

國立交通大學

電機與控制工程學系

碩士論文

利用動態攝影機實現人形追蹤與人臉偵測技術

A Study of Human Tracking and Face Detection
on A Pan-Tilt-Zoom Camera



研究生：朱琳達

指導教授：林進燈 教授

中華民國九十四年七月

利用動態攝影機實現人形追蹤與人臉偵測技術

A Study of Human Tracking and Face Detection
on A Pan-Tilt-Zoom Camera

研究生：朱琳達

Student：Linda Siana

指導教授：林進燈 教授

Advisor：Prof. Chin-Teng Lin

國立交通大學
電機與控制工程學系
碩士論文



A Thesis

Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering and Computer Science

National Chiao-Tung University

In Partial Fulfillment of the Requirements

For the Degree of Master In

Electrical and Control engineering

July 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月


利用動態攝影機實現人形追蹤與人臉偵測技術

研究生：朱琳達

指導教授：林進燈 教授

國立交通大學電機與控制工程研究所

摘 要



本論文提出一套基於適應性動態攝影機控制技術之智慧型人形追蹤與臉部偵測系統，在我們的系統架構下，整個系統可以區分成兩個部分。第一個部分是以影像為基礎之偵測模組，其中包含了人形和人臉的偵測；第二個部分則是針對 pan-tilt-zoom 攝影機所開發的控制模組。首先，我們利用人體的模型來辨識影像中存在的人形，再用膚色資訊和 connected component 找出人臉的位置。利用所找到人臉的資訊及相關特徵，我們完成了一個可用來判斷人臉是否清晰可見的神經網路系統。因此，本論文所提出之技術不但可將被追蹤的人保持在攝影機可視範圍的中間；當移動中人員的人臉不夠清楚或是尺寸小於我們所設定臨界值的時候，系統會自動調整攝影機焦距的倍率，以便得到清晰可辨識的人臉影像。

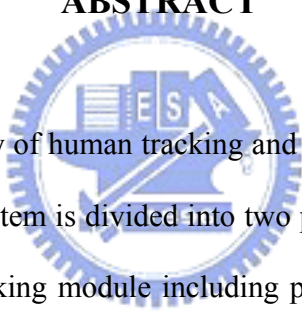
A Study of Human Tracking and Face Detection on A Pan-Tilt-Zoom Camera

Student: Linda Siana

Advisor: Prof. Chin-Teng Lin

Department of Electrical and Control Engineering
National Chiao-Tung University

ABSTRACT

The logo of National Chiao-Tung University is a circular emblem with a gear-like border. Inside the circle, there are stylized Chinese characters and the letters 'ES' and 'A'.

This thesis presents a study of human tracking and face detection on a pan-tilt-zoom camera. In our approach the system is divided into two parts: detection module including human and face detection, tracking module including pan-tilt-zoom camera control. We model a human for human recognition and combine HSI skin color with connected components for face detection. HSI skin color is implemented by using multilayer neural network trained with back-propagation learning algorithms. We use the recognizable face index to describe the definition of detected faces. Active camera will track a human and keep it in the FOV (Field of View) camera. The controls in camera zooming will be automatically started when the detected face of moving humans is not clear or the face index is smaller than a threshold value. We also apply local motion vectors to increasing the capability of our tracking system.

Acknowledgements

First and certainly foremost, I wish to acknowledge my advisor, Dr. Chin-Teng Lin for his support, encouragement, insight, patient guidance and invaluable contribution throughout this study. I was always supported both financially and with the equipment my research required. I would like to thank my Vision Laboratory leader, Her-Chang Pu, for his confidence in my abilities. He was always willing to discuss any type of problems with a positive attitude. I would like to thank the National Chiao Tung University for the scholarship and financial support so I can continue and finished my master degree. I would like to thank Dr. Kai-Tai Song, Dr. Jen-Hui Chuang, Dr. Cheng-Jian Lin, and Dr. Der-Jenq Liu, for their comments and suggestions for editing my thesis.

I also would like to thank my colleagues in Vision Laboratory, Yu-Wen Shou, Kan-Wei Fan, for sharing experience and knowledge during the time of my research, my junior master student Ting-Wei Mei, for help me to do my experimental as a model and discuss any problems related with the experimental results and all my friend that have not been list.


Finally, I dedicate this thesis to my parents, grandmother and auntie for their love, support, confidence and providing me with the opportunity to undertake my studies at National Chiao Tung University, Taiwan. They have provided me with endless support and opportunity to reach new heights.

Linda Siana

National Chiao Tung University

August 2005

Contents

Chinese Abstract.....	i
Abstract.....	ii
Acknowledgements.....	iii
Contents.....	iv
List of Figures.....	vi
List of Tables.....	ix
	
Chapter 1 Introduction.....	1
1.1 Motivation.....	2
1.2 Related Work.....	3
1.2.1 Human Detection Method.....	3
1.2.2 Face Detection Method.....	4
1.2.3 Tracking System.....	7
1.2.4 Zooming System.....	8
1.3 Thesis Organization.....	9
Chapter 2 System Overview.....	10
2.1 Hardware Architecture.....	10
2.2 Software Architecture.....	15

Chapter 3 Detection Algorithm.....	18
3.1 Human Detection	18
3.1.1 Segmentation of moving object	20
3.1.2 Human recognition.....	21
3.2 Face Detection	27
3.2.1 Skin Color	28
3.2.2 Connected Component.....	33
Chapter 4 Tracking System.....	39
4.1 Active Camera	40
4.2 Tracking and Zooming Process	41
4.2.1 Recognizable Face Index.....	44
4.2.2 Local Motion Vector.....	49
Chapter 5 Experimental Results.....	53
5.1 The Experimental System.....	53
5.2 Environment Setup.....	54
5.3 Experimental I.....	55
5.4 Experimental II	58
5.5 Experimental III	63
Chapter 6 Conclusion and Future Work.....	64
6.1 Conclusions.....	64
6.2 Future Works	65
References.....	66

List of Figures

Figure 1.1	Face detection divided into approaches [7]	6
Figure 2.1	System architecture	10
Figure 2.2	Packet structure of RS-232	11
Figure 2.3	(a) The rotation angle of pan range, (b) The rotation angle of tilt range	11
Figure 2.4	Face detection and Tracking System	15
Figure 2.5	Flowchart system	16
Figure 3.1	Human detection system	19
Figure 3.2	(a) Current frame, (b) Previous frame, (c) Difference image, (d) Difference image after thresholded	19
Figure 3.3	Image Projection, (a) Difference image, (b) Horizontal projection, (c) Vertical projection, (d) Projection result	21
Figure 3.4	(a) Division of body according to <i>Canon</i> by Polycleitus, (b) Modern view of human proportions	22
Figure 3.5	Simplified graph of human body ratio	23
Figure 3.6	The measure of human body ratio in Korea	23
Figure 3.7	(a) Human detection and recognition, (b) Parameters of ellipse	25
Figure 3.8	Ellipse human models	26
Figure 3.9	Face detection system	28
Figure 3.10	HSI skin color, (a) training part, (b) testing part	31
Figure 3.11	HSI training set	32
Figure 3.12	MSE of training process	32

Figure 3.13	(a) Arrangement of pixels, (b) pixels that are 4-connectivity, (c) pixels that are 8-connectivity, (d) m-connectivity [48]	34
Figure 3.14	(a) 8-connectivity, (b) Neighbors of 8-connectivity (a) 4-connectivity, (b) Neighbors of 4-connectivity	36
Figure 3.15	Labeling process	36
Figure 3.16	Connected component (a) skin color region, (b) connected component result.....	37
Figure 3.17	Connected component labeling [50]	37
Figure 4.1	Tracking system	39
Figure 4.2	FOV camera	40
Figure 4.3	Step-size regions	41
Figure 4.4	Tracking and zooming process	42
Figure 4.5	Zooming region.....	43
Figure 4.6	Training of recognizable face index	43
Figure 4.7	Testing of recognizable face index	44
Figure 4.8	Recognizable face index training.....	45
Figure 4.9	(a) Original image, (b) After Filtering, (c) After filtering and binerization	45
Figure 4.10	Training face images, (a) side-view faces with percentage index 20%, (b) side-view faces with percentage index 50-60%, (c) front-view faces with percentage index 90%.....	46
Figure 4.11	Recognizable face index	47
Figure 4.12	Zooming procedure.....	48
Figure 4.13	Region of motion vector	49

Figure 4.14	Difference between two frames	50
Figure 4.15	(a) Previous frame, (b) Current frame.....	51
Figure 4.16	(a) Current frame, (b) Previous frame, (c) Difference between two frames.....	51
Figure 4.17	Local motion vector result	52
Figure 5.1	Indoor environments at reference focal length 3.1 mm	54
Figure 5.2	Frames at several focal lengths.....	55
Figure 5.3	Tracking without zooming operation.....	56
Figure 5.4	Human and Face detection sequence	57
Figure 5.5	Tracking and zooming image sequence (zoom-in).....	58
Figure 5.6	Tracking and zooming image sequence (zoom-in and zoom-out).....	59
Figure 5.7	Tracking and zooming step-4.....	60
Figure 5.8	Human tracking and face extraction (1).....	61
Figure 5.9	Human tracking and face extraction (2).....	62
Figure 5.10	Moving object	63

List of Tables

Table 2.1	Camera control specifications.....	11
Table 2.2	RS-232 commands.....	13
Table 2.3	Camera focal length.....	14



Chapter 1

Introduction

Detection and tracking of moving object are important task in computer vision field particularly for visual-based surveillance system. Video surveillance system has been generally applied in recent years. The strong need of it comes from security-sensitive areas such as banks, department store, parking lot, etc. There are two types of video surveillance system: video surveillance system with continuously human monitoring but finding available human resources to sit and watch that imagery is expensive otherwise it can not issue the alarm in real-time. The other is video surveillance system recorded video by camcorder without human monitoring. It is already prevalent in commercial establishment with camera output being recorded periodically or stored in video archives. The system is more efficient and effective if has a camera which can track the movement of human and detect the face because face is important information of human then the system will store or send face image to alert security officers.

Visual tracking has been important topic in computer vision and surveillance system. There are two different categories of tracking, tracking a target with stationary camera and active camera. An active camera can move up-down direction (tilt), left-right direction (pan) by controlling the motor and has ability to do zoom-in/out. Tracking with active camera can keep the target in the camera scene wherever the target moves by

driving the pan-tilt-zoom. Visual tracking is implemented in indoor and outdoor, moving object in these environments are varying such as trees, people, vehicles, animals, etc. and tracking system must distinguish the object that we want to track with other objects. In this thesis we integrated the real time human detection, face detection, tracking human and zooming human face using an active camera (pan-tilt-zoom camera) to achieve the goal of surveillance system.

1.1 Motivation

Face detection determines the location and size of each human face in an image. A successful face detection system can provide valuable insight on how one might approach other similar feature and pattern detection system. The pre-process of face detection is to extract moving object from background (human detection system). There are many existing of human detection system [11, 12, 13, and 14] are used to detect the human that presence in an image but usually the size of human too small to process so the face image somehow is not clear or blur and has low resolution.

Face tracking is used to follow a face through the sequence. There are two types of real-time tracking system that incorporate the face changes over time, in term of changes in scale, position and localize the search for the face. One is to track a target with fixed camera [11, 15, 16, 17, 18, and 19]. The other is to track an object by active camera [20, 21, and 22]. Tracking face with an active camera can keep the object in the scene of camera by controlled camera motor so the object can not disappeared from scene and some active cameras have zoom-in/out operation that can be used to zoom-in/out the object when it is too small to track. Many face detection and tracking work for front view

face although the face has several angles and scales but in real-environment is impossible to detect only front view of face, somehow face or head can be sided-view, backed-view or several angles of view.

Consider for instance video surveillance the images are not necessarily monitored continuously. In many such systems each camera covers a fixed-large area. This has the serious drawback that the resolution of the video pictures is frequently too poor to be able to extract anything but the most basic conclusions as to the appearance of the person in question and as such is virtually useless for convicting criminals in courts. If instead the camera is able to zoom-in on the suspect, the problem of insufficient image resolution can be solved [46].

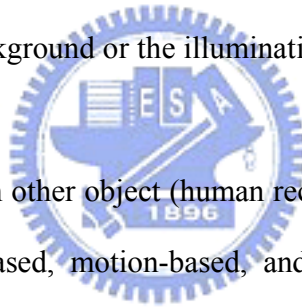
The several reasons above motivates us to develop a face detection and tracking system use an active camera with controlling pan-tilt-zoom, so it can track and zoom the human or human face if the size too small to process. In human detection we use deformable model to modeling the human. This human model can be used to detect moving human that has variation of size. In face detection system we combine skin color, connected component, and recognizable face index to define clearance index of face which can be used for tracking and zooming process.

1.2 Related Work

1.2.1 Human Detection Method

In recent years, many human detection approaches have been developed. There are two parts of human detection system, segmentation of moving object from background

and human detection by distinguishing the human with other moving objects. Several methods for moving object segmentation are optical flow method [24, 25, and 26], stereo based vision, and temporal difference method. Optical flow is used to detect independently moving objects but it has complex computation and sensitive to change of intensity. Optical flow used in [25, 26] used to detect vehicle. Zhao et al [11] exploited stereo based segmentation algorithm to extract object from background and to recognize the object by neural network based recognition. Although stereo vision based technique have been proved to be more robust it requires more than one camera at least two camera and can be used only for short and middle distance detection. Smith et al [27] used background subtraction method to segment isolate human. The serious problem of this approach is the changeable background or the illumination that is almost different in each frame.



To distinguish human with other object (human recognition), several method can be implemented such as shape-based, motion-based, and multi-cue based methods. The shape-based approach uses shape feature to recognize human. Motion based use Fast Fourier Transform and its periodicity against time [28]. Some system integrate multiple feature to recognize human such shape pattern, motion pattern, skin color, etc. Curio et al [29] used the initial detection process that is based on geometry feature of human. Then, motion patterns of limb movements are analyzed to determine initial object hypotheses.

1.2.2 Face Detection Method

In recent years face detection including face recognition and facial expression have attracted much attention though they have been studied for more than 20 years by

psychophysicists, neuroscientists, and engineers. A first step of face processing system is detecting the locations in images where faces are present [30]. Two face detection approaches are feature based and image based approach [7]. Feature based approach divided into three areas: low level analysis, feature analysis and active shape model. Image based approaches divided into linear subspace method [32, 34], neural networks [30, 38] and statistical approach [37]. Most of it apply a window scanning algorithm is in essence just an exhaustive search of the input image for possible face locations at all scales. Figure 1.1 shows details the existing techniques to detect face.

Low level analysis first deals with segmentation of visual features using pixel properties such as gray-scale and color. Edge in low level analysis was applied in earliest face detection work by Sakai et al [31]. In this method, edge need to be labeled and matched to a face model in order to verify correct detection. Beside edge details, gray information within a face can also be used as features. Facial features such as eyebrows, pupils, and lips appear generally darker than their surrounding facial regions [32, 33].

