing the dressing clothes, the similarity is dropped to 40%. The corresponding similarity diagram is shown in region (b) of **Figure 2-12**.

2.2.4 User Identification System and Multimodal fusion

Jain et al. [23] introduced the biometric recognition system. It acquires biometric data from an individual, extracts a feature set from the acquired data, and compares this feature set against the template set in the database.

doi:10.6342/NTU201602427

Three levels of multimodal fusion, similar to that in [23], are employed in classic user identification system. They are feature extraction level, matching score level and decision level.

The data obtained from each biometric modality is used to compute a feature vector. It is defined as fusion at the feature extraction level. Feature reduction techniques may be employed to extract a small number of salient features from the larger set of features.

Techniques such as weighted averaging may be used for combining the matching scores in multiple matches. It is defined as Fusion at matching score.

Each biometric modality makes its own recognition decision based on its own feature vector. A majority vote scheme can be used to make the final recognition decision. It is defined as Fusion at the decision level.

There are three modalities in this chapter. A Euclidean distance is calculated in facial recognition modality. The matching score is calculated in body size recognition modality and dress color modality. Two of them are biometric features and the other one is an external feature. In Manabe et al. [33] approach, FAR (False Acceptance Rate) of decision level fusion is greater than the matching score fusion level. Considering the output of each modality, matching score level fusion is a unique solution to compose

doi:10.6342/NTU201602427

three modalities.

The fuzzy inference system can be used for fusion in the multimodal biometric system [33]. Fuzzy influence system is implemented for modality fusion in this thesis. There are three modalities in this system. The output of each modality is the confidence. However, the unit of similarity for each modality is different (Facial: Distance, Body Size and Dress: Similarity). Matching score level fusion is considered in fuzzy influence system. The block diagram of the multimodal user recognition system is shown in **Figure 2-14**.

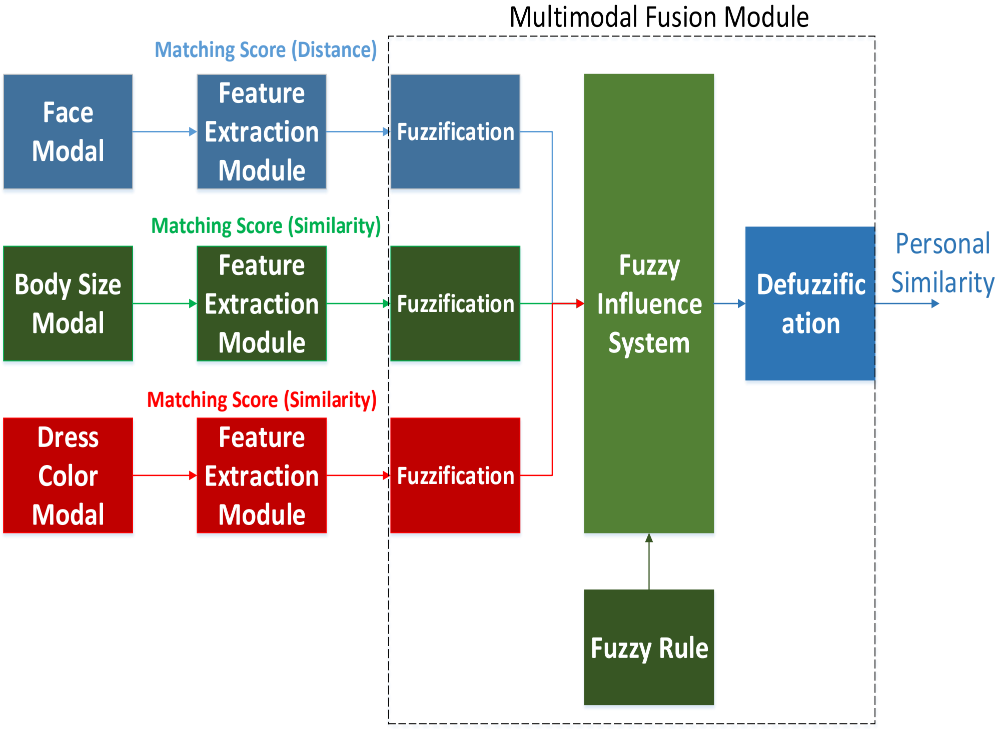


Figure 2-14 Block Diagram of Multimodal Biometric System

Figure 2-14 shows the multimodal user system implemented in this thesis. Face

doi:10.6342/NTU201602427

modality gives the matching score in distance. Body size modality and Dress color modality give the matching score in similarity. The fusion level is conducted after the fuzzification of each modality. A set of if-then rules is set and plugged in the fuzzy influence system. After defuzzification, the personal confidence is calculated.

The membership function for fuzzification in unit-modality is trapezoid membership function and the level of fusion is triangle member function. The membership functions are shown in **Figure 2-15**.

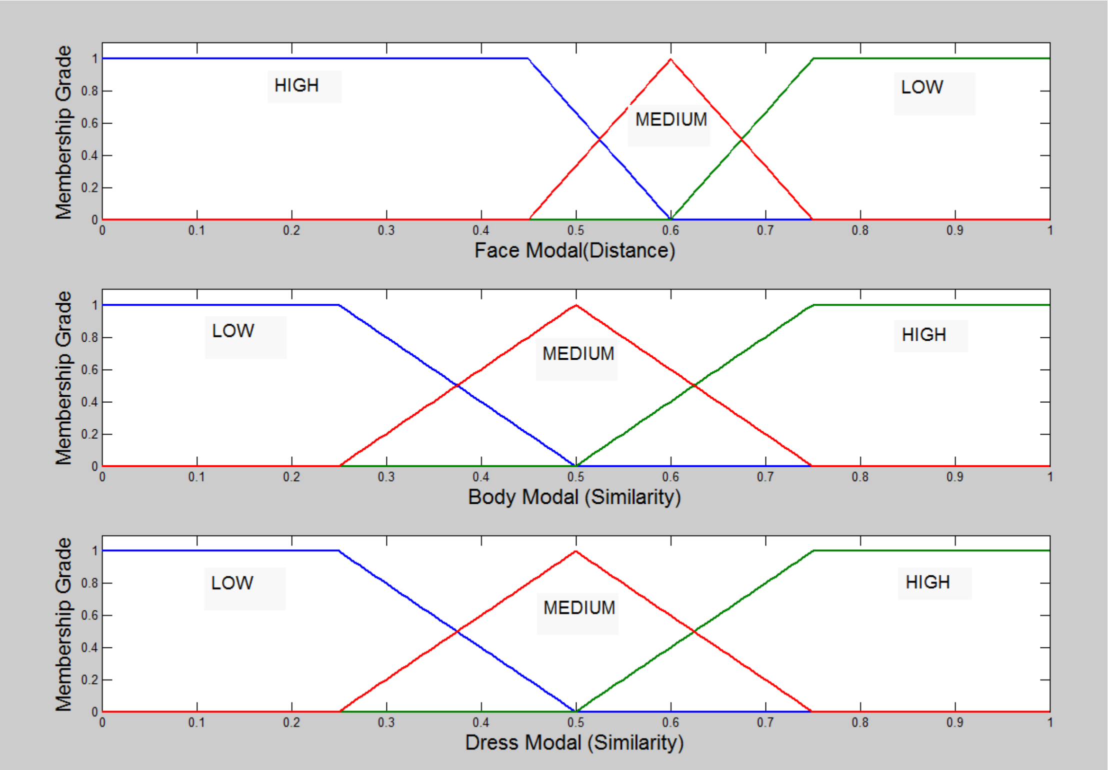


Figure 2-15 Membership function for each modality

Figure 2-15 shows the membership function in each modality. They are in trapezoid form (LOW and HIGH) and triangular form (MEDIUM). The measured

doi:10.6342/NTU201602427

distance range in the face modality is [0~1]. However, the matching distance of face modality is not stable, the output of face modality often falls into the range [0.5~0.7]. This is why the membership function is shifted to the right to ensure the accuracy. The matching score can be obtained from the other two modalities.

The membership function at fuzzy inference output level section in **Figure 2-16**.

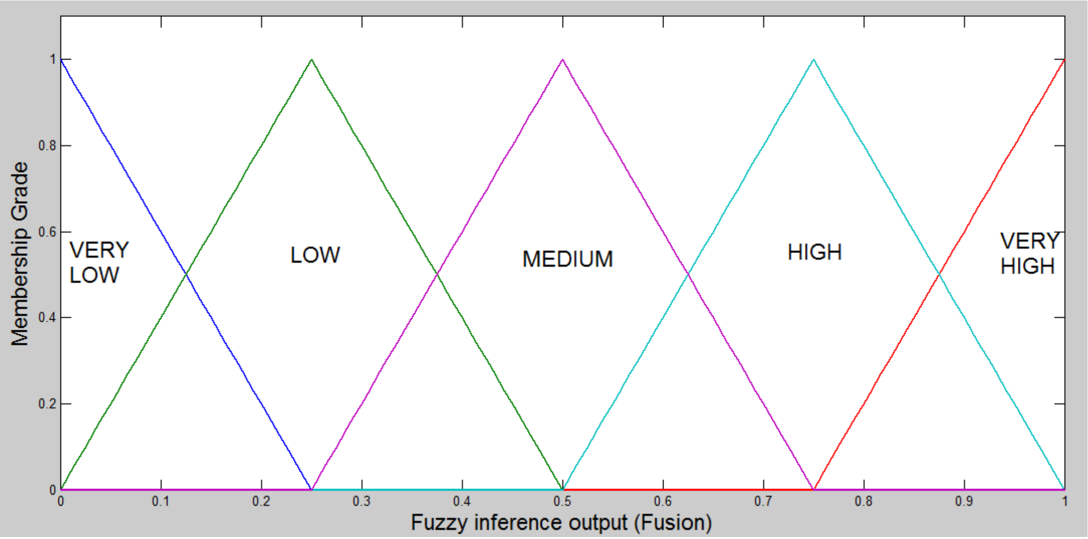


Figure 2-16 Membership function for the fusion result

Figure 2-16 shows the membership function for the fusion result. They are all in triangular form. The range of the user similarity is between 0 and 1. There are five fuzzy sets: VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH. The peak are 0, 0.25, 0.5, 0.75 and 1.

The fuzzy rule described above are implemented in this thesis. The fuzzy influence rules are shown in **Table 2-5**, **Table 2-6** and **Table 2-7**.

doi:10.6342/NTU201602427

Table 2-5 Fuzzy rule with dress modality in HIGH level

Body Face	H	M	L
H	VH	H	H
M	H	M	M
L	H	M	L

Table 2-6 Fuzzy rule with dress modality in MEDIUM level

Body Face	H	M	L
H	H	M	M
M	M	M	M
L	M	M	L

Table 2-7 Fuzzy rule with dress modality in LOW level

Body Face	H	M	L
H	H	M	L
M	M	M	L
L	L	L	VL

In the tables shown above, VH denotes VERY HIGH, H denotes HIGH, M denotes MEDIUM, L denotes LOW and VL denotes VERY LOW.