Another low level analysis is color method. Human face can be modeled as skin color model. Some methods of skin color that have been used for detect face are RGB color model, HSI color model, YCbCr, etc. [32-36]. If the use of video sequence is available, motion information is a convenient means of locating of moving object. The straightforward way to achieve motion segmentation is using frame difference analysis. This approach is simple and able to discern a moving foreground efficiently regardless of the background content [7].

Features generated from low level analysis are likely to be ambiguous. For instance, in locating facial regions using skin color model, background objects of similar color can

also be detected. In many face detection techniques, the knowledge of face geometry has been employed to characterize and subsequently verify various features from their ambiguous state. There are two approaches in face geometry, feature searching based on the relative positioning of individual facial feature [33, 34]. The confidence of feature existence is enhanced by the detection of nearby features. The technique in the second approach group features as flexible constellations is using various face models [34]. Active shape models depict the actual physical and hence higher level appearance of features. Once released within a close proximity to a feature, active shape model will interact with local image feature like edge and brightness then gradually deform to take the shape of the feature [7].

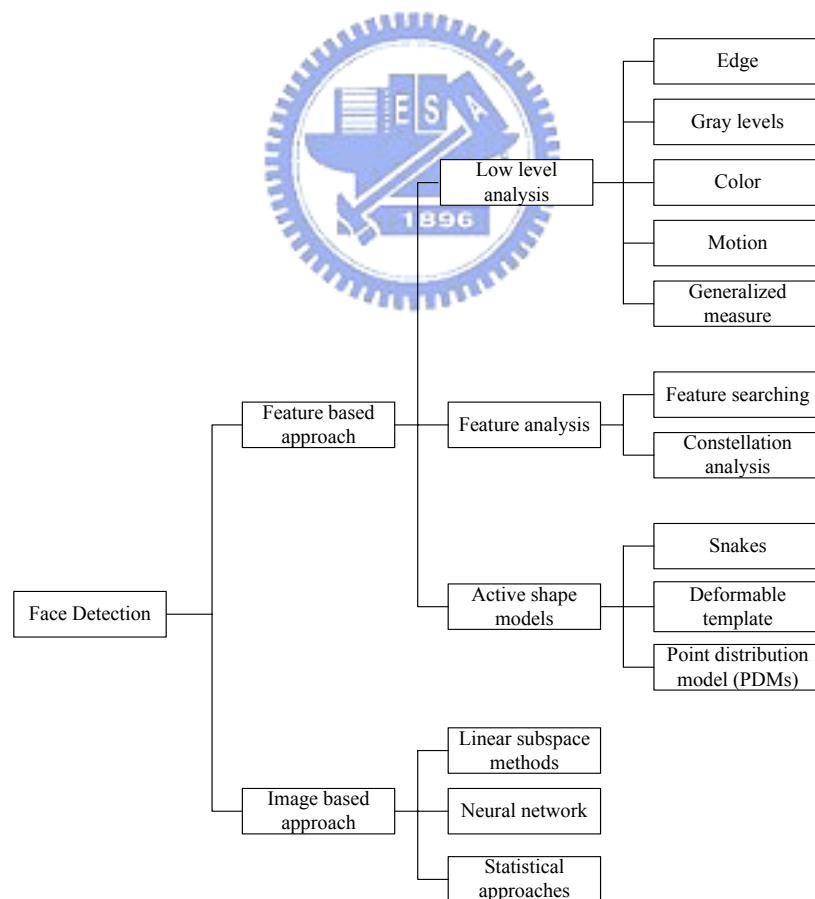


Figure 1.1 Face detection divided into approaches [7]

There are generally three types of active shape models, the first type uses a generic active contour that is called *snakes* and first introduced by Kass et al. *Deformable template* were then introduced by Yuille et al to take into account the a priori of facial features and has better performance of snakes. Cootes et al later proposed the use of new generic flexible model which they termed smart *snaked* and provide an efficient interpretation of the human face.

1.2.3 Tracking System

Tracking is used to follow an object through the sequences in this case the object is face or moving human that changes over time, in terms of changes in scale and position, and to localize the search for the face. As mentioned in motivation and contribution part, there are two type of real-time tracking system. One is tracking object with fixed camera and the other is with an active camera.

A real-time detection system for color image sequence is present by G. L. Foresti [10]. The approach is using in video-based surveillance system for monitoring indoor scene. J. Steffens [14] presents real-time face detection and tracking system for face recognition by using stereo CCD camera to capture an image. L. Wang [8] presents a real-time face tracking by using digital camera which fixed on a tripod. For single tracking, it provides a robust tracking result even in presence of slight scale and orientation changes of human face. K. Schdwerdt and J. L. Crowley [40] discuss a new robust tracking technique by applying histogram of intensity normalized color. The face tracking procedure described in this method has certain advantages, such as grater stability, higher precision and less jitter, over conventional tracking techniques using

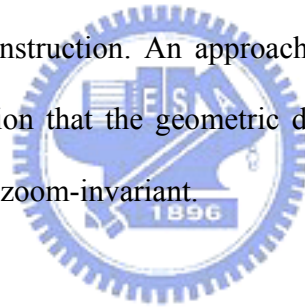
color histograms. R. C. Verma et al. [5] present a new probabilistic method for detecting and tracking multiple faces in a video sequence. D. Comaniciu and V. Ramesh [21] present a real-time system for detection and tracking human face with an active camera, this paper focuses on techniques that prune the face candidates for case when the background color is similar to the skin color. J. Yang and A. Waibel [39] present a real-time face tracker using a stochastic model to characterize skin-color distribution of human face. The frames are captured by active camera, this system hardly achieve good performance because time delay of pan-tilt-zoom camera control. M. Scheutz et al [13] present a real-time system for a mobile robot that can reliably detect and track people, the camera has pan-tilt controller.

1.2.4 Zooming System

On automated tracking of scene motion, auto zoom control is a rather unexplored area [45]. However, it is an area which progress over the last few years in the theory of structure from motion and self calibration of cameras lays open to practical investigation. Under human operator control, camera zooming is initiated in two ways [45], the first is called purposeful zooming, where some higher level process indicates that it would be valuable either to zoom-in to collect more object details, or to zoom-out to obtain surrounding context. The other way zooming is used is more reactive. In this case, the camera operator adjusts the zoom to preserve the image size of the target object as it moves away from or towards the camera.

Many tracking system are used active camera work only on pan-tilt although the camera has zoom operation otherwise using zoom without pan-tilt operations. Some

literatures [45, 46, 51] work in zoom operation by using affine camera. B. Turdoff and D. Murray present a method for visual control of the zoom setting of an active camera during tracking. The method assumes an affine projection and tracking achieved using affine transfer, a process which is fundamentally invariant to zoom. H. Shah and D. Morrell [51] proposed an adaptive zoom by adaptively changing the camera focal length. The target tracker is implemented using Rao-Blackwellized particle filter. The simulation of it demonstrates that the adaptive zoom algorithm has a smaller average squared position estimate error than a comparable fixed zoom algorithm. The proposed system is doing tracking by using pan-tilt-zoom. E. Hayman [46] developing and analyzing algorithms that function in spite of zoom, algorithm for visual tracking, camera calibration and Euclidean reconstruction. An approach grounded in visual geometry is adopted, motivated by the notion that the geometric descriptions of point (corner) and line (straight edge) features are zoom-invariant.



1.3 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 describes system overview including software and hardware architecture. Chapter 3 shows detection system including human and face detection system. Chapter 4 describes the tracking, zooming module and camera control system. Chapter 5 shows the experimental results. Chapter 6 is the conclusions of this thesis and the future works.

Chapter 2

System Overview

2.1 Hardware Architecture

The system uses an active camera Sony EVI-D100/P that has pan-tilt-zoom function to acquire image frames. These frames are captured and processed by Personal Computer (PC). The specification of the computer is AMD Athlon XP 2000+ 1.67 GHz, 512 Mb RAM. As shown in Fig. 2.1, the active camera has two interfaces which are RS-232 and video interface. RS-232 interface is used to send a command to control the camera movement including pan-tilt and zoom operation. Meanwhile video interface is an analog input that needs video grabber card, so PC (personal computer) can read out the image data.

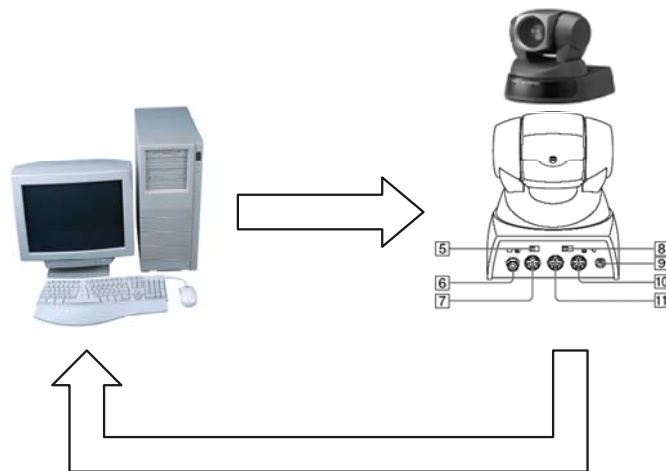


Figure 2.1 System Architecture

The parameters of RS-232 are communication speed 9600 bps, 8 data bits, 1 start bit and stop bit, non parity and MSB first. The packet structure of it shows in Fig. 2.2.

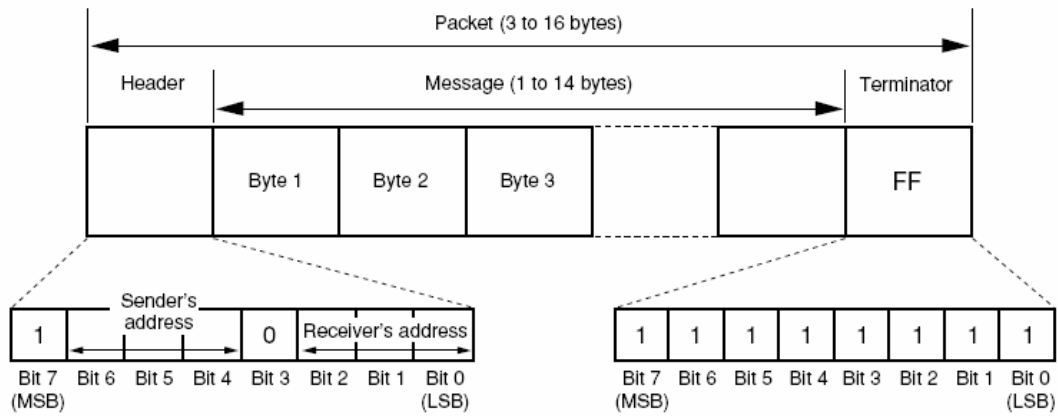


Figure 2.2 Packet structure of RS-232

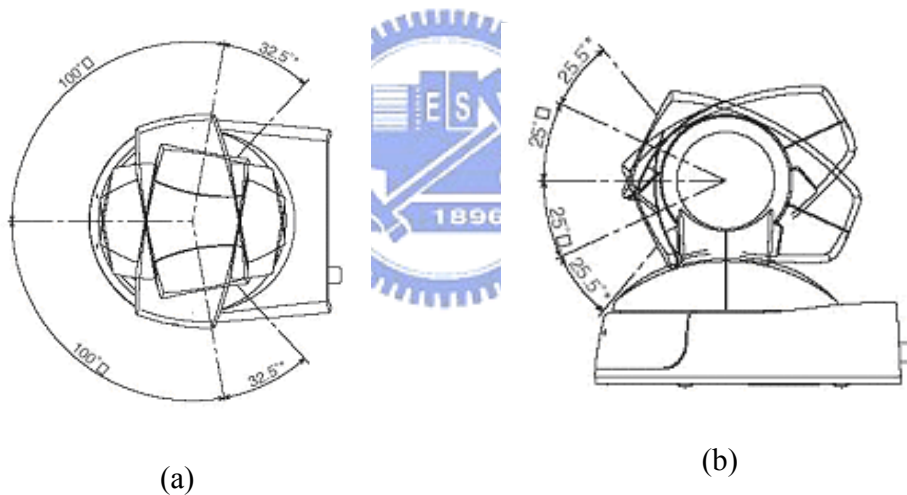


Figure 2.3 (a) The rotation angle of pan range, (b) The rotation angle of tilt range

Table 2.1 Camera control specifications

Camera control	Angle	Maximum speed
Pan	Horizontal ± 100 degrees	300 degrees/s
Tilt	Vertical ± 25 degrees	125 degree/s

In order to track object (human/face) we control the angle of pan-tilt and focal length of zoom-in/out. The specification of camera control shows in Table 2.1 and Fig. 2.3. EVI-D100/P uses 10x optical zoom lens and its digital zoom function allows zoom up to 40x. The horizontal angle of view is approximately 65 degrees (wide-end) to 6.6 degrees (tele-end). The digital zoom increases the picture element size and reduces the resolution. The camera focal length is 3.1-31 mm and the object is two meter away from the front surface of the lens, and minimum object distance WIDE (zoom-out) 100mm, TELE (zoom-in) 600mm.

Table 2.2 shows some RS-232 commands for drive pan-tilt-zoom camera and its return commands. The pan-tilt camera moves by its relative position so it moves from current position to desire position by using different reference coordinate on the other hand we take current position as reference position. It is different with absolute position condition which still has same reference coordinate wherever camera moves.

Home command is used to drive pan-tilt to original reference position in 3D coordinate the position is (0,0,0). Reset command is used to drive pan-tilt to maximum and minimum position. Pan control is used to drive camera to left and right side in the other hand tilt control is to drive camera to up and down side. When pc sends a command through RS-232, it will return two commands to indicate that has been accepted and executed (completion message) otherwise RS-232 will return an error message. When command messages are sent to camera, it is normal to send the next command message after waiting for the completion message or error message to return, however EVI-D100/P has two buffers (memories) for commands so within these buffers the camera can execute command and receive new command at the same time.

Table 2.2 RS-232 commands

Command Set	Command	Command Packet	Comments
Pan-tilt drive	Relative position	81 01 06 030VV WW 0Y 0Y 0Y 0Y 0Z 0Z 0Z 0Z FF	VV: pan speed 01 to 18 WW: tilt speed 01 to 14 YYYY: pan position FA60 to 05A0 (center 0000) ZZZZ: tilt position FE98 to 0168 (center 0000)
	Home	81 02 06 04 FF	Back to origin position (x,y,z)=(0,0,0)
	Reset	81 01 06 05 FF	
Zooming	Direct	81 01 04 47 0p 0q 0r 0s FF	pqrs: zoom position
Command Set	Command	Reply message	Comments
General command	81 01 04 38 02 FF	90 41 FF / 90 42 FF(ACK) 90 51 FF / 90 52 FF(completion)	- Return ACK when a command has been accepted - Return completion when a command has been executed
Pan-tilt status	81 09 06 10 FF	90 50 pq rs FF	pq rs=02 04: pan-tilt moving pq rs=02 08: pan-tilt operation is completed
Zoom position	81 09 04 47 FF	90 50 0p 0q 0r 0s FF	pqrs: zoom position

When pan-tilt camera is still moving the system does not send another command until RS-232 return a command that indicate pan-tilt operation is completed and zooming operation also do the same thing. The system will process new zooming commands until RS-232 returns a command that zooming process is already achieved the desired focal length. The zoom has four steps depend on the focal length value which camera can be used to do zoom-in or zoom-out on this focal length. Table 2.3 shows the camera focal length.

The camera in reference condition or condition without zooming is using focal length equal to 3.1 mm. The zoom-in step one will increase focal length from 3.1 mm to 4.65 mm, zoom-in step two will increase focal length from 4.65 mm to 6.2 mm, and so on. On the other hand, zoom-out step one will decrease focal length from 4.65 mm to 3.1 mm, zoom-out step two decrease focal length from 6.2 mm to 4.65 mm, and so on.

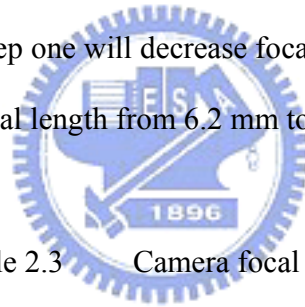


Table 2.3 Camera focal length

Step	Zoom-in (focal length-mm)	Zoom-out (focal length-mm)
1	4.65	3.1
2	6.2	4.65
3	9.3	6.2
4	12.4	9.3

2.2 Software Architecture

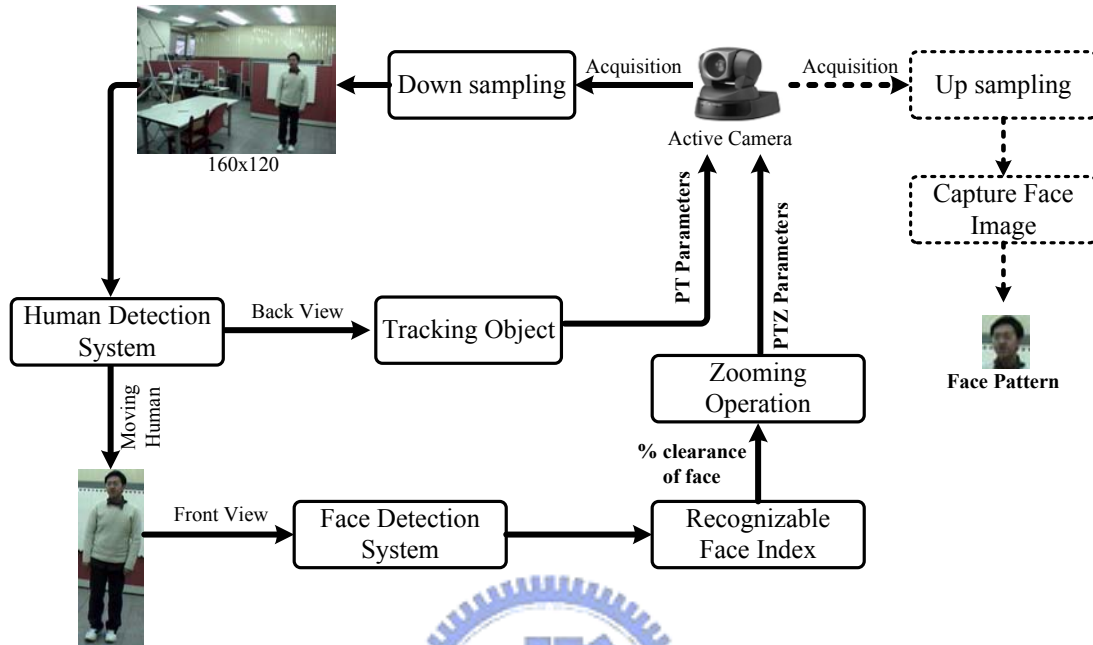


Figure 2.4 Face detection and Tracking System

Input frames are captured by Sony active camera EVI-D100/P with resolution 320x240 then pre-processed it by using down-sampling. Its result is processed in human detection module to extract the moving human region as shown in Fig. 2.4. The system use face detection module to extract human face and apply recognizable face index to calculate percentage face index that indicate clearance of human face for zooming and tracking application. The front view and back view in Fig. 2.4 are the conditions of moving human depend on human face position to camera. On back-view condition, the system only do tracking otherwise front-view condition the system will do tracking and zooming then the system will capture face image region for further application of surveillance system such as face recognition.

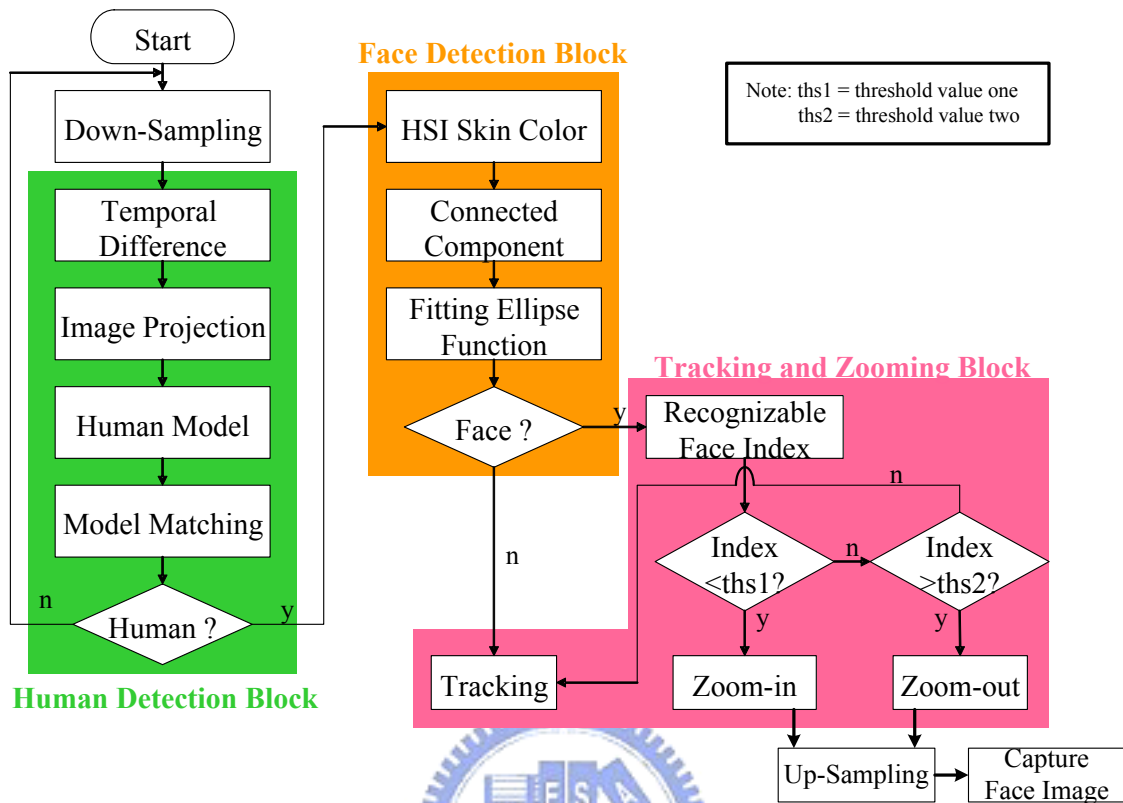


Figure 2.5 Flowchart system

Figure 2.5 shows flowchart of our system. The system is divided into three blocks: human detection block, face detection block, tracking including pan-tilt-zoom camera control. The left block is human detection block, up-center block is face detection block and block under face detection is tracking and zooming block. The human detection system consists of temporal difference, image projection, human model and model matching. The human detection result is moving human bounding box and its position in an image. The face detection system consists of HSI skin color, connected component and fitting ellipse function, the result is human face location. The tracking and zooming block consist of recognizable face index, pan-tilt-zoom camera control and local motion vector.

The human detection system is used to detect moving human in a scene and distinguish it from other objects. We use deformable human model to achieve this goal. Projection an image in human detection system is processed to find and extract location of moving object by applying change detection technique. It works by using difference between two frames: current frame $I(x,t_i)$ and previous frame $I(x,t_0)$ that representing in monitored scene. The thresholded output is a binary image represents the moving object.

The face detection system is used to extract face pattern from moving human. It applies HSI skin color that is trained by back-propagation neural networks to find human skin region and labeling it by connected component, its label is arranged from large size to small. We fit the best possible ellipse to every label to find the human face position. The clearance of face is obtained from recognizable face index. Large percentage index means face could be recognized than small percentage face index.

Sometimes some positions of moving human are back-view so the face detection system can not extract the human face. In this condition the system still can do tracking by using moving human information. In the human tracking system, camera drives pan-tilt to follow movement of human and will track face in a further step when the face is visible (front-view, several angle view of face). Figure 2.5 show the zooming operation block is worked automatically when face detection system obtain face index smaller than threshold one and larger than threshold value two. The zooming operation includes zoom-in and zoom-out that has four steps by change the focal length camera, its work step by step from zoom step one until step four and it will stop if the recognizable face index obtain face index between threshold one to threshold two.

Chapter 3

Detection Algorithm

In this chapter we will describe how to determine location of moving human and its face in image sequence. The detection system is divided into two parts: human detection and face detection. We use deformable human model for human detection and HSI skin color performed by using multilayer neural networks, connected component and fitting ellipse function for face detection.

3.1 Human Detection



Human detection system is subdivided into two parts: segmentation of moving object and distinguishing human with other objects using template matching. Figure 3.1 shows our human detection system. Successive two frames are taking as input of segmentation process which are current frame $f(n+t)$ and previous frame $f(n)$. These frames are captured by active camera with frame rate 30 frame/second after down-sample these frames by two, we apply segmentation process which consist of temporal difference, median filter and image projection. The result of its process is used to model a human. So by using the human model we can distinguish moving human from other moving objects.

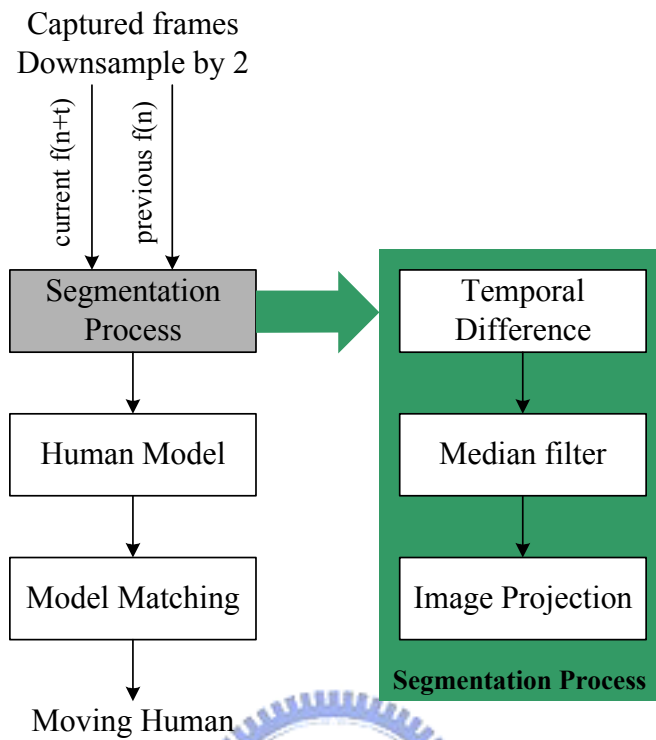


Figure 3.1 Human detection system

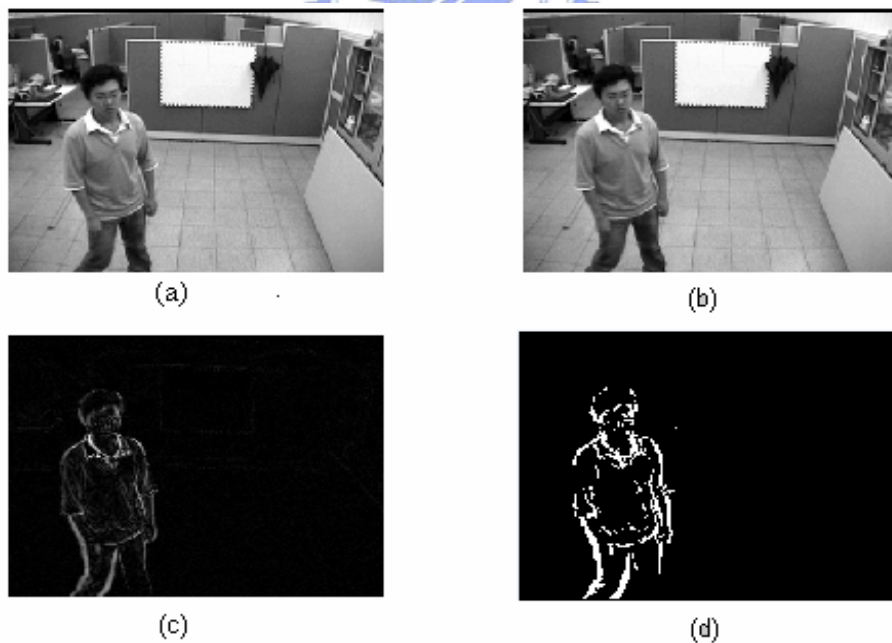


Figure 3.2 (a) Current frame, (b) Previous frame, (c) Difference image, (d) Difference image after thresholded

3.1.1 Segmentation of moving object

The main idea of segmentation moving object is to extract foreground from background. It consists of temporal difference, median filter and image projection. Temporal difference uses difference between current and previous frame as shown in Fig. 3.2. The input and result of difference image are gray-level images. After filtering the result is automatically thresholded and the output is a binary image as shown in Fig.3.2 (d) whose pixels can assume as two possible states, a static or moving pixel.

When foreground already extracted from background, size of image still same with input image and important data or moving object somehow is appeared in a region of an image which is shown in Fig. 3.2 (d). Then apply horizontal and vertical projection to extract only the moving object region. Vertical projection is used to find Xstart and Xstop position, and by applying these positions in horizontal projection we find Ystart and Ystop position. The result is a rectangular bounding box from left-top coordinate (Xstart,Ystart) to right-bottom coordinate (Xend,Yend) as show in Fig. 3.3(d).

Figure 3.3 consists of image after filtering and binerization which is same with image that show in Fig. 3.2(d), graph of horizontal projection, graph of vertical projection and bounding box of moving object in binary image and color image. The binary bounding box is extracted from difference image in Fig 3.3(a) and color bounding box is extracted from current frame shows in Fig 3.2(a).