The linguistic fuzzy rules shown above address the relationship between the level of each unit-modality recognized and level of fusion. There are total 27 rules. The level of unit-modality is divided into three levels: LOW, MEDIUM, and HIGH. The level of fusion consists of five results: VERY BAD, LOW, MEDIUM, HIGH, and EXCELLENT. The level of fusion is classified as VERY HIGH and VERY LOW if all of the unit-modalities are classified as in HIGH and LOW, respectively. The level of fusion is classified as HIGH if two of the unit-modalities are classified in HIGH. The level of fusion is classified as LOW if two of the unit-modalities are classified as LOW. The level of fusion is classified as MEDIUM if two or more of the unit-modalities are classified as MEDIUM, or three modalities are classified as HIGH, MEDIUM, and LOW, respectively.

It is a Mamdani-Type (Min.-Max) Fuzzy Inference in this system. The defuzzifier of fuzzy fusion system is centroid method. This method is also called center of gravity defuzzifier or center of area defuzzifier. The defuzzifier is given as follows.

In continuous field case:

$$y^{*} = \frac{\int_{Y} \mu_{A}(y) \cdot y \, dy}{\int_{Y} \mu_{A}(y) \, dy} \tag{2.16}$$

In discrete field case:

$$y^{*} = \frac{\sum_{i=1}^{L} \mu_{\mathrm{A}}(y_i) \cdot y_i}{\sum_{i=1}^{L} \mu_{\mathrm{A}}(y_i)} \tag{2.17}$$

where A is an aggregated output fuzzy set. This is the most widely adopted defuzzification strategy, which is reminiscent of the calculation of expected values of probability distributions. y^{*} is the solution after defuzzification from the probability distributions. One of the fused results is shown in **Figure 2-17**.

doi:10.6342/NTU201602427

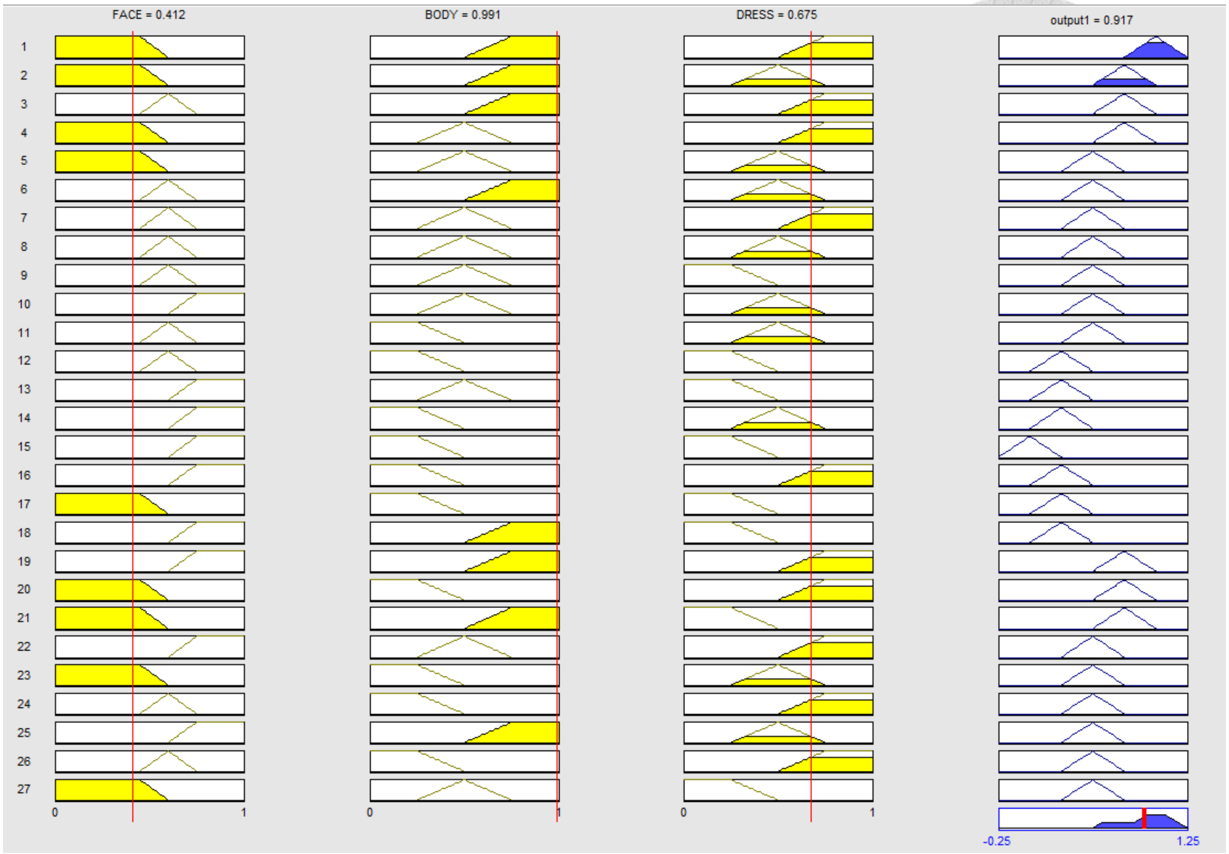


Figure 2-17 The simulation result of multimodal fusion in Fuzzy influence system

Figure 2-17 shows the result of multimodal fusion by fuzzy influence system. The simulated input values of face modality, body modality and dress color modality are 0.412, 0.991 and 0.675, respectively. After the multimodal fusion, the output after defuzzification is 0.917.

2.3 Summary

Comparing to the approach proposed by Chen [11], the skeletal joint returned by

Kinect can be used for the tracking joint [21]. As the hardware of the robot is upgraded, SLAMMOT and Kinect SDK can operate simultaneously.

We proposed the fusion of three modalities. OpenCV function is implemented in face modality. The length of limbs in skeletal data is captured by Kinect and recognized by using Support Vector Machine. There are twenty images of clothes captured by Kinect and matched with the initial dress color data by cosine similarity matching. The fuzzy influence system is implemented for fusing three modalities. It can recognize the specific user correctly. The multimodal user recognition system can still work if one of the modalities fails.

doi:10.6342/NTU201602427

Chapter 3 Following, Guiding and User Re-Tracking

This chapter is divided into two parts: following and user re-tracking. We introduce the whole human following system briefly. The robot tracks the target with combining the information of visual sensor and Laser Range Finder.

In order to improve the quality of interaction between human and the robot, a basic lost target re-tracking approach is also proposed in this section. If the target is lost in the sight of the following robot, the robot can re-tracks the target automatically. It is useful in the corner turning case and crowded case.

3.1 Human Following System

Figure 3-1 states the relationship between multimodal user recognition system (painted in blue) and function (painted in green). SLAMMOT and visual information are implemented in the system. The multimodal user recognition system is used for checking the tracked user is the target user or not. After the tracked user passes the recognition, the moving object (viz. human) tracked in SLAMMOT is assigned as the target pedestrian for the robot’s following.

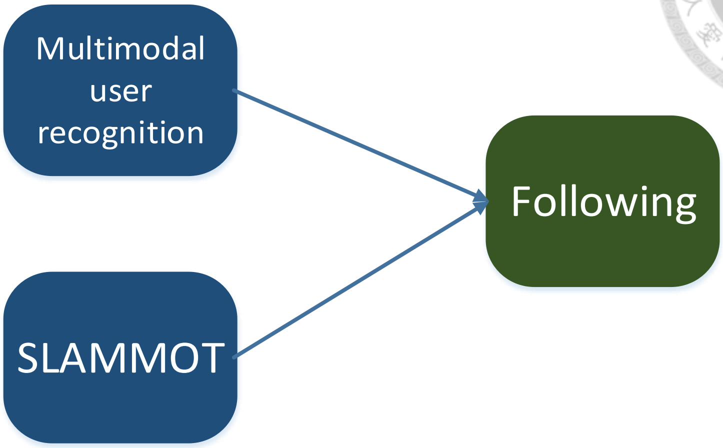


Figure 3-1 The relationship between multimodal user recognition system and function

3.2 User Re-Tracking

3.2.1 User Re-tracking mechanism

There are a small amount of research for re-tracking a lost target. Ota et al. [37] proposed an approach for re-tracking a lost target human after the human has turned the corner. The robot goes through a predicted trajectory which is modeled as a logarithmic function with 5 points. Lai [28] proposed a re-tracking scenario that the robot turns left or right according to the disappeared direction of the user. Another simple re-tracking

doi:10.6342/NTU201602427

mechanism exists in some research, such as Luo et al. [32] and Misu et al. [35] They have the similar state transition model.

Clearly, the robot has to "search" the user after the robot loses the tracking target. The state transition diagram implemented in this thesis is shown in **Figure 3-2**.

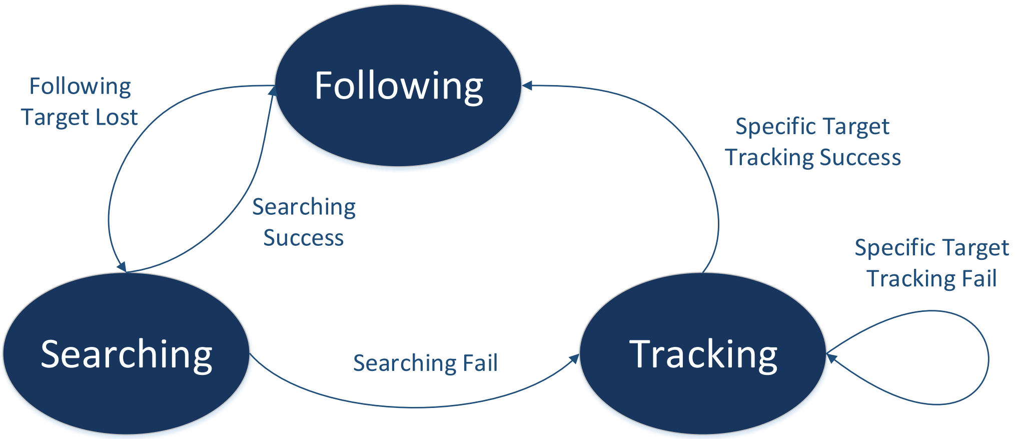


Figure 3-2 State Transition Diagram

There are three states in the specific target following system: Tracking, Following, and Searching. The "Tracking" state is activated when the system initializes. If a specific user is tracked, the state is transferred to "Following" state. Maybe the target is lost during the robot follows the specific user. "Searching" state is activated and searching for the lost target. The robot keeps following if the robot searches the lost target again. Otherwise, the robot will go back to the point where the robot loses the target. The searching behavior scenario of the robot will be introduced in the next

doi:10.6342/NTU201602427

section.

3.2.2 User Re-tracking Searching Scenario

In general, the robot searches for the lost target by turning the position of the robot to the direction where the target disappears, such as the research of Lai [28]. It is a simple approach to solve this problem. However, this approach is applicable for the open environment. The approach proposed by Ota et al. [38] is applicable when the user turns into the corner.

Two cases are considered. The first one is that the target is lost in the visual sensor. However, the laser sensor can still track the user as the sensing range of laser is wider than Kinect. The similar situation is shown in **Figure 3-3**.

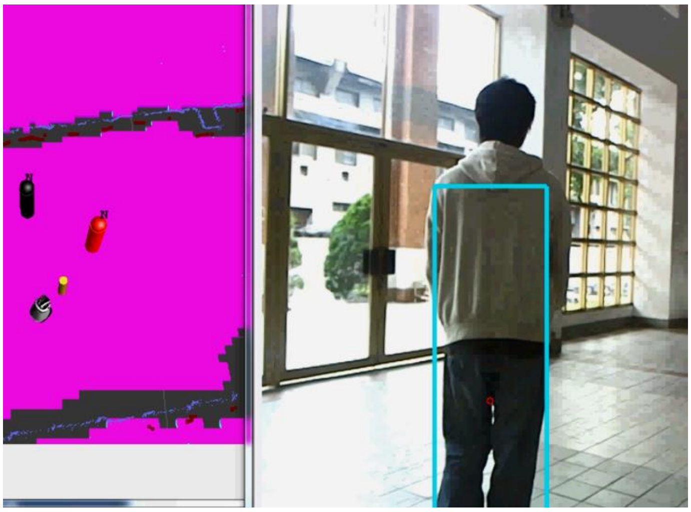


Figure 3-3 SLAMMOT Tracking and the user is out of the range of the visual sensor

Figure 3-3 shows that the laser tracks the pedestrian out of the range of the visual

doi:10.6342/NTU201602427

sensor. If one of the sensors loses the tracking target, the other sensor is activated and tracks the user. The system still works even if one of the sensors loses the following target.

The other case is that the robot is moved in a crowded environment. The robot may loses the tracking target in a crowded environment. The last position where the tracking target appears is needed. The relationships between the robot and the tracking user is shown in **Figure 3-4**.

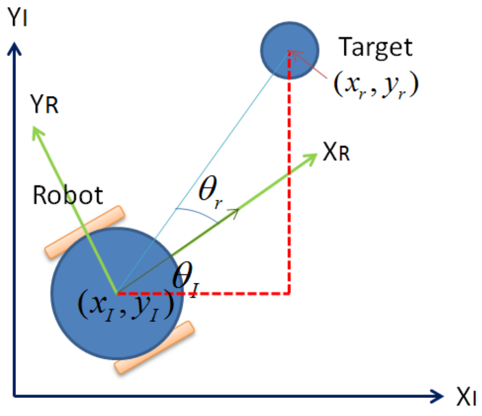


Figure 3-4 The position of the robot and target in global coordinates

Figure 3-4 shows the position of the robot and the target in global coordinates. If the target is lost at the coordinate (x_r,y_r) with orientation $\theta_I + \theta_r$ relative to the robot, SLAMMOT is implemented to locate the coordinates of the following user. In other words, if the user disappears from the sight of Kinect sensor and laser sensor of the robot simultaneously, the last position where the tracking target appears is recorded. That position is assigned as the goal that the robot searches for the lost target. After the robot has reached the goal, the robot rotated to find the user. If the user is still not found, the robot stays and wait the user to come back. It means the robot enters in the tracking transition state. The whole process is described in **Figure 3-5**.

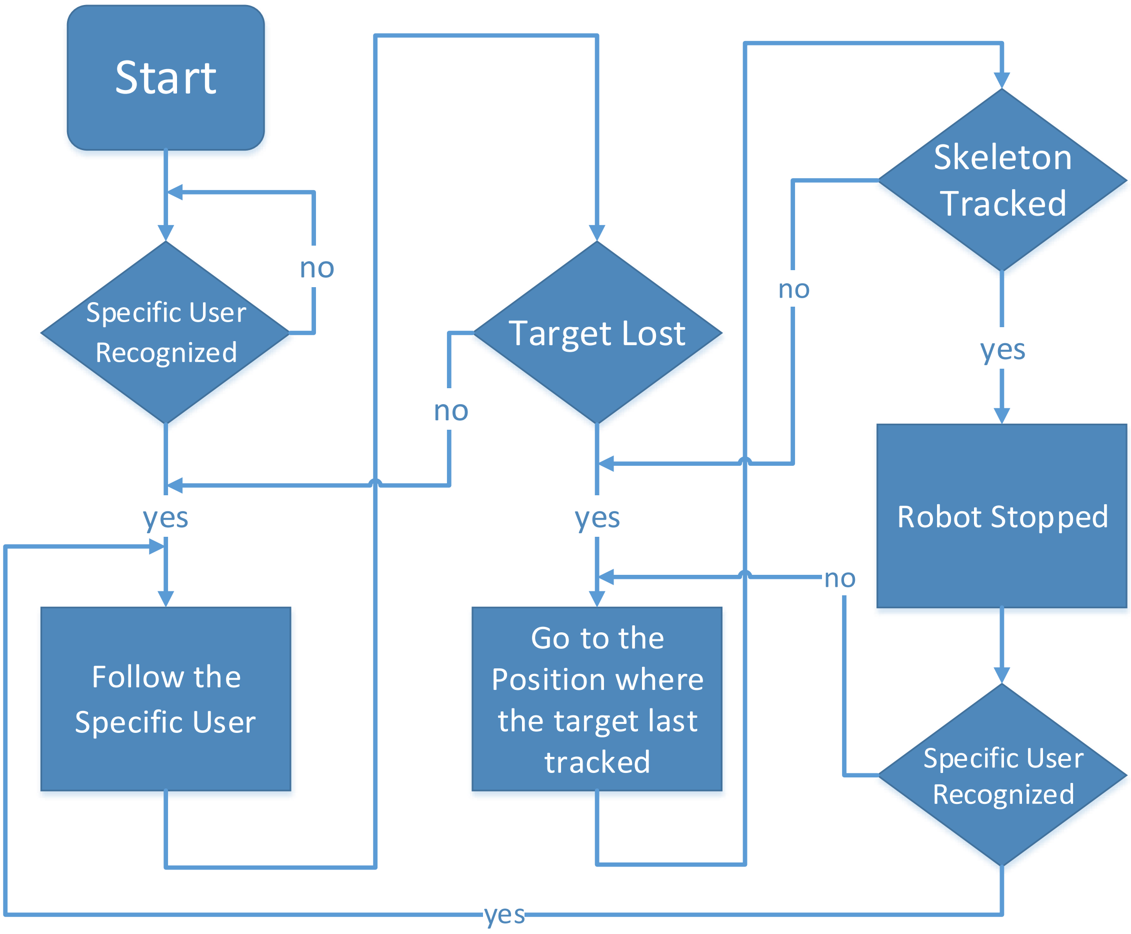


Figure 3-5 Flowchart of the processes in Specific User Re-tracking

3.4 Summary

This section presented the human following system, and the user re-tracking

doi:10.6342/NTU201602427

mechanism. The user re-tracking mechanism is developed on the basis of SLAMMOT. The final detected position of the user before disappeared becomes the goal where the robot searches for the lost target. If the robot searches for the target user successfully, the robot will follow the target again, or the robot will go back to the tracking state to wait for the target user.

doi:10.6342/NTU201602427

doi:10.6342/NTU201602427

Chapter 4 Simulations and Experiments

4.1 Software Platform

A software platform is developed for simulations and real experiments. The development and testing environment is Microsoft® Visual C++™ 2010 under Microsoft® Windows™ 7. The Open Graphics Library (OpenGL) is used for simulation display. Additional utility is provided by Open Source Computer Vision (OpenCV) with version 2.4.9. The proposed algorithms are embedded in the integrated simulation platform based on MFC (Microsoft Foundation Class).

4.2 Hardware Platform

The real experiments were conducted on the robot Bunny (see **Figure 4-1**). Bunny is a differential-drive, two-wheel mobile robot. It is equipped with a laser range finder, a head camera, motor encoders, and an RGB-depth camera. This specification of Bunny's shown in **Table 4-1**.

The laser range finder is a SICK laser LMS-291 with sensing range of 80 meters and angle resolution of 0.5 degrees. The camera is Logitech C820 web camera. It has a resolution of 640×480 pixels.

A Kinect sensor is installed in Bunny's chest as shown in **Figure 4-1** and **Figure**

doi:10.6342/NTU201602427

4-2. The Kinect has a resolution of 640×480 pixels and returns the depth data in 5 meters of the images.

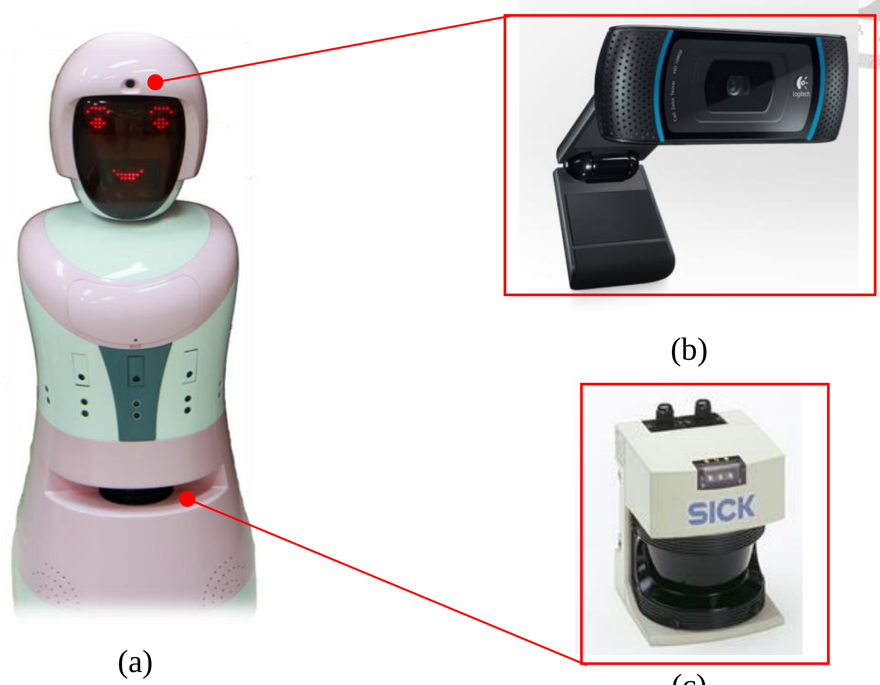


Figure 4-1 Photos of hardware (a) The robot Bunny, (b) Logitech C820 Webcam, (c) Laser Range Finder SICK LMS-291