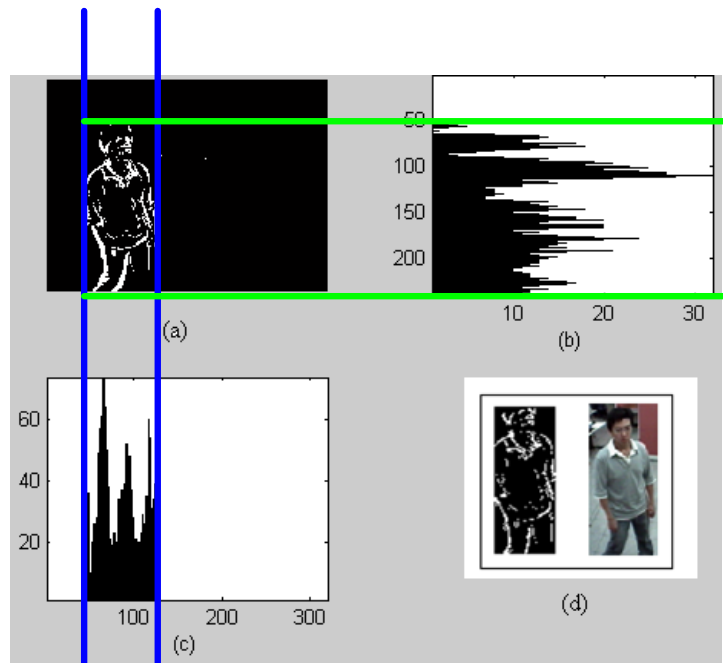


Figure 3.3 Image Projection, (a) Difference image, (b) Horizontal projection, (c) Vertical projection, (d) Projection result



3.1.2 Human Recognition

We apply human model to distinguish moving human from other moving objects. There are many methods in fields of human recognition. Its technique can be classifier into shape-based, motion-based and multi-cue-based. We combine multi-cue and human shape to model a human.

Around 540 B.C., Polycleitus published his treatise on ideal proportions of the human body known as the *Canon* [47]. He divided the human body into seven equal parts (seven and a half counting the foot), each part being equal to the height of the head. Modern view of human proportions in [47] is determined by related measurement one to another human body parts and the proportions put forth by Polycleitus.

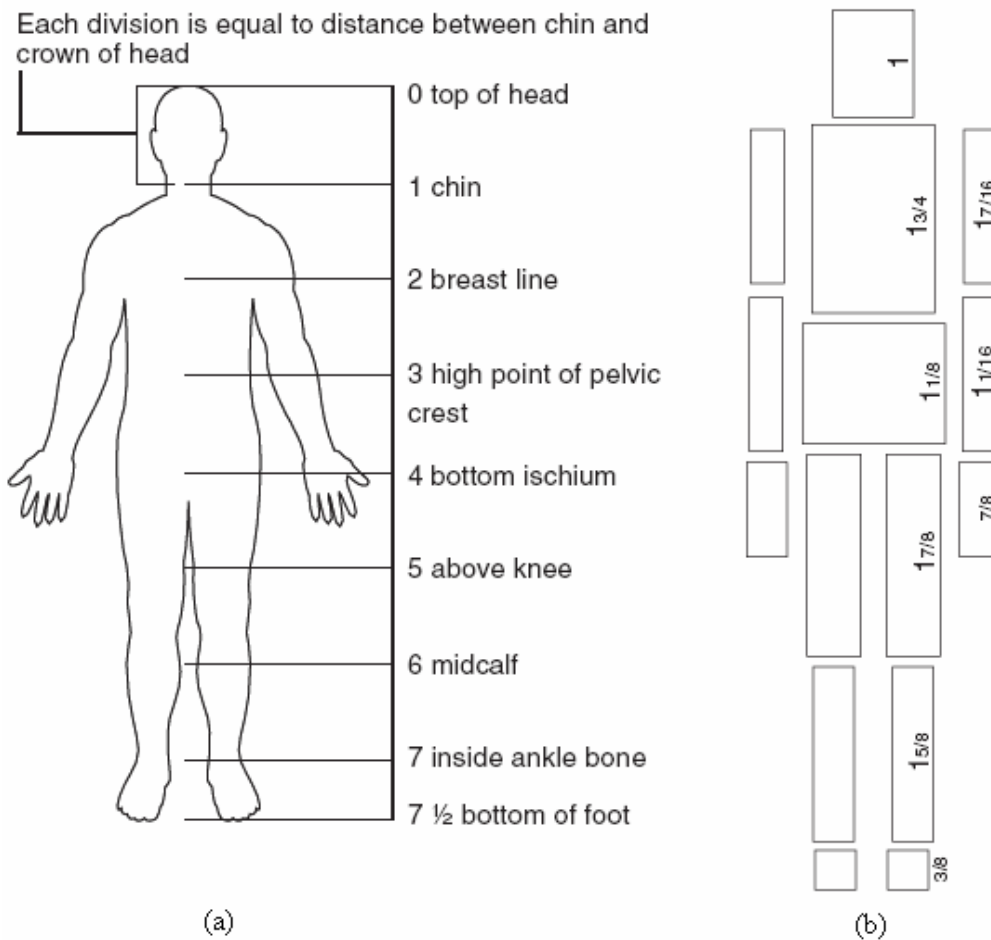


Figure 3.4 (a) Division of body according to *Canon* by Polycleitus, (b) Modern view of human proportions

The philosopher Leonardo da Vinci has indicated that the ratio of human being's head and body is 1 to 7. The head-body ratio structure is painted by Leonardo da Vinci with the architecture theory of Vitruvian. Although this artistic creation is not finished, it expresses the relation of human body clearly. Figure 3.5 is simplified graph of human body ratio and another human body ration is shown in Fig. 3.6. It is human body ratio has been used in Korea.



Figure 3.5 Simplified graph of human body ratio

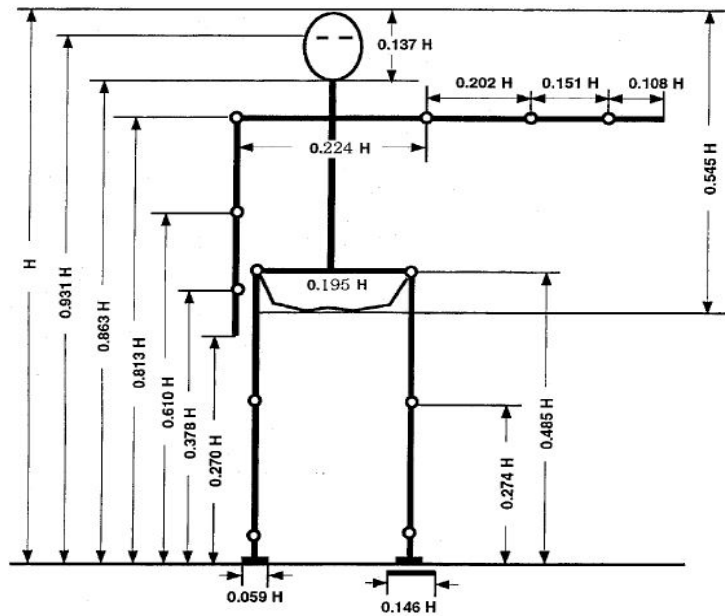


Figure 3.6 The measure of human body ratio in Korea

In human body dimensions the most interested parameters are height and width of human body. The length of head is proportional to the height of human body. The width parameter is defined as width of the torso. According to modern view and Leonardo da Vinci the ratio of head and body approximates to 1:7, so we use this ratio to model our human.

We model human shape by applying two ellipsoids corresponding to head and torso. It is because human head and torso are more fixed than other human parts such as arms and legs. Each ellipsoid is controlled by two parameters called long axis and short axis. The long axis parameter determines height of the head and torso, short axis parameter defines width of the head and torso. This model is sufficient to capture the gross shape variations of most humans in the scene. The length and width of head and torso are defined by proportional relationship of body dimension in human being. The human head and torso can be generally described as being roughly elliptic in nature. The elliptic parameters are,

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1 \quad (3.1)$$

Given point is (x,y) and the parameters are the center point (x₀,y₀), the semi major axis a, the semi major axis b of the ellipse.

The proposed human model is shown in Fig. 3.7. It is created by define moving object as height and width, and use human body proportional ratio which is approximately 1:7 then take 4/8 height of moving object to model human head and upper body. We modify general ellipse function to construct ellipse head model and upper body model.

$$\frac{\left(x - \frac{width}{2} + (o-oo)\right)^2}{\left(\frac{d}{2}\right)^2} + \frac{\left(y - \frac{3d}{4}\right)^2}{1.44\left(\frac{d}{2}\right)^2} \leq 4\left(\frac{d}{2}\right)^3 + 1.44\left(\frac{d}{2}\right)^4 \quad (3.2)$$

$$\frac{\left(x - \frac{width}{2}\right)^2}{\left(\frac{dd}{2}\right)^2} + \frac{\left(y - \frac{4}{5}dd + d\right)^2}{2.56\left(\frac{dd}{2}\right)^2} \leq 4\left(\frac{dd}{2}\right)^3 + 2.56\left(\frac{dd}{2}\right)^4 \quad (3.3)$$

Equation (3.2) is ellipse function for head of human model and Eq. (3.3) for upper side of human body. Width of head is (o-oo), diameter of head is d and diameter of body is dd. These parameters are shown in Fig. 3.7(b).

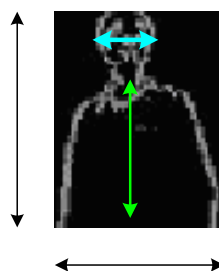
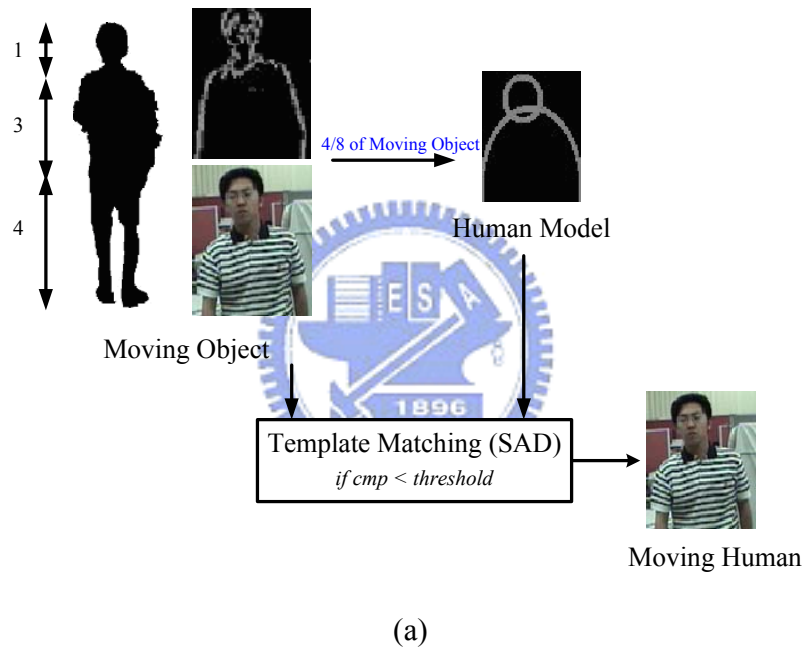


Figure 3.7(a) Human detection and recognition, (b) Parameters of ellipse



Figure 3.8 Ellipse human models

Figure 3.8 shows some ellipse human model. The size of head and body of human model changeable depend on size of head and body that is already detected in human detection system.

Correlation between moving object and human model is calculated by template matching, it is used to localize and to identify a pattern or template in a full content image. Nowadays, the template matching techniques is widely applied to the video coding in order to perform the faster compression, pattern recognition, and visual tracking. There are several methods that are popularly used for template matching: normalized cross-correlation (NCC) method, sum absolute difference (SAD), and sum squared difference (SSD). The mathematic formula of it as follow

$$NCC(x,y) = \frac{\sum_{j=0}^{B-1} \sum_{i=0}^{B-1} I(x+i, y+j)T(i, j)}{\|I\| \cdot \|T\|} \quad (3.4)$$

$$SAD(x, y) = \sum_{j=0}^{B-1} \sum_{i=0}^{B-1} |I(x+i, y+j) - T(i, j)| \quad (3.5)$$

$$SSD(x, y) = \sum_{j=0}^{B-1} \sum_{i=0}^{B-1} (I(x+i, y+j) - T(i, j))^2 \quad (3.6)$$

T and I denote the template and the image block in the same size of the template respectively. B denotes the size of the matching block, (x,y) is the reference point under current comparison process within the searching area, (i,j) is the local coordinate in the template and in the image block within the searching area with the reference point (x,y) . The larger value of NCC will have the higher similarity between the template and aforementioned image block becomes. Unlike NCC method, SAD and SSD are image subtraction method. If the value of SAD or SSD are closer to zero, then the image pattern are more similar. Consequently, the position (x,y) with respect to the maximum NCC value is the most likely position of the target that we want in the searching area, whereas the position (x,y) according to the minimum SAD or SSD value is the most likely the target position desirable.

The correlation of moving object in our approach is calculated by matching the human model and moving object using SAD as shown in Eq. (3.5), if the value of SAD is closer to zero, the moving object and human model is more similar.

3.2 Face Detection

The face detection system is shown in Fig. 3.9. It is divided into three blocks: HSI skin color connected component and fitting ellipse head model. The HSI skin color obtains three skin regions, head and two hands. The connected component will have three different labels and fitting ellipse function using these labels to find human head as shown in Fig. 3.9.

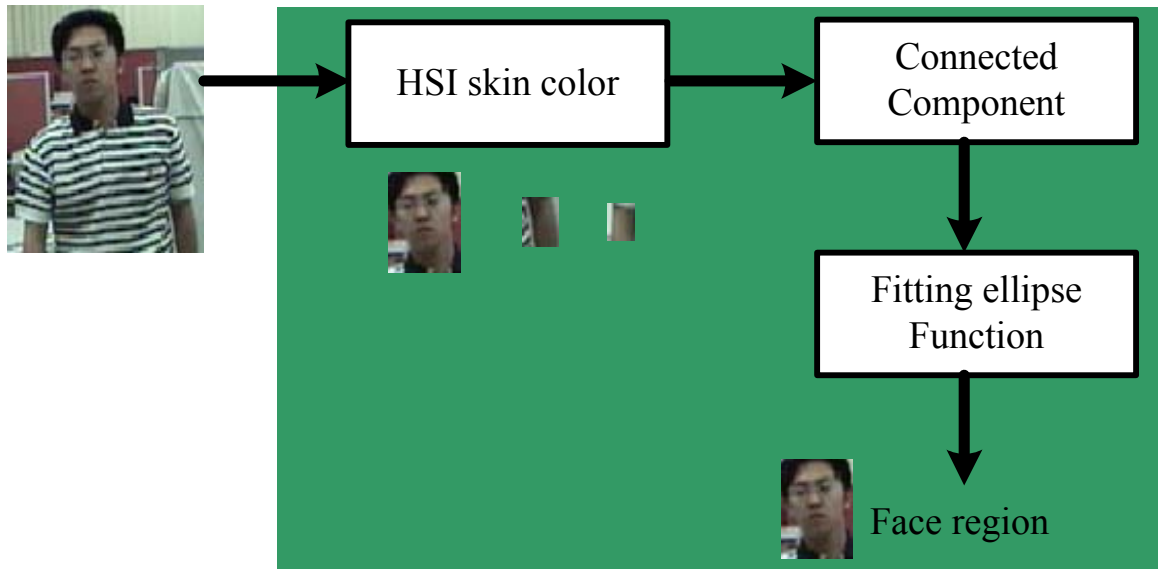


Figure 3.9 Face detection system

3.2.1 Skin color

Human skin color has been used and proven to be an effective feature in many applications from face detection to hand tracking [30]. Although different people have different skin color, several studies have shown that the major difference lies largely between their intensity rather than their chrominance. Several color space have been utilized to be label pixels as skin including RGB, HSV, YCbCr.

1. RGB skin color

In RGB color space, every pixel values are very sensitive to lightness and normalization RGB color space is reduced the effect of lightness. The transfer function is shown below,

$$R_n = \frac{R}{R + G + B} \quad (3.7)$$

$$G_n = \frac{G}{R+G+B} \quad (3.8)$$

The skin color distribution can be writing as combination of three functions below,

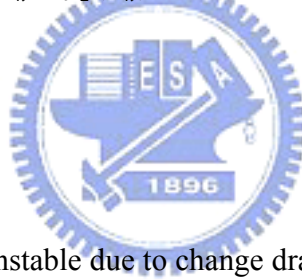
$$f_1(R_n) = -1.3767R_n^2 + 1.0743R_n + 0.1452 \quad (3.9)$$

$$f_2(R_n) = -0.776R_n^2 + 0.5601R_n + 0.1766 \quad (3.10)$$

$$w = (R_n - 0.33)^2 + (G_n - 0.33)^2 \quad (3.11)$$

RGB skin region represents as S region defined in Eq. (3.12)

$$S = \begin{cases} 1, & \text{if } (G_n < f_1(R_n) \& (G_n < f_2(R_n) \& w > 0.00004 \& (R > G > B)) \\ 0, & \text{Otherwise} \end{cases} \quad (3.12)$$



2. YCbCr skin color

Color appearance is often unstable due to change drastically under varying lighting or luminance condition. Consequently the original RGB image which incorporates luminance information must be transformed into a color space that separates the luminance and chrominance components (YCbCr). In this transformation, luminance information is stored in Y component and chrominance information is stored in Cb (chromatic blue) and Cr (chromatic red) components. Cb-Cr is natural model associated to jpeg images and for mpeg stream. Transformation from RGB color components to Cb-Cr color space is given by the following matrix:

$$\begin{pmatrix} Y \\ Cb \\ Cr \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (3.13)$$

3. HSI skin color

When humans view a color object, we describe it by its hue, saturation, and brightness [48]. Hue is color attribute that describes a pure color (yellow, orange, red). Saturation gives a measure of the degree to which a pure color is diluted by white light in the other hand, the saturation refers to the relative purity or the amount of white light mixed with a hue. Brightness is a subjective descriptor that is practically impossible to measure. It embodies the achromatic notion of intensity and is one of the key factors in describing color sensation. The following formulas show how to convert from RGB space to HSI:

$$\begin{aligned} I &= \frac{1}{3}(R + G + B) \\ S &= 1 - \frac{3}{R + G + B} [\min(R, G, B)] \\ H &= \cos^{-1} \left[\frac{\frac{1}{2} [(R - G) + (R - B)]}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \right] \end{aligned} \quad (3.14)$$

The HSI color space is an ideal tool for developing image processing algorithms based on color descriptions that are natural and intuitive to humans [48] and more stable due to drastically under varying lighting or intensity than RGB and YCbCr skin color model. For that reason our system is used HSI model to find skin region in an image. HSI skin color is performed by neural networks trained by back-propagation algorithm. The neural networks consist of offline training and online skin classification as shows in Fig. 3.10.

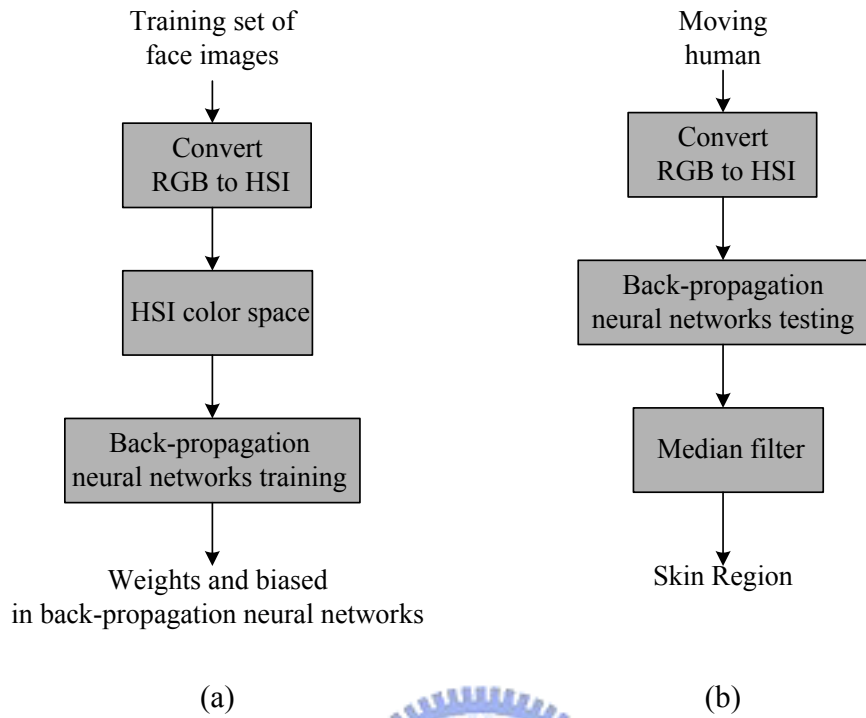


Figure 3.10 HSI skin color, (a) training part, (b) testing part.

The network consists of input layer, two hidden layers and output layer. Back-propagation algorithm is based on the error correction learning rule. The error signal at the output neuron j at iteration n is defined by

$$e_j(n) = d_j(n) - y_j(n) \quad (3.15)$$

$d_j(n)$ refers to the desired response for neuron j and is used to compute $e_j(n)$, $y_j(n)$ refers to output of neuron j at iteration n .

The model of each neuron in the neural networks includes a nonlinear activation function. The important point to emphasize here is the smooth nonlinearity [49]. A

commonly used form of nonlinearity that satisfies this requirement is a sigmoidal nonlinearity defined by the logistic function:

$$y_j = \frac{1}{1 + \exp(-v_j)} \quad (3.16)$$

The training set is composed of 250 RGB face images which have been extracted from real-time video capture then converts it to HSI color space by Eq. (3.14) and using its components as input training. Some of training set are shown in Fig. 3.11.



Figure 3.11 HSI training set

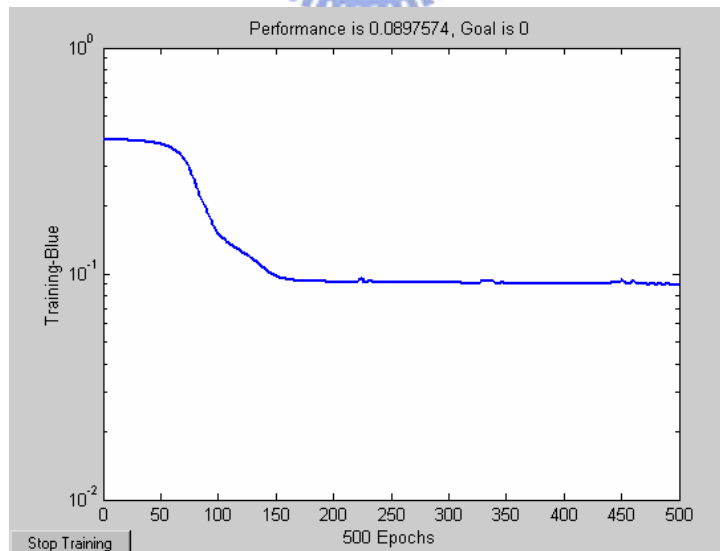


Figure 3.12 MSE of training process

Training part is processed to get the weighting and bias values. The mean square error (MSE) during training process is shown in Fig. 3.12. The goal of MSE is 0 and the training performance is 0.0898.

Testing part in Fig. 3.10(b) is applying in real-time system (online). The process is similar with training process. Input is moving human image and converts it into HSI color channel and classification by using weights and biases that obtained from training part then apply median filter to get smooth skin image and reduce some noise.

3.2.2 Connected Component

Connectivity between pixels is a fundamental concept that simplifies the definition of numerous digital image concepts, such as regions and boundaries. To establish whether these two pixels are connected, it is determined by their neighbors and finds their gray levels satisfy a specified criterion or similarity [48]. For instance, in binary image with values 0 and 1, two pixels maybe 4-neighbors, but they are said to be connected only if they have the same value.

Let V be the set of gray-level values used to define adjacency. In a binary image, $V = \{1\}$ if we are referring to adjacency of pixels with value 1. We consider three types of adjacency/connectivity [48]:

1. 4-connectivity

Two pixels p and q with values from V are 4-connectivity if q is in the set $N_4(p)$.

2. 8-connectivity

Two pixels p and q with values from V are 8-connectivity if q is in the set $N_8(p)$.

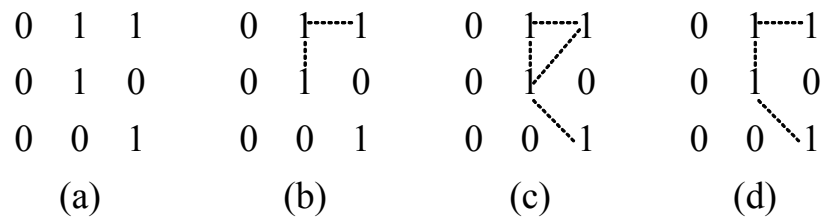


Figure 3.13 (a) Arrangement of pixels, (b) pixels that are 4-connectivity, (c) pixels that are 8-connectivity, (d) m-connectivity [48]

3. m-connectivity

Two pixel p and q with values from V are m-connectivity if

- (i) q is in $N_4(p)$, or
- (ii) q is in $N_D(p)$ and the set $N_4(p) \cap N_4(q)$ has no pixels whose values are from V .

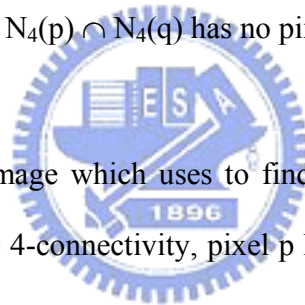


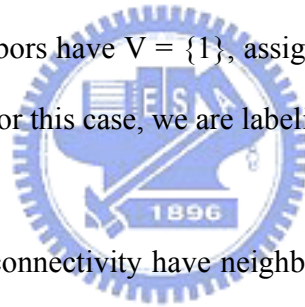
Figure 3.13(a) shows binary image which uses to find the connectivity between every pixel. Figure 3.13(b) shows the 4-connectivity, pixel p has 4-connectivity to its neighbor whose in horizontal or vertical position and contain $V = \{1\}$. If pixel p has connectivity to neighbor pixel in horizontal, vertical, or diagonal position, it will define as 8-connectivity. The last figure is m-connectivity, it is a modification of 8-connectivity which introduced to eliminate the ambiguities that often arise when 8-connectivity is used. The three pixels at the top of Fig.3.13(c) show ambiguous of 8-connectivity, as indicated by the dashed lines. This ambiguity is removed by using m-connectivity, as shown in Fig. 3.13(d).

Connected component works by scanning an image, pixel-by-pixel in order to identify connected pixel regions [50]. Its works on binary or gray-level images and different measures connectivity are possible. Choice of the connectivity is among 4, 8, 6,

10, 18, 26 connectivity which are 4 and 8-connectivity for 2D connected component extraction and the others for 3D connected component extraction. However in this thesis, the input is binary images and the connectivity is 8-connectivity. The connected components labeling operator scan the image by moving along a row until it comes to a point p where denotes the pixel to be labeled at any stage in the scanning process for which $V = \{1\}$. When it is true, it examines the four neighbors of p which already been encountered in the scan. Based on this information, the labeling of p occurs as follows:

[50]

- (i) If all four neighbors are 0, assign a new label to p , else
- (ii) If only one neighbors has $V = \{1\}$, assign its label to p , else
- (iii) If one or more of the neighbors have $V = \{1\}$, assign one of the labels to p and make a note of the equivalence. For this case, we are labeling p with minimum label value.



Connected component with 8-connectivity have neighbors as shown in Figure 3.14. The center of 3x3 mask is a pixel that we want to assign to a new label and it has eight neighbor which indexes '1' to '8'. When we scan an image from left to right and up to down, pixel which already labeled is shown in Fig. 3.14(b) which has indexes 1, 2, 3, and 4, and other neighbors have not assigned in a new label. So, based on this information, labeling rules will be valid in these neighbors. We take other connectivity such as 4-connectivity as another example and it shows in Fig 3.15 (c) and (d), the labeling rules will be valid in neighbor whose have index 1 and 2.

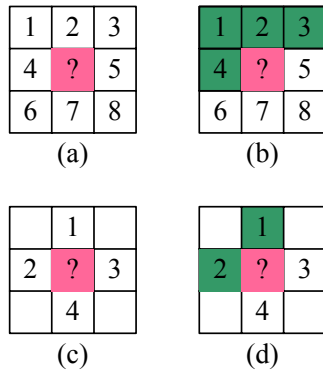


Figure 3.14(a) 8-connectivity, (b) Neighbors of 8-connectivity (a) 4-connectivity, (b) Neighbors of 4-connectivity

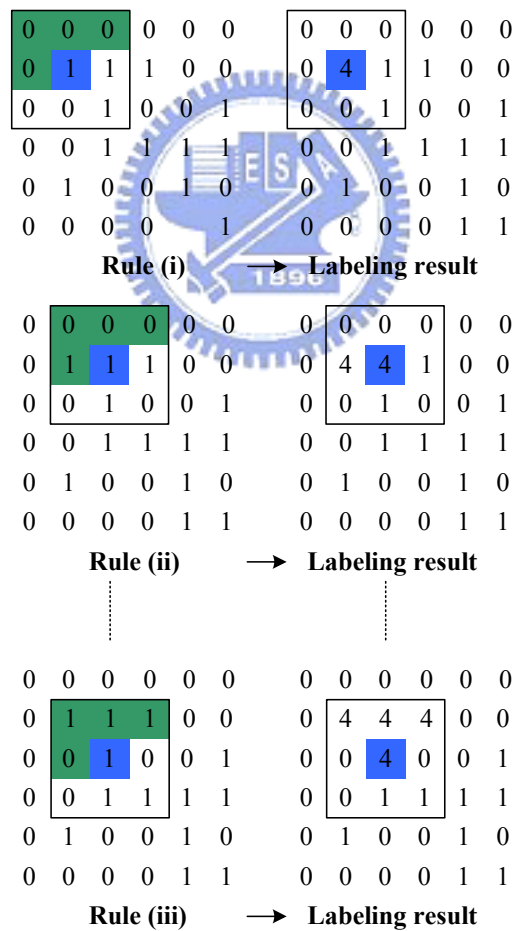


Figure 3.15 Labeling process

Figure 3.15 shows labeling process by using rule (i), (ii) and (iii). In rule (i) all neighbors of pixel p (center of 3x3 mask) is '0' so we assign it to a new label. This figure shows the new label is '4'. Rule (ii) is applying in next pixel. Pixel p in this position only has one neighbor with $V = \{1\}$, so we label pixel p same with its neighbor and the label is '4'. Scan image pixel by pixel, the last image is shown rule (iii) condition where pixel p has more than one neighbors contains $V = \{1\}$ so we label it to minimum label of its neighbors. In our example all of its neighbors have same labels '4' so pixel p is labeled to '4'. If there is another labels value such as '2' we take its label as pixel p label.