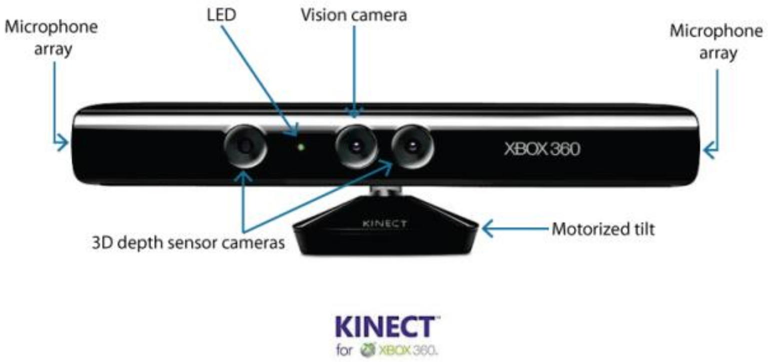
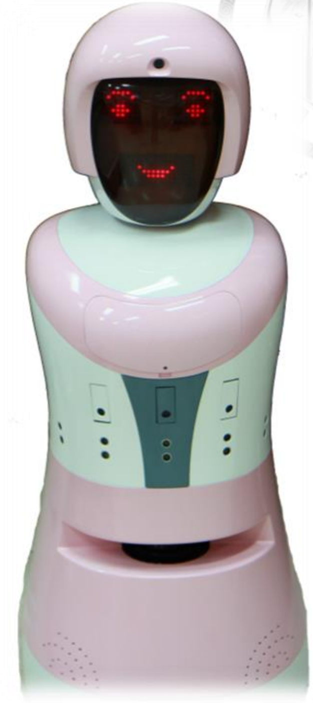


Figure 4-2 Kinect RGB-D sensor

doi:10.6342/NTU201602427

Table 4-1 The specification of the mobile robot

Item	
Name	**Bunny**
Height	**130cm**
Weight	**48.5kg**

4.3 Experimental Results

There are two cameras on Bunny, Logitech C820 webcam and Kinect. Because two cameras are implemented, the target user is asked to stand in the range of the 0.8m to 2.5m from the front of Bunny. It is the stable sensing range of the Kinect, as indicated in **Figure 4-3**.

doi:10.6342/NTU201602427

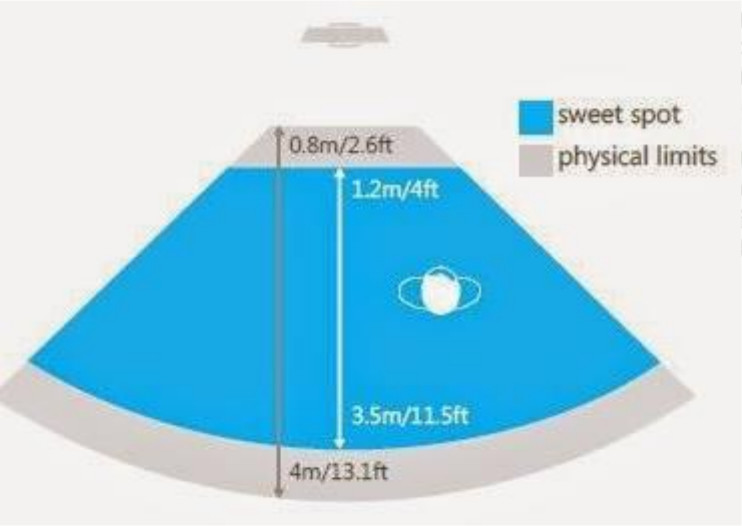


Figure 4-3 Stable sensing range of Kinect [4]

The robot follows the specific user in a distance of 1.8m. It is a good sensing distance for Kinect. The real scenario of the specific target following is shown in **Figure 4-4**.

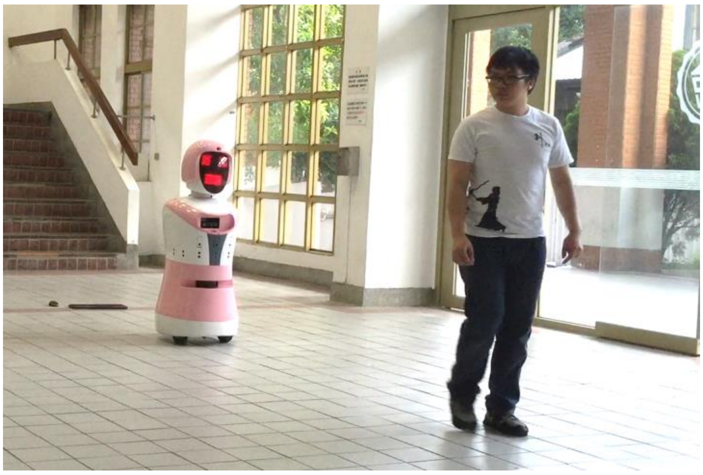


Figure 4-4 The real scenario of the specific target following

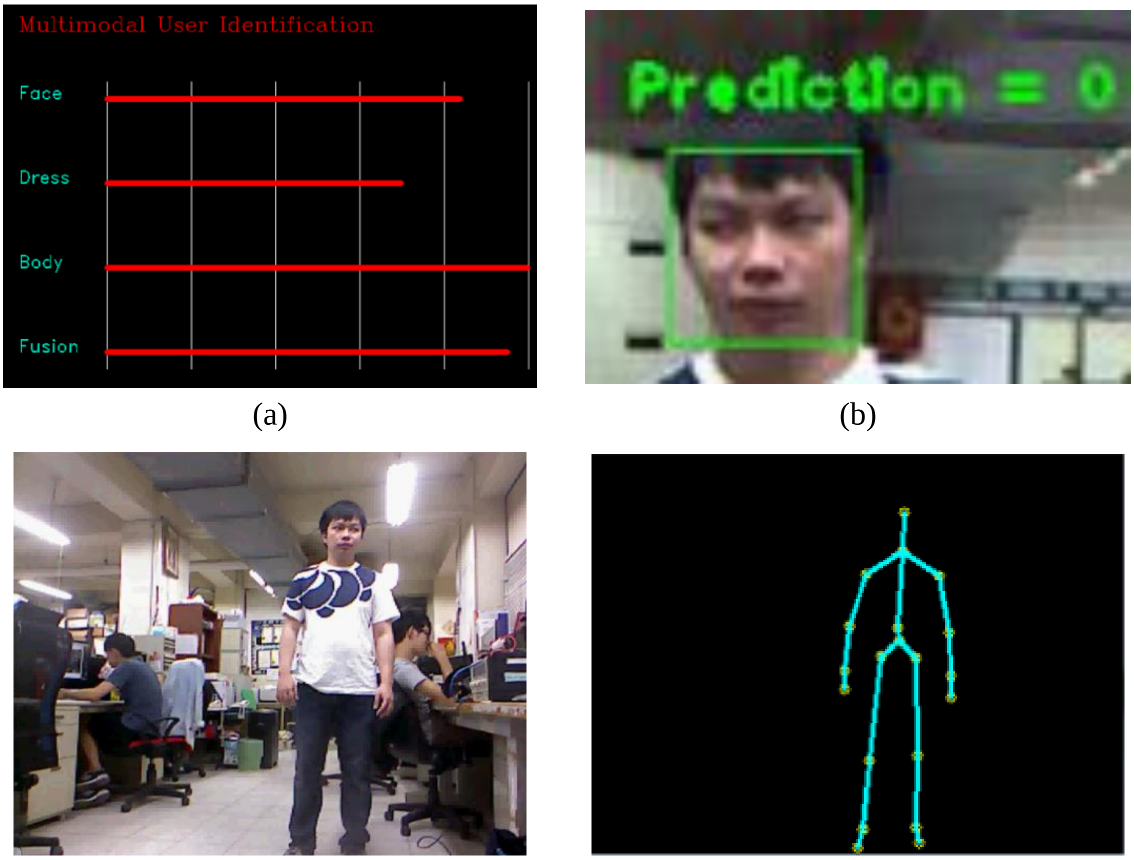
4.3.1 Case Study 1: Multimodal user recognition

The unit-modality experiment results was given explained in Chapter 2. The

doi:10.6342/NTU201602427

multimodal fusion experiments results were explained in this section.

Three modalities are implemented in this thesis. The fusion score is calculated by fuzzy influence system. In order to ensure the multimodal user recognition system still working after one of the modalities fail, the threshold is set at 0.75. It means the user is accepted if his matching score is larger than 0.75. It corresponds to HIGH level of the output of the fuzzy influence system.



(a)

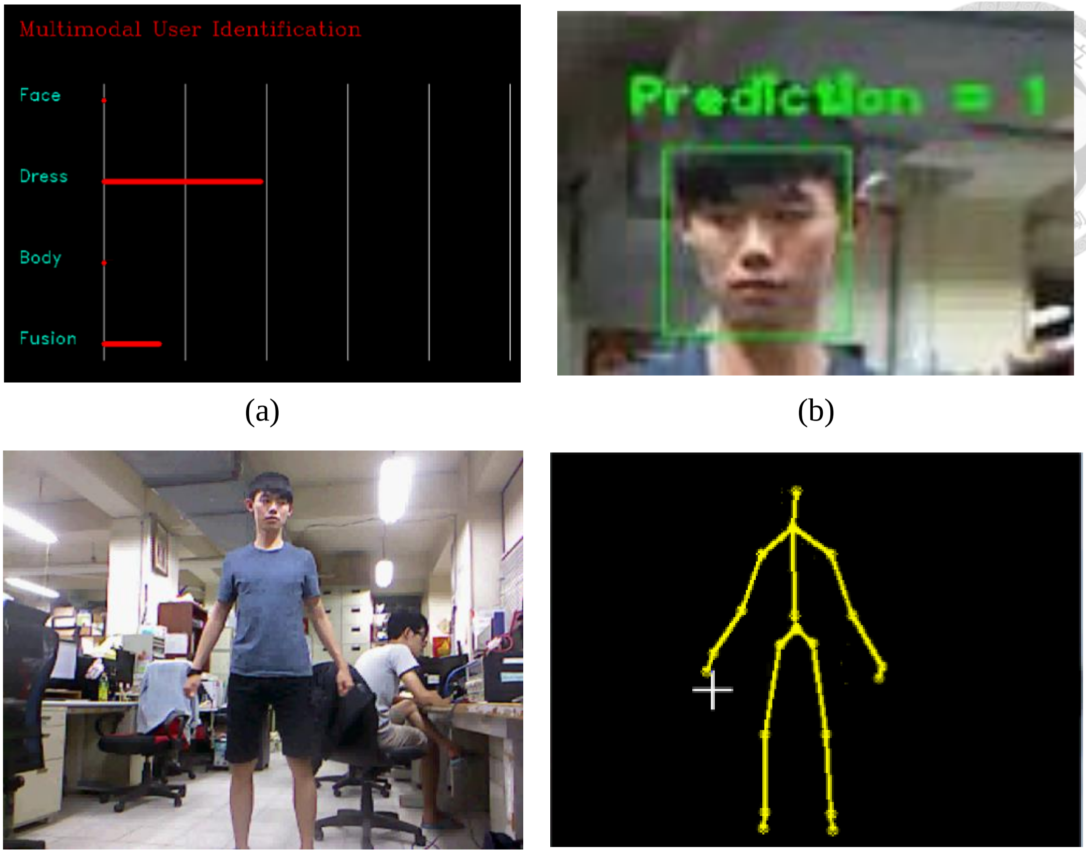
(b)

(c)

(d)

Figure 4-5 The result of correct multimodal user recognition (a) the score bar of each modality and multimodal fusion, (b) face recognition, (c) dress color recognition, (d) skeleton image

doi:10.6342/NTU201602427



(a) (b)

(c) (d)

Figure 4-6 The results of incorrect multimodal user recognition (a) the score bar of each modality and multimodal fusion, (b) face recognition, (c) dress color recognition, (d) skeleton image

Figure 4-5 and **Figure 4-6** show the results of correct and incorrect multimodal user recognition, respectively. The fusion scores are shown in of **Figure 4-5** (a) and **Figure 4-6** (a). All of the unit-modality scores and multimodal fusion scores are very high when the correct target are recognized. On the other hand, the scores are very low when the incorrect target is recognized.