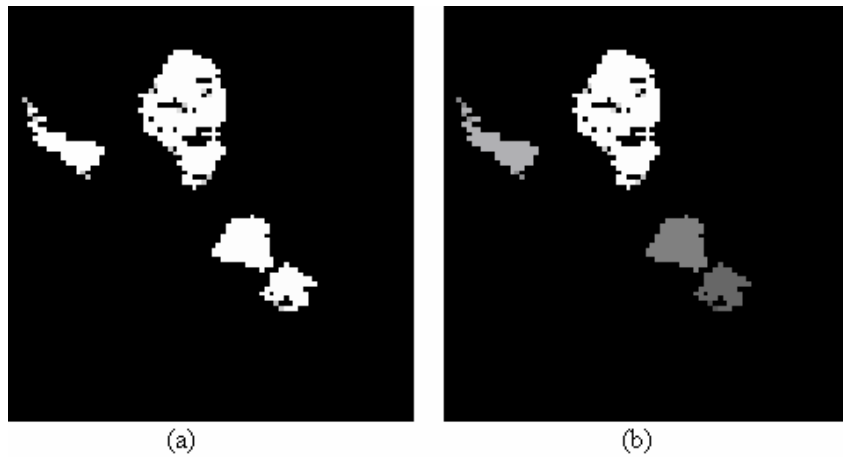


Figure 3.16 Connected component (a) skin color region, (b) connected component result



Figure 3.17 Connected component labeling [50]

Figure 3.16 and Fig. 3.17 show the result of connected component. Figure 3.16 shows skin regions obtained from HSI skin color and its connected component labeling result. It has four different colors which indicate different labels. The labels are arranged by large size to small which label '1' has largest size and label '4' has smallest size. Every label is fitting into ellipse model and this process is concerned to component that has small label or large skin region because face region is larger than human hand as shows in Fig. 3.16. The ellipse function is fitting to this label to find the face region. The fitting process will stop if there is obtained match skin region although the system still has another labels which its skin region smaller than match region.



Chapter 4

Tracking System

Tracking system uses PTZ (pan-tilt-zoom) camera to track object position on an image by drive pan-tilt camera to keep the object in FOV (field of view) camera. The system divided into two types, tracking using human information and face information. First condition will work automatically if face of moving human is back-view position and the other condition will work if face of moving human is front-view or side-view position. Figure 4.1 shows our tracking system module, it is included pan-tilt-zoom. The zooming system will work by using recognizable face index information.

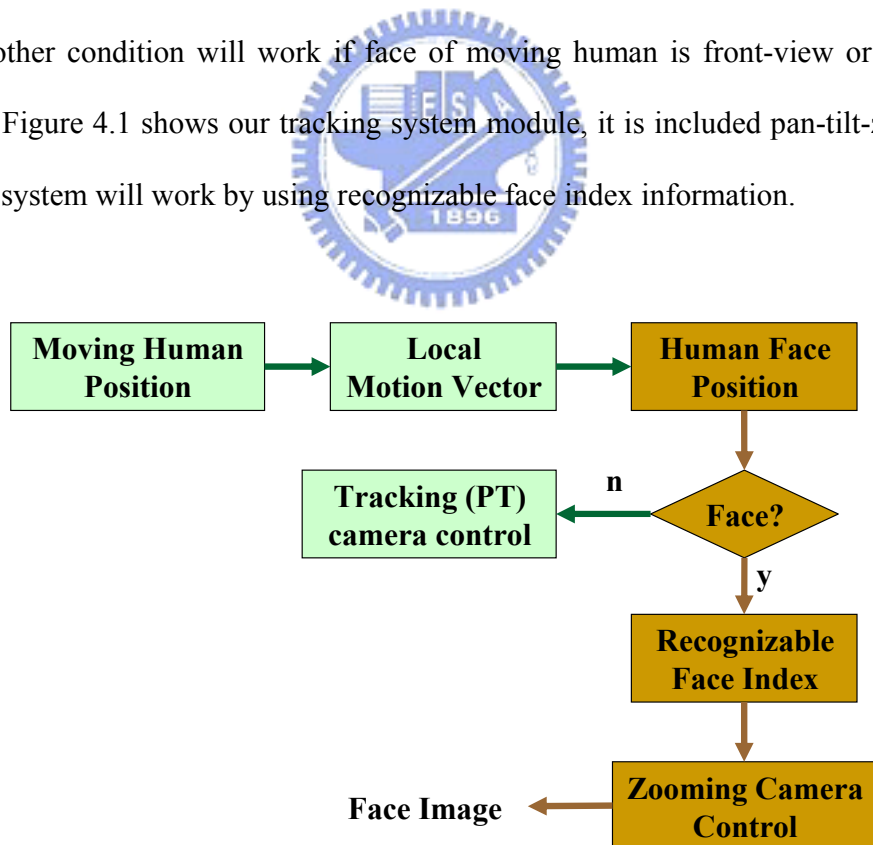


Figure 4.1 Tracking system

4.1 Active Camera

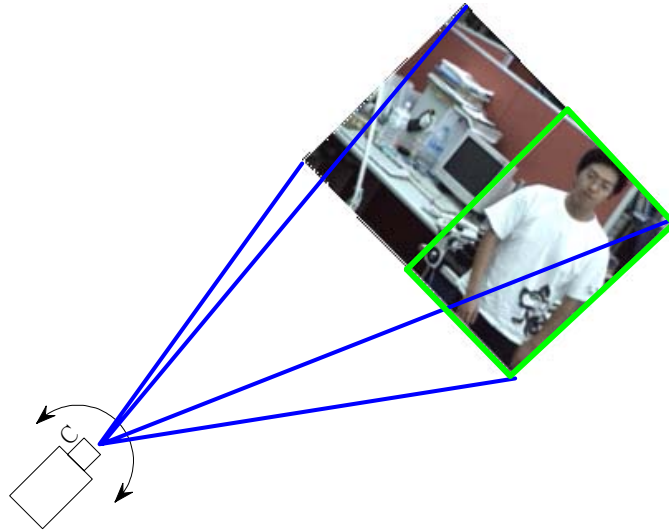


Figure 4.2 FOV camera

FOV camera is shown in Fig. 4.2. Image in this figure is captured after human detection part and the human region is entire the bounding box. In this condition, position of human is not in the center of FOV or image, so camera will drive pan-tilt by θ angle therefore human position in FOV center. In FOV camera, tilt upward is negative and pan right is positive.

The Camera drives pan-tilt by using angle values but human and face position in an image by pixel. We convert a pixel value to an angle by multiply it with a scale value which is called 'step size'. The step size is divided into x-axis direction and y-axis direction, if center of moving object close to center of image, step size will decrease, otherwise if far from center of image, step size will be increased.

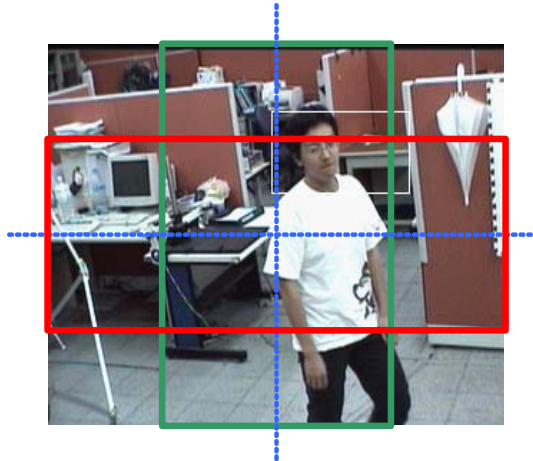


Figure 4.3 Step-size regions

Based on these rules we divide an image into two regions (x-axis and y-axis) as shown in Fig. 4.3. Every region is represented by a step-size value and the region inside bounding box has step-size smaller than regions outside bounding box.

4.2 Tracking and Zooming Process

Tracking and zooming are working together to follow a moving human. Figure 4.4 shows the tracking and zooming flowchart. It is started by detect location of moving object, if there is a human then we apply face detection to this image, otherwise the system will return and capture new images until we obtain moving human. If face obtained from face detection system has index smaller than a threshold value and its position in zooming region as shown in Fig. 4.5, the system is worked automatically to zoom-in face region therefore the result is contained a face with larger size and more clearly, otherwise if face index larger than other threshold value the system will do zoom-out.

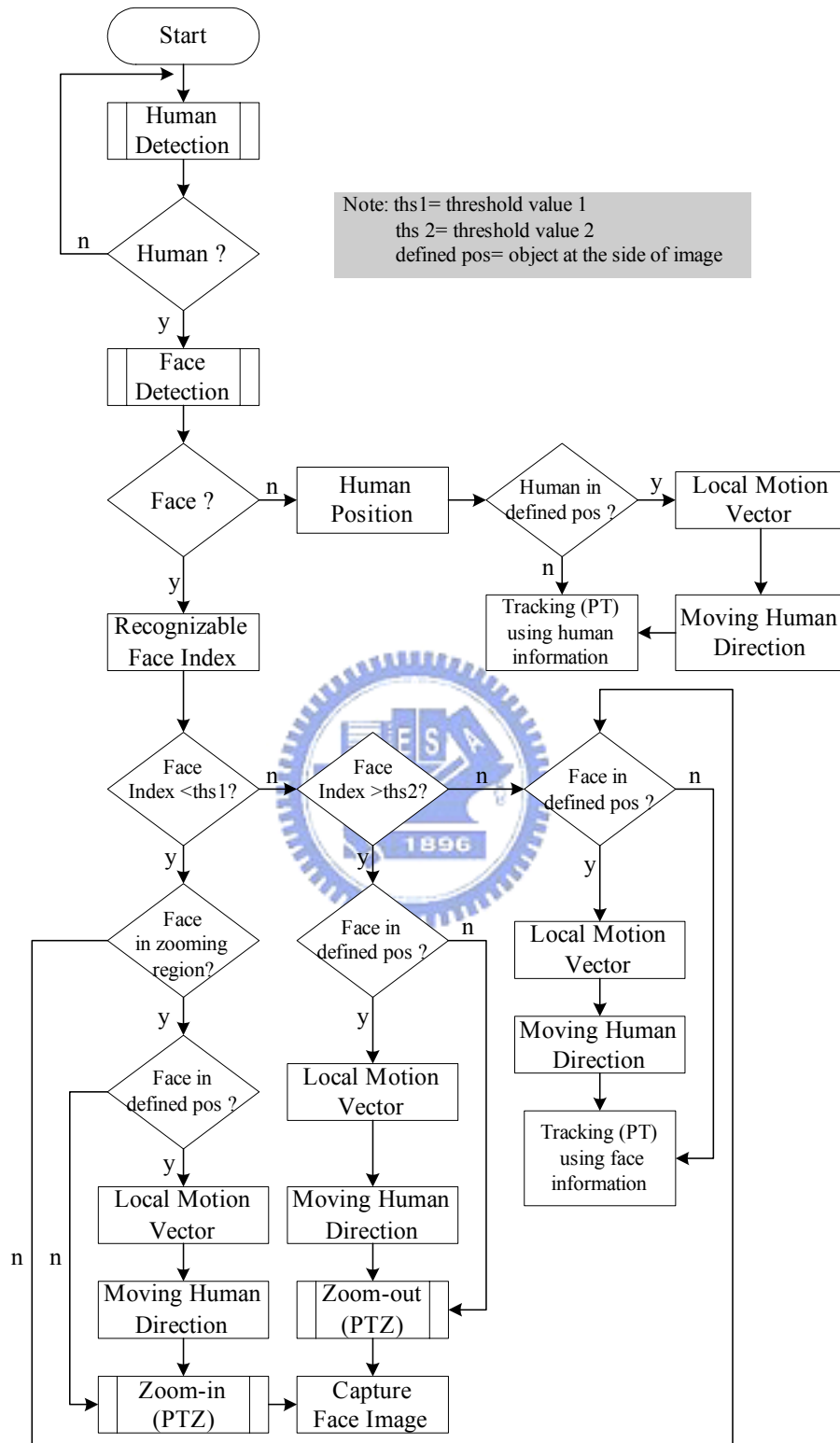


Figure 4.4 Tracking and zooming process

We also apply local motion vector to find the direction of moving object then uses the direction information to drive the pan (left-right direction) of camera.

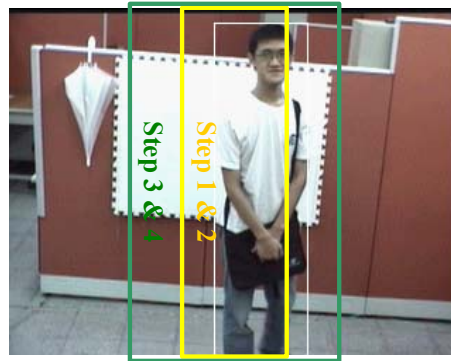


Figure 4.5 Zooming region

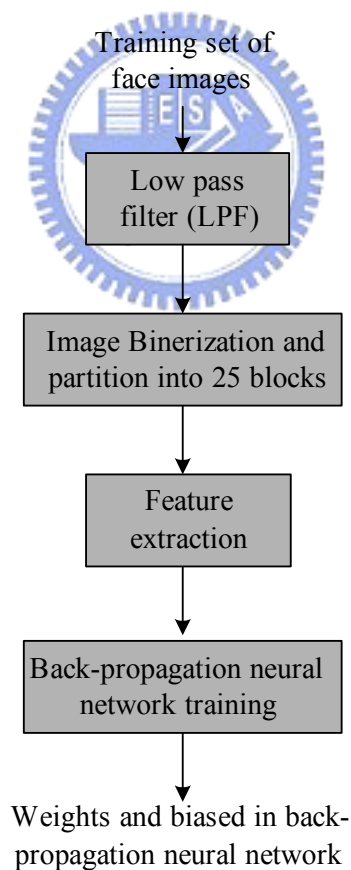


Figure 4.6 Training of recognizable face index

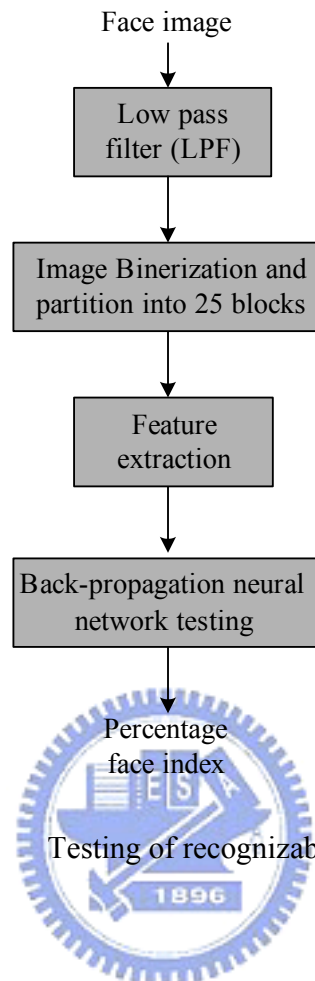


Figure 4.7 Testing of recognizable face index

4.2.1 Recognizable Face Index

Recognizable face index is used to give an index that face image can be recognized. The index is a percentage value i.e., high percentage value indicates the face is very clear and can be recognized. When we work in real-time face detection, our system still can detect a face although its size is small, i.e. 15x15 pixels, but a face image with this size is very difficult to recognize as face or non face because some details or features face is not available or too small to extract. In this case face index is needed as an indicator to the system therefore we can obtain a clear and high resolution face image.

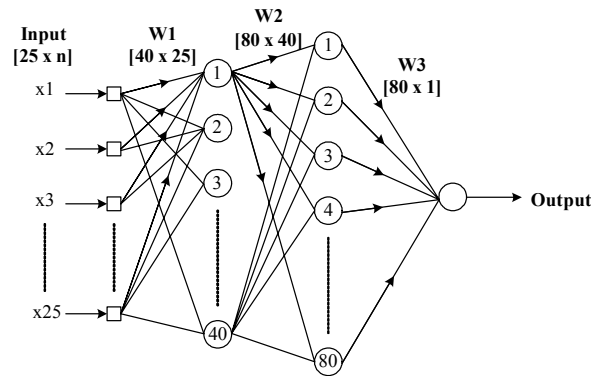


Figure 4.8 Recognizable face index training

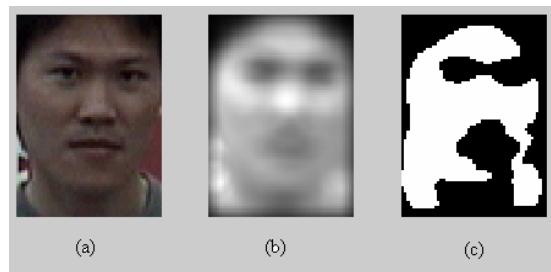


Figure 4.9(a) Original image, (b) After Filtering, (c) After filtering and binerization

Recognizable face index is performed by multilayer feedforward networks trained with back-propagation learning algorithm. It works offline but testing process works online. Figure 4.6 and Fig. 4.7 shows its training and testing process. The network consists of three layers. Input layer has 25 dimensions, two hidden layers, and output layer. Face images is captured by real-time face detection for several positions and classify it into three classes of percentage values. Front-view face is classified into index 90-100%, side-view 50-60% and 0-20%. Figure 4.9(a) shows a front-view face which has index 90%. The pre-processing steps are including low pass filter, image binerization, and find image features. Low pass filter is used to reduce small detail of face and find smooth image as shown in Fig. 4.9(b). By thresholded it with its median value we find

binary image as show in Fig. 4.9(c). This binary image is divided into 25 non-overlapping blocks. The feature of every block is calculated by averaging face region (pixels) in every block. These features will train as input neural network.

First hidden layer of neural network has forty neurons, second hidden layer has eighty neurons and output layer is a face index result. The activation function is sigmoid function that shows in Eq. (4.1)

$$\varphi(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (4.1)$$

Figure 4.10 shows some face images which used for recognizable face index training data. These face images already classified into three percentage values, first row images are face images with percentage value 20%, second row images are face images with percentage value 50-60%, and the last row images are face images with percentage value 90%.

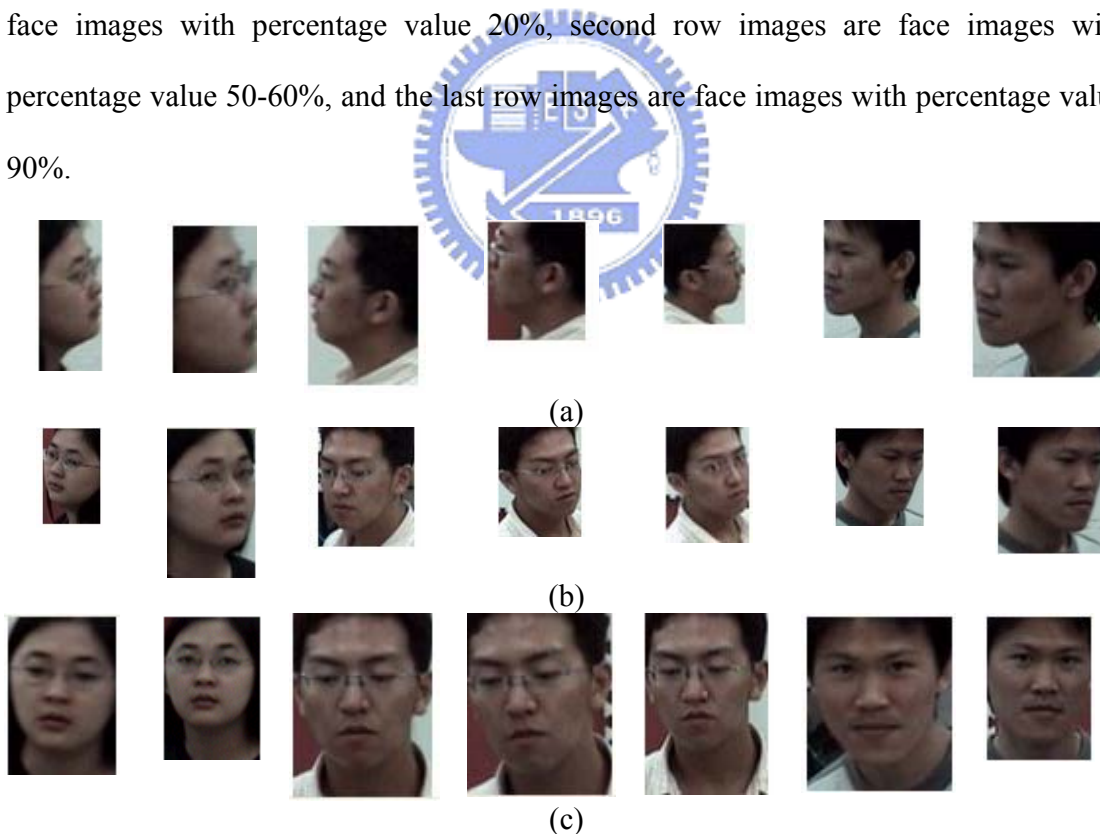


Figure 4.10 Training face images, (a) side-view faces with percentage index 20%, (b) side-view faces with percentage index 50-60%, (c) front-view faces with percentage index 90%

The testing process is worked online, its process similar with training part which consisted filtering and feature extraction. Apply weighting values and biases which obtained from training part in the testing image and the output is percentage of face index. Some face index result shows in Fig. 4.11. The index of first image is 60%, second image is 20% and the last is 91%. Small index indicates a face is difficult to recognize because the information is not enough. Meanwhile, face index 60% and 100% shows a face is clear and can be recognized.



Recognizable face index is applied in zooming process, as shown in Fig. 4.12. If face index smaller than threshold value 'ths1' camera will zoom-in face region, otherwise if face index larger than threshold value 'ths2' camera will zoom-out face region. This threshold values is changeable depends on users define. Zooming process has four steps which are step 1 until step 4 and it will zooming face region according to face index so on until we get clear and high resolution image.

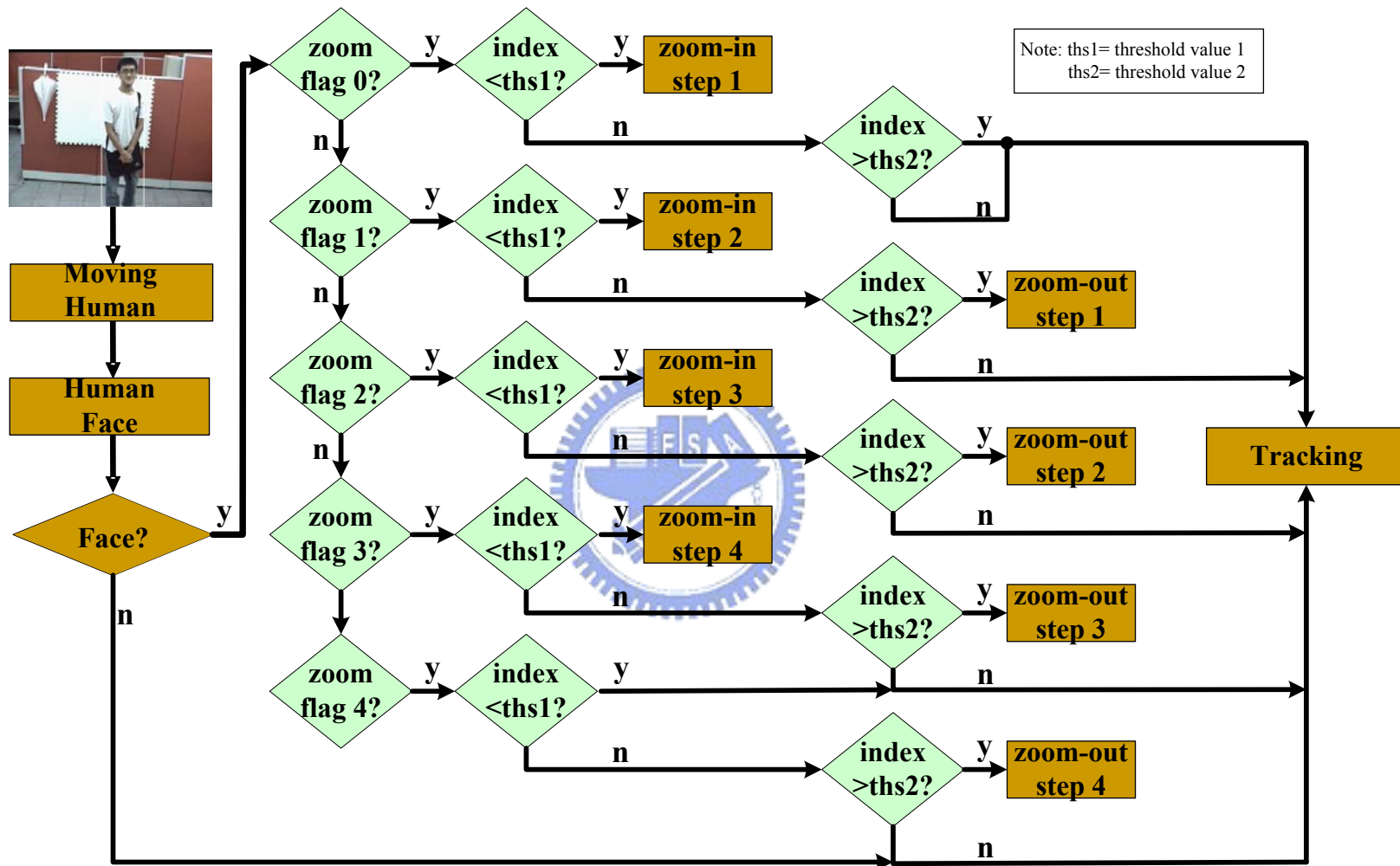


Figure 4.12 Zooming procedure

4.2.2 Local Motion Vector

Motion vectors are typically used to compress video by storing the changes of an image from one frame to the next. It can be used to find the movement of a pixel or block of pixels in a frame $f(n)$ to next frame $f(n+1)$ to extract moving object direction. We use the direction to predict position of moving object in the next frame. To compute motion vector in entire image it needs high computational cost and spends long time. To achieve real-time system, we combined motion vector and segmentation to reduce the computational cost. The motion vector is using binary difference image as reference position to find the motion vector.

We apply local motion vector to predict the direction of moving human. It works automatically when the object is in an area as shown in Fig. 4.13. Because in this region the moving human easily walks outside the FOV camera when the moving human changes its direction, so we predict the direction of moving human to make sure camera movement can follow the moving object by using local motion vector. We only work in x-axis direction, because in our experiment only a human which walks from left to right or right to left can cause it to walk outside the FOV, meanwhile if the human walks away or approaches, it is still in FOV.