After the target is recognized correctly, the ID number of skeleton is locked and

doi:10.6342/NTU201602427

the corresponding 3D coordinates (center of hip) are returned. The returned 3D position becomes the joint position of the robot.

There are 563 frames of image are captured for online testing. The results of online testing of each modalities and multimodal fusion are shown in **Table 4-2**, **Table 4-3**, **Table 4-4** and **Table 4-5**.

Table 4-2 The confusion matrix of face modality in online testing

Condition Predicted	Accept	Reject
Accept	**189**	1
Reject	133	**240**
Accuracy	**76.199%**	

Table 4-3 The confusion matrix of dress modality in online testing

Condition Predicted	Accept	Reject
Accept	**231**	0
Reject	81	**241**
Accuracy	**83.84%**	

Table 4-4 The confusion matrix of body modality in online testing

Condition Predicted	Accept	Reject
Accept	**243**	27
Reject	69	**214**
Accuracy	**81.172%**	

Table 4-5 The confusion matrix of multimodal fusion in online testing

Condition / Predicted	Accept	Reject
Accept	**232**	0
Reject	80	**241**
Accuracy	**84.0142%**	

The accurate, FAR and FRR are shown in **Table 4-6**. The mathematical functions of False Acceptance Rate and False Rejection Rate are shown in **Table 2-1**.

Table 4-6 The accurate, FAR and FRR of each modalities and multimodal

Variable / Modality	Accuracy (%)	False Acceptance Rate	False Rejection Rate
Face	76.199	0.00415	0.413
Dress	83.84	0	0.2596
Body	81.172	0.112	0.2212
Multimodal fusion	84.0142	0	0.2564

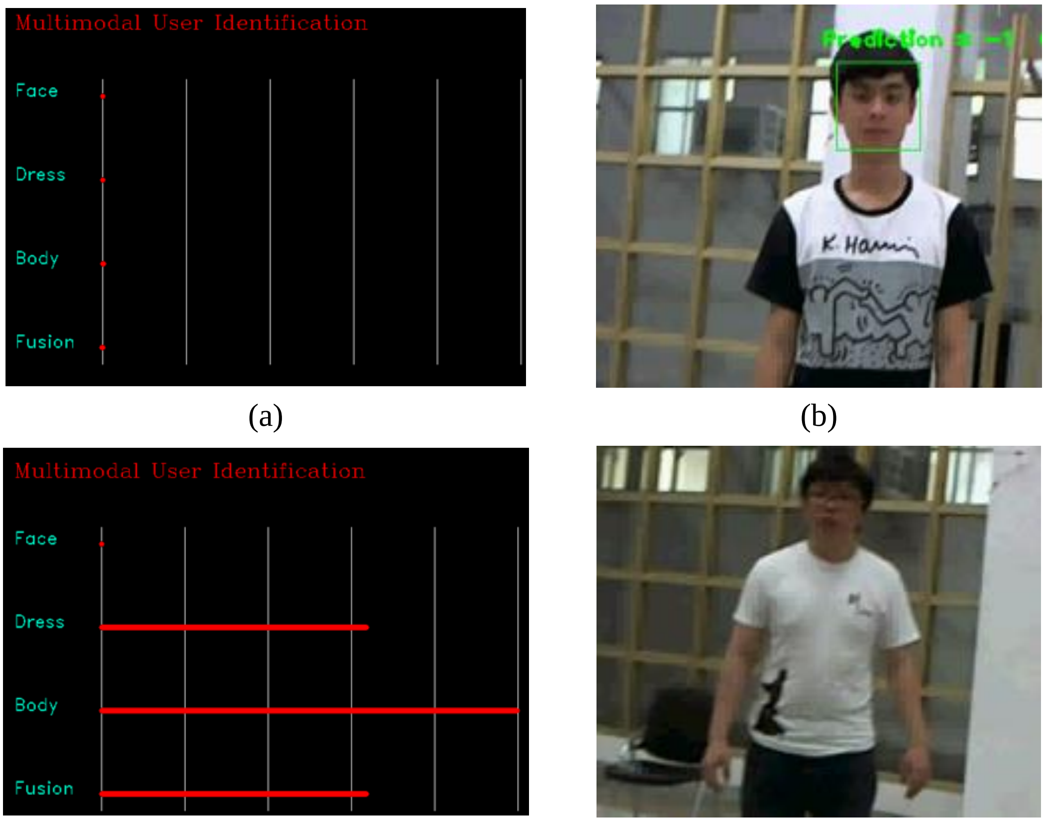
Comparing with each unit-modality and multimodal, the accurate of multimodal is the highest and FAR is lowest. Although the FRR of multimodal is higher than the body modality's, the FAR of body modality is higher than the multimodal fusion. It denotes that the multimodal fusion can improve the accuracy, FAR and FRR. The robot can recognizes a target user correctly.

4.3.2 Case Study 2: Following a Specific User

In order to make Bunny follow the real target, the multimodal user recognition module is integrated with the human following function of Bunny.

The target person is asked to stand in front of the robot and let the robot recognize the person. After successful recognition, the target person will walk around the lobby of the college of engineering building. The robot follows the user and walks around the lobby smoothly.

The following series of figures show the situation of the operating multimodal user recognition system. The robot would follow the specific user when the user accesses the recognizing of the multimodal user recognition system.


Multimodal User Identification
Face
Dress
Body
Fusion
Prediction = -1
Multimodal User Identification
Face
Dress
Body
Fusion

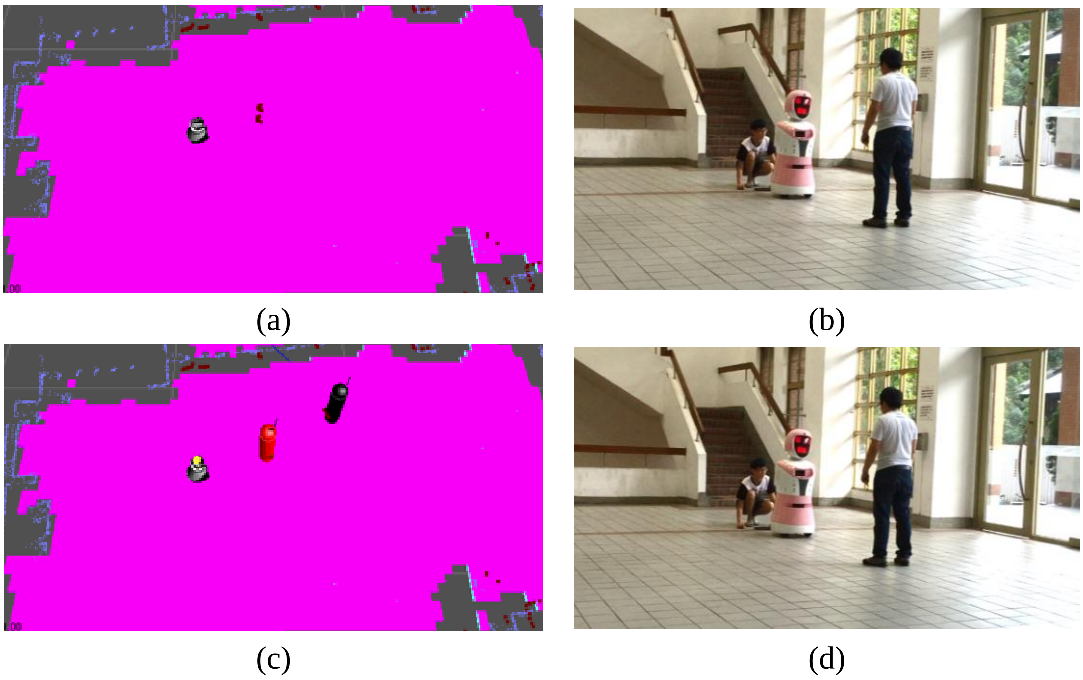
(a) (b)

(c) (d)

Figure 4-7 The multimodal user recognition system recognizes the users

The recognition result of the incorrect target is shown in **Figure 4-7** (a). The fusion score is very low. The result of the correct target is shown in **Figure 4-7** (c). The fusion score is high but there is no score in face modality. Due to the sensing range of Kinect, the user is standing 1.8m from the robot causes the images of detected faces were very blurry. Although the face modality is very unstable, the target user is still recognized successfully by the fusion of other two modalities, such as the results in **Figure 4-7** (c).

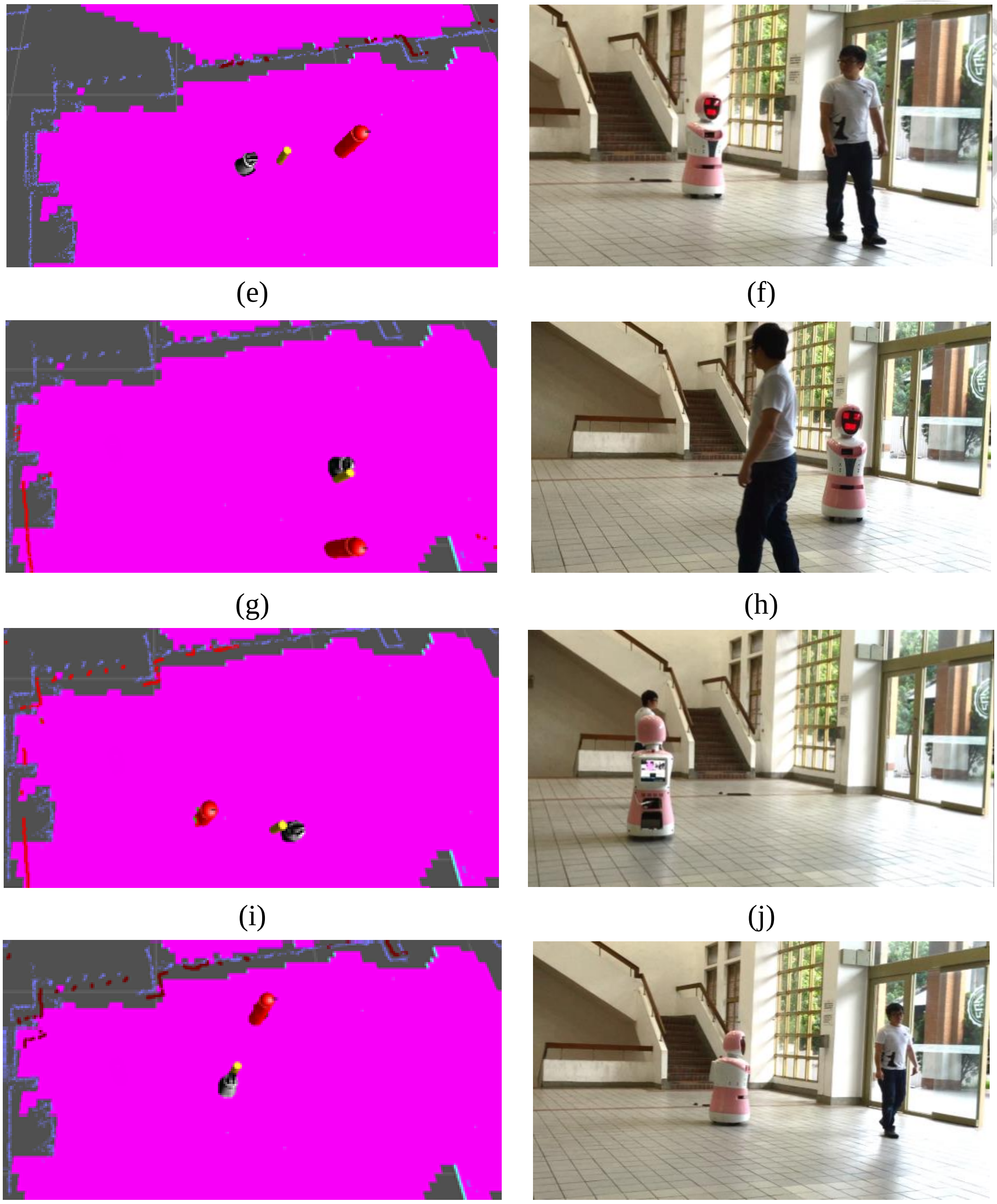
The following series of figures show that the robot follows the specific user after the user is recognized by the multimodal user recognition system. The target goal of the robot is set about 1.8m away from the user. If the distance between the robot and the user is less than 1.8m, the robot stops and waits for the moving of the user The choice of 1.8m distance is due to the physical limitation of Kinect.


(a)

(b)

(c)

(d)


(e) (f)

(g) (h)

(i) (j)

(k) (l)

Figure 4-8 The robot follows the specific user

The simulation of SLAMMOT are shown in **Figure 4-8** (a), (c), (e), (g), (i) and (k), and the experiment results are shown in **Figure 4-8** (b), (d), (f), (h), (j) and (k).

In the simulation interface of SLAMMOT, the orange and black cylinders denote

doi:10.6342/NTU201602427

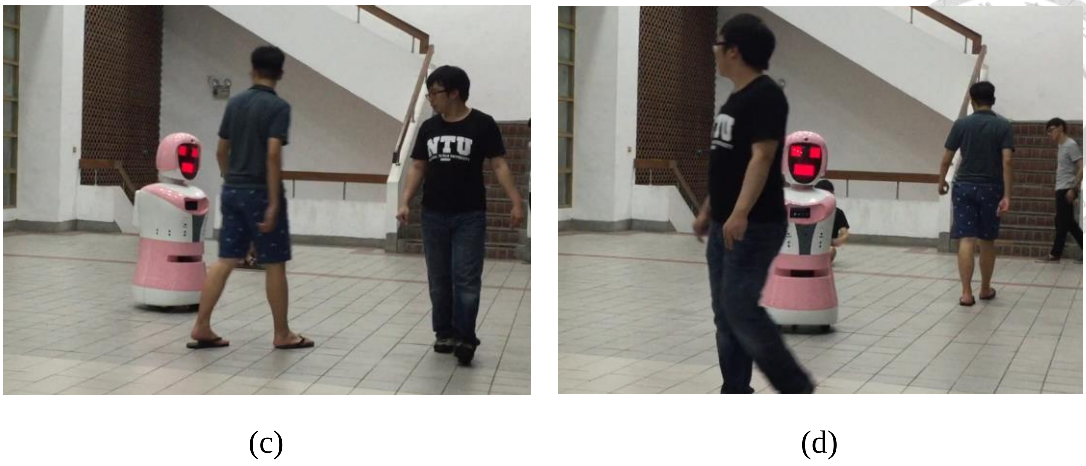
the followed person and other pedestrians, respectively. The white cylinder denotes the robot, and yellow cylinder denotes the target goal of the robot. The target goal is set at 1.8m away from the user.

When a user is recognized successfully, the robot calculates the difference between the sensor data of Kinect and Laser in a Euclidean distance. The sensor data of Kinect contains the position of tracking joints. The sensor data of laser contains the position of pedestrians relative to the robot. The pedestrian with the smallest Euclidean distance from the Kinect sensor data becomes person who will be followed (red cylinder). The other pedestrians are rejected (keep in the black cylinder). **Figure 4-8** (c) shows that the pedestrian is defined as the tracking target (shown in the red cylinder).

Due to the function of SLAMMOT, the tracked pedestrian can be labeled as target (red cylinder) and non-target (black cylinder). The robot can follows the target user in a crowded environment. The following series of figures show the situation of crowded environment.

(a)

(b)



(c) (d)

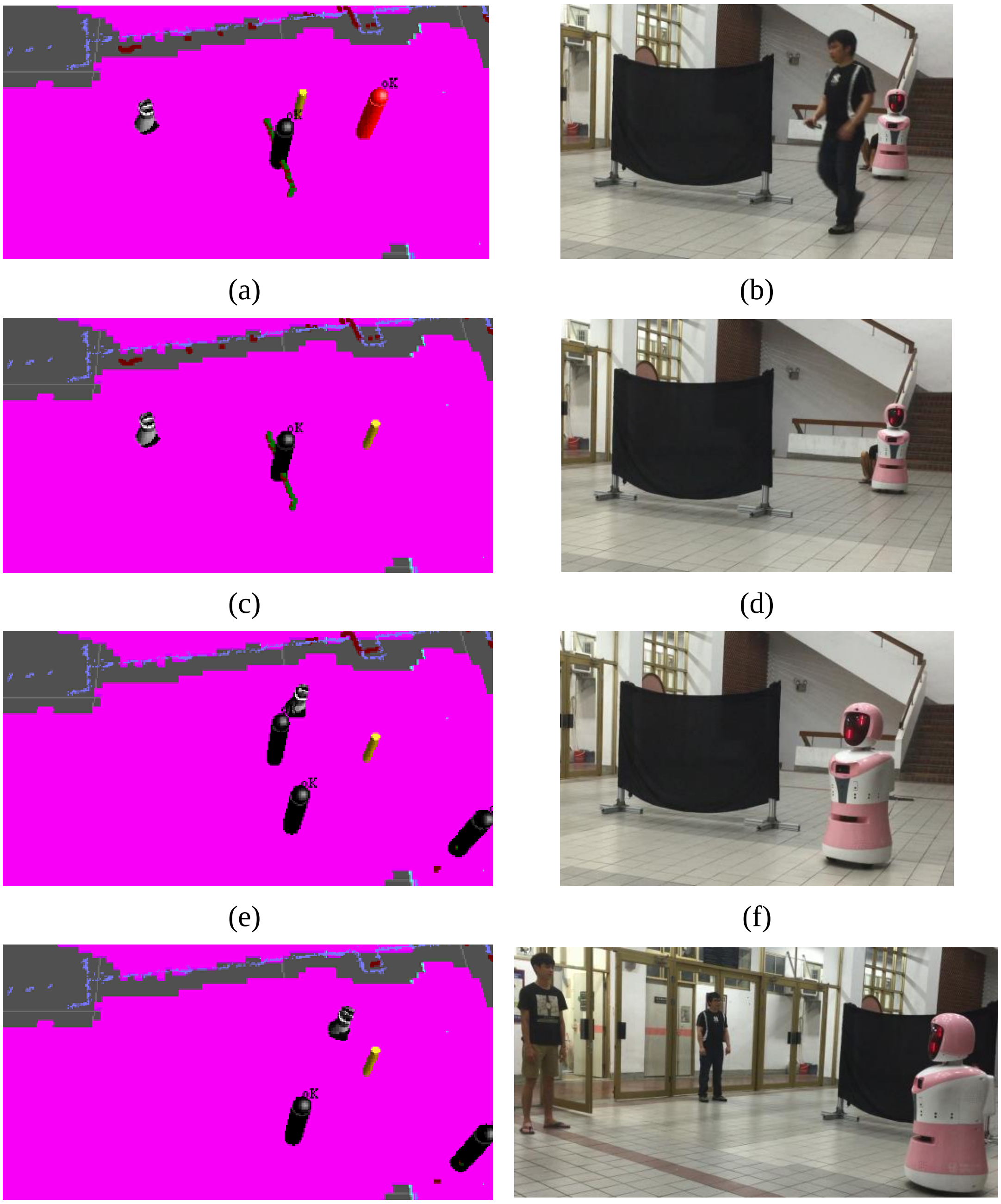
Figure 4-9 The robot follows the target user (in black dress) in crowded environment

Figure 4-9 shows the robot follows the target user in a crowded environment. The other pedestrian passes through between the robot and the target user in **Figure 4-9** (c). The robot still follows the target user in **Figure 4-9** (d).

4.3.3 Case Study 3: User Re-Tracking

In the real world, the robot may get lost due to several reasons: the target turns into the corner, in a crowded environment, blocked sight, etc. Although the robot has an ability to re-track the user, the robot might follow another pedestrian.

The following series of figures show the real situation when the robot loses the tracking target with the situation of the target user turning into the corner. The lost position of the target user becomes the target goal of the robot. The robot moves to the target goal and stops when the Kinect detects the pedestrian skeleton. The multimodal user recognition system starts to recognize the presently detected pedestrians. The robot will follow the user again when the pedestrian passes the verification by the multimodal user recognition system, or the robot stops and waits for the correct user.



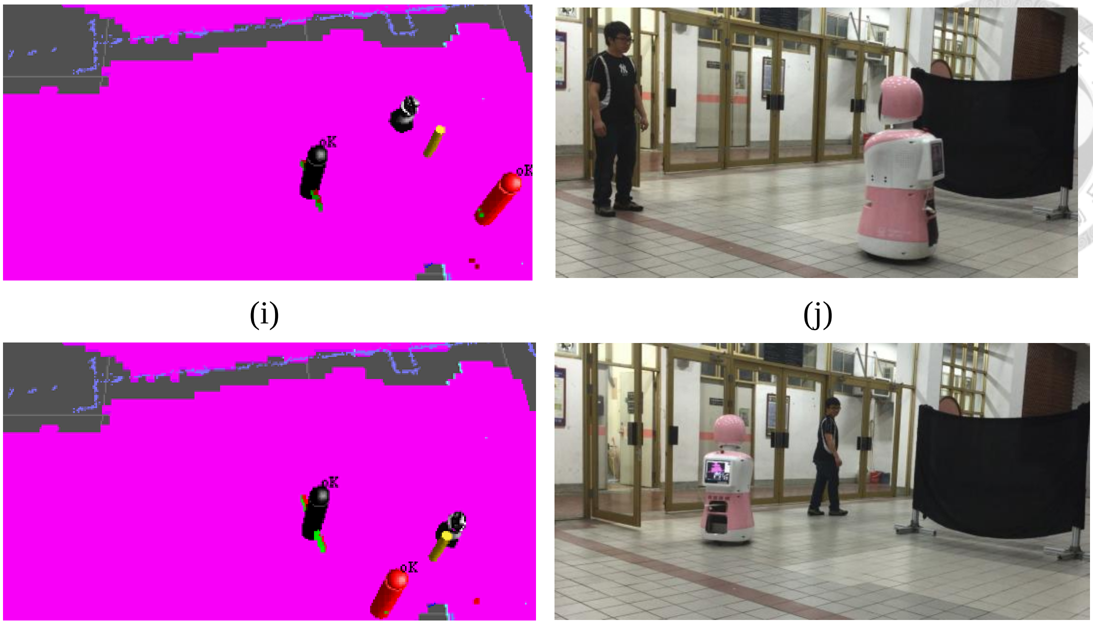
(a) (b)

(c) (d)

(e) (f)

(g) (h)

doi:10.6342/NTU201602427



(i) (j)

(k) (l)

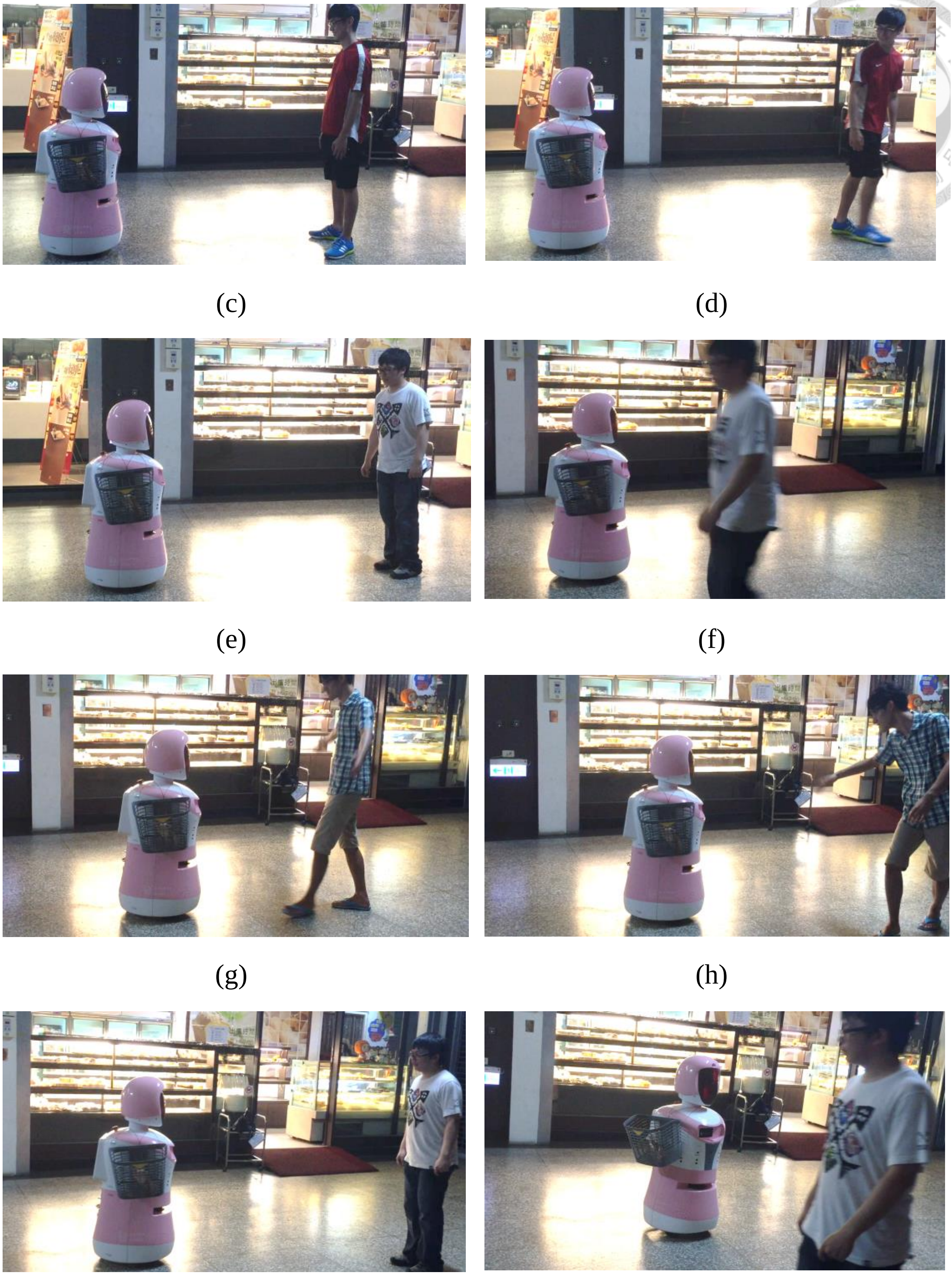
Figure 4-10 The robot re-tracks the target user

The simulation results of SLAMMOT are shown in **Figure 4-10** (a), (c), (e), (g), (i) and (k), and the experiment results are shown in **Figure 4-10** (b), (d), (f), (h), (j) and (k). The user re-tracking scenario is constructed as follows. A large black cloth is erected to simulate a corner. A user is asked to walk and turn into a corner. The laser range finder and Kinect cannot track the desired target because the black cloth blocks the sight, i.e. the street corner. It is shown in **Figure 4-10** (b). In **Figure 4-10** (c), the last position of the target is set as a target goal (yellow cylinder). The robot moves to the goal, which is shown in **Figure 4-10** (e) and **Figure 4-10** (f). The robot finds the other pedestrian in **Figure 4-10** (g) and the pedestrian does not passes the verification of multimodal user recognition system. Then, the pedestrian is defined as a black cylinder in the simulation interface of SLAMMOT. The correct target is back in the sight of the robot and passes the verification of multimodal user recognition system in **Figure 4-10** (i). Then, the correct target is defined as a red cylinder in the simulation interface of SLAMMOT. After the successful recognition, the robot follows the user again.

4.4 Summary

This thesis is focused on presenting a specific user recognition system for a human following robot. Hence, all experiments are used to simulate some situations the robot may encounter when it loses the master. To explain the relationship between the experiments and practical usages, we equipped a basket on the robot and simulated practical shopping scenarios. The following series of figures show the real situation that the robot follows the master only and how the robot reacts when another pedestrian is in the sight of the robot.

(a) (b)



(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure 4-11 The real situation in the First Student Activity Center

Figure 4-11 shows the series of the real situations in the First Student Activity

Center. The master buys goods from a shop and puts the goods into the basket, as shown in **Figure 4-11** (a) and (b). If someone else appears between the robot and the master, the robot will stop, as shown in **Figure 4-11** (c) and (d). Upon the master coming back, the robot will follow the master again. If the master makes a turn suddenly, the robot is not fast enough to track the master, as shown in **Figure 4-11** (f). In **Figure 4-11** (g) and (h), the other pedestrian comes into the sight of the robot. The robot will not move since the pedestrian does not pass the recognition process. Once the master is back in the sight of the robot, the robot will follow the master again, as shown in **Figure 4-11** (i) and (j).

doi:10.6342/NTU201602427

Chapter 5 Conclusions and Future Works

5.1 Conclusions

Servant robots should have the ability to follow its master and provide services whenever its master asks. Our purpose is to build a practical assistant robot that can be adopted in malls, department stores or shopping centers. Therefore, this thesis proposed an application of a shopping assistant robot based on the thesis of Chen. We targeted on an accurate specific user recognition and user re-tracking mechanism after the desired target is lost in the environment.

We implemented three modalities, face, body size, and dress color matching, in the specific user recognition system. We combined these three modalities by fuzzy influence system. The combined system can still work even if one of the modalities fail. The proposed multimodal user recognition system can recognize the specific user.

We also proposed a simple mechanism for the lost target re-tracking. The last detected position of the user becomes the goal where the robot searches for the user. By combining SLAMMOT with multimodal user recognition, the robot can always localize itself in the map and re-track the lost target by referring the information provided by SLAMMOT and multimodal user recognition system.

doi:10.6342/NTU201602427

5.2 Future Works

Improving the user recognition process

The multimodal user recognition system can recognize the user accurately in a static condition. But, it may not be accurate in a dynamic condition. It is easy to recognize a false result when the robot is moving, especially in body modality. Better features can be implemented for the body modality.

Gesture Control

An approach of specific user following is provided in this thesis, but the interactivity of the robot can be improved. Some situations should be considered in the shopping mall, such as some assigned areas in a shopping mall that the robot cannot pass through. The robot should stand by. We should “tell” the robot that the robot should stop. Gesture control is one of the interaction approaches.

doi:10.6342/NTU201602427

References

[1] Face Recognition with OpenCV. Retrieved Jun., 2016, from http://docs.opencv.org/2.4/modules/contrib/doc/facerec/facerec_tutorial.html

[2] LibSVM. Retrieved Dec., 2015, from https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3] Self-propelled luggage Retrieved Apr., 2014, from http://www.google.tl/patents/US20140107868

[4] Skeletal Tracking. Retrieved Jan., 2016, from https://msdn.microsoft.com/en-us/library/hh973074.aspx

[5] Tracking Users with Kinect Skeletal Tracking. Retrieved Jun., 2012, from https://msdn.microsoft.com/en-us/library/jj131025.aspx

[6] Would you trust a robot to take care of your grandpa? Retrieved Jun., 2014, from http://robohub.org/would-you-trust-a-robot-to-take-care-of-your-grandma/

[7] S. An, X. Ma, R. Song, and Y. Li, "Face Detection and Recognition with SURF for Human-Robot Interaction," *Proc. of IEEE International Conference on Automation and Logistics*, Shenyang, China, pp. 1946-1951, 2009.

[8] R. M. Araujo, G. Gra, and V. Andersson, "Towards Skeleton Biometric Identification using the Microsoft Kinect Sensor," *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, Coimbra, Portugal, pp. 21-26, 2013.

[9] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D Sensors," in *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Berlin, Germany: Springer, 2012 pp. 433-442.

[10] F. Bohnert, I. Zukerman, S. Berkovsky, T. Baldwin, and L. Sonenberg, "Using interest and transition models to predict visitor locations in museums," *AI Commun.*, Vol. 21, No. 2-3, pp. 195-202, 2008.

[11] M. N. Chen, *Development of Shopping Assistant Robots*, *Master Thesis*, Department of Mechanical Engineering, National Taiwan University, 2013.

[12] S. Y. Chung, *Spatial Understanding and Motion Planning for a Mobile Robot*, *Doctoral Dissertation*, Department of Mechanical Engineering, National Taiwan University, pp. 59-84, 2010.

[13] S. Y. Chung and H. P. Huang, "SLAMMOT-SP: Simultaneous SLAMMOT and Scene Prediction," *Advanced Robotics*, Vol. 24, No. 7, pp. 979-1002, 2010.

[14] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.

[15] A. H. Dallal, A. S. Derbala, and M. F. Taher, "A mobile robot for hospitals controlled using WLAN," *Proc. of Biomedical Engineering Conference*, Giza,

doi:10.6342/NTU201602427

Egypt, pp. 100-103, 2012.

[16] N. Deng, Y. Tian, and C. Zhang, *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions,* British, U.K.: Chapman and Hall/CRC 2012.

[17] P. Fan, A. Men, M. Chen, and B. Yang, "Color-SURF: A SURF descriptor with local kernel color histograms," *Proc. of IEEE International Conference on Network Infrastructure and Digital Content,* Beijing, China, pp. 726-730, 2009.

[18] S. Fazli, H. M. Pour, and H. Bouzari, "Particle Filter Based Object Tracking with Sift and Color Feature," *Proc. of Second International Conference on Machine Vision*, Dubai, UAE, pp. 89-93, 2009.

[19] W. K. Fung, Y. Y. Leung, M. K. Chow, Y. H. Liu, Y. Xu, W. Chan, T. W. Law, S. K. Tso, and C. Y. Wang, "Development of a hospital service robot for transporting task," *Proc. of IEEE International Conference on Robotics and Intelligent Systems and Signal Processing,* Changsha, Hunan, China, Vol. 1, pp. 628-633, 2003.

[20] J. H. Guo, K. L. Su, C. C. Wang, and C. J. Wu, "Laser Range Finder Applying in Motion Control System of Mobile Robots," *Proc. of Fourth International Conference on Innovative Computing, Information and Control,* Kaohsiung, Taiwan, pp. 536-539, 2009.

[21] B. Ilias, S. A. Shukor, S. Yaacob, A. Adom, and M. M. Razali, "A Nurse Following Robot with High Speed Kinect Sensor," *Journal of Engineering and Applied Sciences,* Vol. 9, No. 12, pp. 2454-2459, 2014.

[22] Y. Iwamura, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets?," *Proc. of 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* Lausanne, Switzerland, pp. 449-457, 2011.

[23] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 14, No. 1, pp. 4-20, 2004.

[24] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "An affective guide robot in a shopping mall," *Proc. of 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* La Jolla, California, U.S.A., pp. 173-180, 2009.

[25] H. J. Kim and J. Lee, "Designing diving beetle inspired underwater robot(D.BeeBot)," *Proc. of 13th International Conference on Control Automation Robotics & Vision (ICARCV),* Singapore, pp. 1746-1751, 2014.

[26] M. G. Kim, H. Lee, J. Lee, S. S. Kwak, and Y. Joo, "Effectiveness and service quality of robot museum through visitors experience: A case study of RoboLife

doi:10.6342/NTU201602427

Museum in South Korea," *Proc. of International Symposium on Micro-NanoMechatronics and Human Science (MHS)*, Nagoya, Japan, pp. 1-5, 2015.

[27] R. Laganière, *OpenCV 2 Computer Vision Application Programming Cookbook*, Packt Publishing, U.K., 2011.

[28] I. H. Lai, *The following robot with searching and obstacle-avoiding, Master Thesis*, Department of Electrical Engineering, National Central University, 2009.

[29] Y. Lee, K. Lee, H. Jee, Y. Gil, W. Choi, D. Ahn, and S. Pan, "Fusion for Multimodal Biometric Identification," in *5th International Conference of Audio- and Video-Based Biometric Person Authentication*, Berlin, Germany: Springer, 2005 pp. 1071-1079.

[30] M. Liem, A. Visser, and F. Groen, "A hybrid algorithm for tracking and following people using a robotic dog," *3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, Netherlands, pp. 185-192, 2008.

[31] R. C. Luo, N. W. Chang, S. C. Lin, and S. C. Wu, "Human tracking and following using sensor fusion approach for mobile assistive companion robot," *Proc. of 35th Annual Conference of IEEE on Industrial Electronics*, Porto, Portugal, pp. 2235-2240, 2009.

[32] R. C. Luo, C. H. Huang, and T. T. Lin, "Human tracking and following using sound source localization for multisensor based mobile assistive companion robot," *Proc. of 36th Annual Conference on IEEE Industrial Electronics Society*, Glendale, Arizona, U.S.A., pp. 1552-1557, 2010.

[33] Y. Manabe, R. Saito, Y. Shimada, and K. Sugawara, "Frontal View Person Verification by Soft Biometric Features of Gait, Face and Body," *J. SOFT*, Vol. 24, No. 5, pp. 988-1001, 2012.

[34] C. Martin, E. Schaffernicht, A. Scheidig, and H. M. Gross, "Multi-modal sensor fusion using a probabilistic aggregation scheme for people detection and tracking," *Robotics and Autonomous Systems*, Vol. 54, No. 9, pp. 721-728, Sept. 2006.

[35] K. Misu and J. Miura, "Specific Person Detection and Tracking by a Mobile Robot Using 3D LIDAR and ESPAR Antenna," *Proc. of the 13th International Conference of Intelligent Autonomous Systems,*, Padua, Italy, pp. 705-719, 2016.

[36] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An Adaptive Color-Based Particle Filter," *Image and Vision Computing*, Vol. 21, No. 1, pp. 99-110, 2003.

[37] M. Ota, H. Hisahara, H. Takemura, and H. Mizoguchi, "Recovery function of target disappearance for human following robot," *Proc. of First International Conference on Innovative Engineering Systems*, Alexandria, Egypt, pp. 125-128, 2012.

doi:10.6342/NTU201602427

[38] M. Ota, T. Ogitsu, H. Hisahara, H. Takemura, Y. Ishii, and H. Mizoguchi, "Recovery function for human following robot losing target," *Proc. of 39th Annual Conference of the IEEE of Industrial Electronics Society*, Vienna, Austria, pp. 4253-4257, 2013.

[39] S. Ouellet, F. Grondin, F. Leconte, and F. Michaud, "Multimodal biometric identification system for mobile robots combining human metrology to face recognition and speaker identification," *Proc. of The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Edinburgh, Scotland, U. K. , pp. 323-328, 2014.

[40] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multimodal Person Reidentification Using RGB-D Cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 26, No. 4, pp. 788-799, 2016.

[41] M. P. Paulraj, R. B. Ahmad, C. R. Hema, F. Hashim, and S. Yusoff, "Active stereo vision based system for estimation of mobile robot orientation using affine moment invariants," *Proc. of International Conference on Electronic Design*, Penang, Malaysia, pp. 1-7, 2008.

[42] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, Vol. 2, No. 11, pp. 559-572, 1st Nov. 1901.

[43] M. V. Rajasekhar and A. K. Jaswal, "Autonomous vehicles: The future of automobiles," *Proc. of IEEE International Transportation Electrification Conference*, Chennai, Tamil Nadu, India, pp. 1-6, 2015.

[44] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, Vol. 323, No. 6088, pp. 533-536, Sept., 1986.

[45] J. Satake, M. Chiba, and J. Miura, "Visual Person Identification Using a Distance-dependent Appearance Model for a Person Following Robot," *International Journal of Automation and Computing*, Vol. 10, No. 5, pp. 438-446, 2013.

[46] M. Shiomi, T. Kanda, D. F. Glas, S. Satake, H. Ishiguro, and N. Hagita, "Field trial of networked social robots in a shopping mall," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, Missouri, U.S.A., pp. 2846-2853, 2009.

[47] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Interactive Humanoid Robots for a Science Museum," *IEEE Intelligent Systems*, Vol. 22, No. 2, pp. 25-32, 2007.

[48] A. Sinha, K. Chakravarty, and B. Bhowmick, "Person identification using skeleton information from kinect," *Proc. of Intl. Conf. on Advances in Computer-Human Interactions*, Nice, France, pp. 101-108, 2013.

[49] S. Sugiyama, K. Baba, and T. Yoshikawa, "Guide robot with personal

doi:10.6342/NTU201602427

identification method using dress color information via KINECT," *Proc. of IEEE International Conference on Robotics and Biomimetics*, Guangzhou, China, pp. 2195-2200, 2012.

[50] S. Sugiyama and T. Wada, "Improved personal identification method for guide robots using dress color information via KINECT," *Proc. of International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Hanoi, Vietnam, pp. 111-116, 2013.

[51] R. Vijay, A. Kapuria, C. Datta, G. Dubey, C. Sharma, and G. Taank, "Ecosystem of a shopping mall robot - Neel," *Proc. of 9th International Conference on Ubiquitous Robots and Ambient Intelligence*, Daejeon, South Korea, pp. 120-125, 2012.

[52] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 511-518, 2001.

[53] C. C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous Localization, Mapping and Moving Object Tracking," *Int. J. Robotics Research*, Vol. 26, No. 9, pp. 889-916, 2007.

[54] X. Wu, H. Gong, P. Chen, Z. Zhi, and Y. Xu, "Intelligent household surveillance robot," *Proc. of IEEE International Conference on Robotics and Biomimetics*, Bangkok, Thailand, pp. 1734-1739, 2009.

[55] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada, K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba, "Home-Assistant Robot for an Aging Society," *Proceedings of the IEEE*, Vol. 100, No. 8, pp. 2429-2441, 2012.

[56] L. Yang, X. Song, Y. Li, H. Shan, and J. Guo, "Design and Experimental Research on Intelligent Household Assistive Robot for the Elderly," *Proc. of Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, Qinhuangdao, China, pp. 1316-1319, 2015.

[57] Y. Yoo and C. Woojin, "Detection and following of human legs using the SVDD (Support Vector Data Description) scheme for a mobile robot with a single Laser Range Finder," *Proc. of International Conference on Electrical, Control and Computer Engineering*, Pahang, Malaysia, pp. 97-102, 2011.

[58] Y. Zheng, "Using Score Fusion for Improving the Performance of Multispectral Face Recognition," in *Signal and Image Processing for Biometrics* Berlin, Germany: Springer, 2014 pp. 107-128.