Figure 4.13 Region of motion vector

Local motion vector is worked by scanning binary difference image pixel by pixel. Difference image in Fig. 4.14 contains '1' and '0' value, '1' is edge of moving object and '0' is background. If pixel p has $V = \{1\}$ and refers to Fig. 4.14, motion vector is applied in shaded pixel in position $(x,y) = (6,4)$ by using its position as center 3x3 image block of previous frame as shown in Fig. 4.15(a). By using SAD method in Eq. (4.2) we find its correlation in 7x7 image block of current frame. The center position of 7x7 image block same with center position of 3x3 image block shows in Fig. 4.15(b). Based on this rule, the movements of 3x3 image block only in three pixel of any direction.

$$SAD(i, j) = \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} |I_{previous}(x+k, y+l) - I_{current}(x+i+k, y+j+l)| \quad (4.2)$$

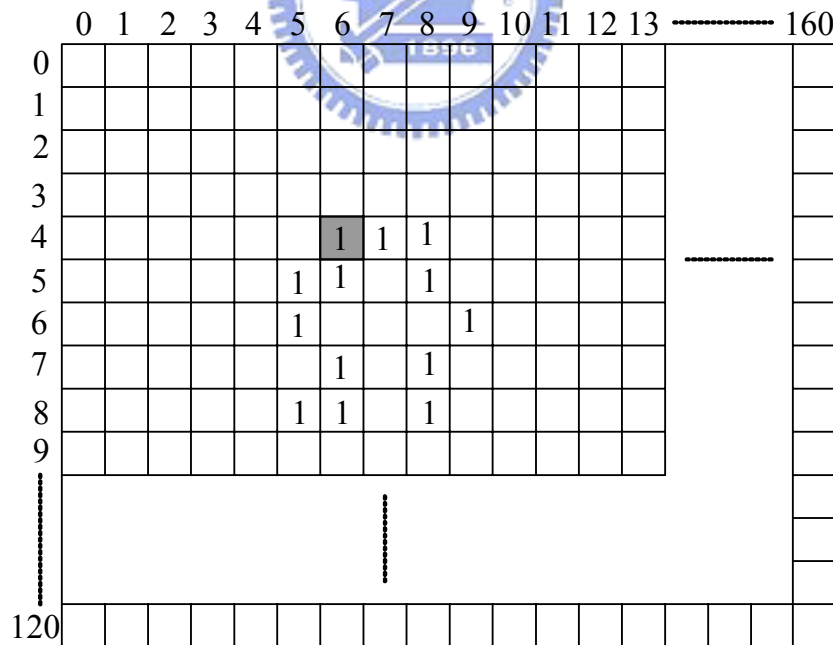


Figure 4.14 Difference between two frames

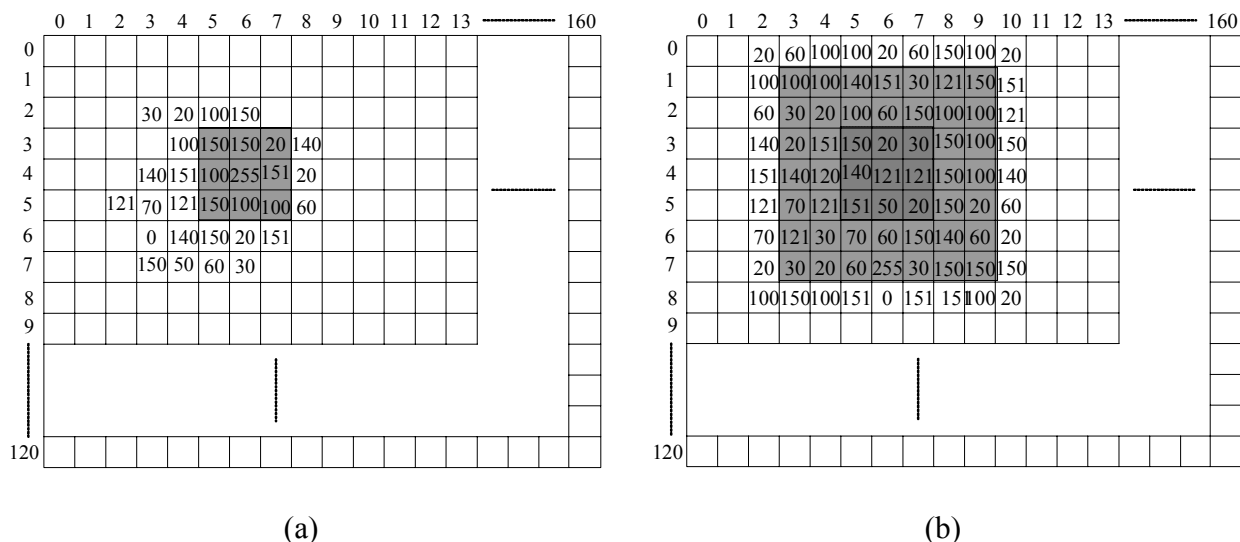


Figure 4.15 (a) Previous frame, (b) Current frame

Local motion vector is simulated by Matlab. Figure 4.16 shows two image frames, current and previous frame, and difference between these two frames. The local motion vector of it shows in Fig. 4.17. Directions of motion vector are left, right and fix. Most of directions in this image are from right to left hence we decided the object is moved from right to left side. The tracking system receives information about next position and left direction and so camera can move pan-tilt to new position and if index of human face larger than 50% than we capture that face.

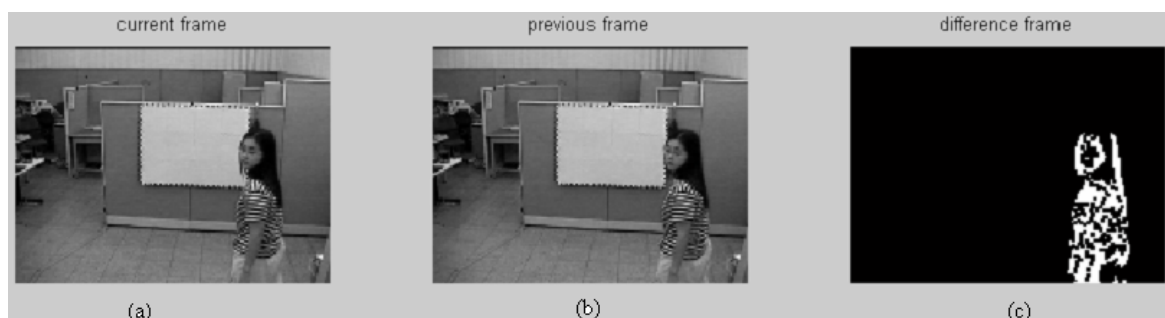


Figure 4.16 (a) Current frame, (b) Previous frame, (c) Difference between two frames

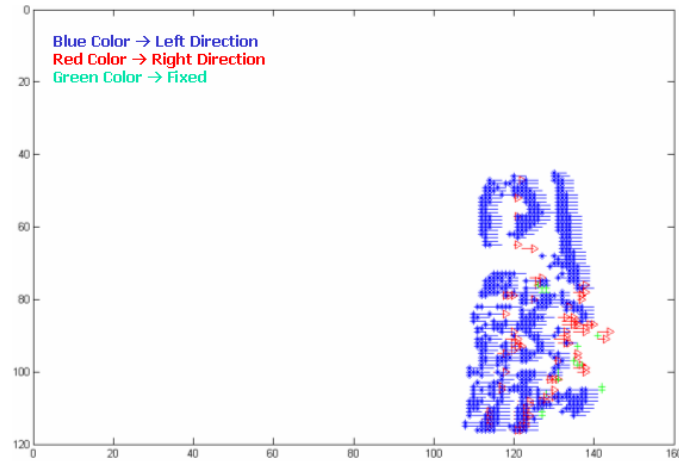


Figure 4.17 Local motion vector result

So, we can resume the tracking and zooming conditions as,

- Tracking us worked automatically when human in back-view condition and outside region zooming as shown in Fig. 4.5
- Zoom-in is worked automatically when the face index is smaller than 50% and moving object position in zooming region.
- Zoom-out is worked automatically when the face index is larger than 50% and moving object position in zooming region.

Chapter 5

Experimental Results

This chapter will show detection and tracking under condition without zooming camera control and with zooming camera control.

5.1 The Experimental System

The experimental system is using following components,

- The system uses SONY EVI-D100 active camera for capture image sequence. The sequence is composed of 320x240 color images acquired at a frame rate 20 frames per second.
- The system is developed in Borland C++ Builder 6 and has been tested on color image sequences acquired on indoor environments.
- The system has been implemented on an AMD Athlon XP 2000+ 1.6 GHz CPU computer under Microsoft XP and 512 Mb RAM.
- The active camera has two interfaces which are RS-232 and video-in. RS-232 interface is used to drive pan-tilt-zoom camera and video-in interface is an analog input that needs video grabber card so computer can read out the image data.

5.2 Environment Setup

The environment of our experimental locates in our laboratory. The complexity of the environment is enough to verify our system while tracking and detecting moving human. Figure 5.1 shows several images of the environment using focal length 3.1 mm which a normal condition without zoom-in/out operation. For zoom-in and zoom-out condition, Fig. 5.2 shows several images from zoom step one (focal length 4.65 mm) to zoom step four (focal length 12.4 mm).



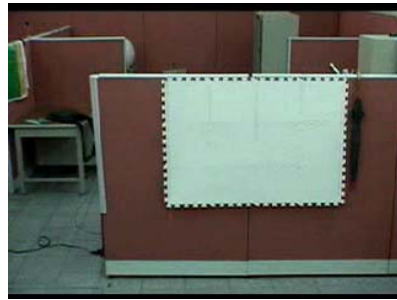
Figure 5.1 Indoor environments at reference focal length 3.1 mm



Reference focal length 3.1 mm



zoom-in step 1 = 4.65 mm



zoom-in step 2 = 6.2 mm



zoom-in step 3 = 9.3 mm



zoom-in step 4 = 12.4

Figure 5.2 Frames at several focal lengths

5.3 Experimental I

This experimental will detect and track without recognizable face index and zooming camera control. It is worked by drive pan-tilt camera to follow movement of moving human. The results are shown in figure below,



Figure 5.3 Tracking without zooming operation

This figure shows the system is successfully tracked the human by drive the pan-tilt but its face is too small to recognize and has low resolution. Some of human faces that have been extracted from this image sequence are shown in Fig. 5.4.



Figure 5.4 Human and Face detection sequence

The human has been detected within a few frames after entering the field of view camera. The bounding box is shown moving human region and the detected face is shows in the small upper-left window. The camera is tracking the human during back-view face or tracking the human face during front-view face.

The detected face images more difficult to recognize because size of image face too small and has low resolution so the face feature is more difficult to extract.

5.4 Experimental II

This experimental will detect and track moving human with recognizable face index and zooming camera control. The zooming process includes zoom-in and zoom-out which has four steps. The steps are depended on camera focal length. Zoom-in and zoom-out have same focal length which are the changeable focal length in zoom-in is opposite with zoom-out. The tracking and zooming experimental II results are shown in figure below,



Reference position (3.1 mm)



zoom-in step 1



zoom-in step 2



zoom-in step 3

Figure 5.5 Tracking and zooming image sequence (zoom-in)



zoom-in step 3



zoom-out step 3



zoom-out step 3



zoom-in step 3



zoom-in step 4



zoom-in step 4

Figure 5.6 Tracking and zooming image sequence (zoom-in and zoom-out)



Figure 5.7 Tracking and zooming step-4

Figure 5.5 shows zoom-in image sequence from reference focal length 3.1 mm until zoom in-step three. In this condition the camera is not increase the focal length to zoom-in step four because human face already clear and has face index larger than 50%. Figure 5.6 is continued from Fig. 5.5 which is the first image has same step with the last image in Fig. 5.5 (zoom-in step three condition). In this condition we obtain face index larger than 70% so the system automatically do zoom-out step three. When it is already worked we obtain face index smaller than 50% so system do zoom-in step three and zoom-in step four. In this condition the system can not increase the camera focal length because we limited focal length until step four (12.4 mm) therefore in this condition the system only do tracking and zoom-out depend on face index. Figure 5.7 is shown tracking its image sequence for several positions.



The face images which extracted from process above are shown in Figure 5.7. Face images are obtained from experimental II more clearly and have high resolution than experimental I.



Figure 5.8 Human tracking and face extraction (1)



Figure 5.9 Human tracking and face extraction (2)

The image frames in Fig. 5.8 and Fig. 5.9 are captured after down-sample by two so the face region in these images looks like smaller than face regions that have been extracted in up-side image. The up-side image is face image that extracted from original image.

5.5 Experimental III

This system is processed in indoor environment so the possible moving objects are human, animal, chair, etc. In our experimental, tracking and human detection is tested by using chair. We push a chair from one position to other position as shown in Fig. 5.10 and the system is not track it although it moves.



Figure 5.10 Moving object

Chapter 6

Conclusions and Future Works

6.1 Conclusions

The experimental results show that the system is capable of tracking moving humans and zooming operations can use the information in face indices, both of which will make the system work well on the faces with the size larger than 25x25. Face images obtained from this system are clearer and have a higher resolution than those from the system without zooming operations. The CPU performance approximately 10-35% depends on the resolution of moving objects which the average value is 20%. The local motion vector applied in this system can work well to predict the directions of moving objects although the driving pan-tilt becomes slow.

There are several contributions made out of this research,

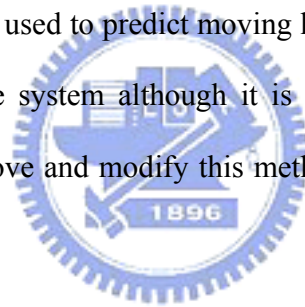
1. The system can distinguish human with other moving object based on deformable human model.
2. The system provides intelligent camera control (pan-tilt-zoom) technique to obtain the clear human faces and the face images in the higher resolution for some applications like face recognition.
3. We proposed recognizable face index to classify face based on the percentage clearance of those faces, such as front view face 90-100%, side view 30-50%, and side view 0-10%.

4. The system is based on tracking framework that involves a real-time system

6.2 Future Works

So far, our tracking system works in indoor environment with one moving target although it still can detect multiple humans and extract their faces but the tracking and zooming system only work on one moving human which is randomly chosen from the scenes. In order to solve that problem, we must deal with the system in a more complex situation which might have multiple targets and track moving humans that has information which the system want to know.

The local motion vector is used to predict moving human directions but it causes the high computational cost in the system although it is already combined with temporal differences. So we must improve and modify this method to decrease its computational cost.



References

- [1] M. J. Seow, R. Gottumukkal, D. Valaparla, and K. V. Asari, "A robust face recognition system for real time surveillance," in *Proc. Of the IEEE International Conference on Information Technology: Coding and Computing*, vol. 1, 2004, pp. 631-635.
- [2] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39-51, Jan. 1998.
- [3] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042-1052, 1993.
- [4] A. L. Yuille, D. S. Cohen, and P. W. Hallinan, "Feature extraction from faces using deformable templates," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1989, pp. 104-109.
- [5] R. C. Verma, C. Schmid, and K. Mikolajczyk, "Face detection and tracking in a video by propagating detection probabilities," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1215-1228, Oct. 2003.
- [6] H. P. Graf, T. Chen, E. Petajan, and E. Cosatto, "Locating faces and facial parts," in *Proc. of the International Workshop on Automatic Face and Gesture Recognition*, 1995, pp. 41-46.
- [7] E. Hjeltnes and B. K. Low, "Face detection: A survey," in *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236-274, Sept. 2001.

- [8] L. Wang, W. Hu, and T. Tan, "Face tracking using motion-guided dynamic template matching," in *the 5th Asian Conference on Computer Vision*, 2002.
- [9] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa, "A System for Video Surveillance and Monitoring," Robotics Institute, Carnegie Mellon University, Technical Report CMU-RI-TR-00-12, May 2000.
- [10] G. L. Foresti, C. Micheloni, L. Snidaro, and C. Marchiol, "Face detection for visual surveillance," in *Proc. of the 12th IEEE International Conference on Image Analysis and Processing*, Sept. 2003, pp. 115-120.
- [11] L. Zhao and C. E. Thorpe, "Stereo- and neural network-based pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, pp. 148-154, Sept. 2000.
- [12] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," in *Computer vision and image understanding*, vol. 80, no. 1, pp. 42-56, 2000.
- [13] M. Scheutz, J. McRaven, and Gy. Cserey. "Fast, reliable, adaptive, bimodal people tracking for indoor environments," in *Proc. of the IEEE International Conference on Intelligent Robots and Systems*, vol. 2, Oct. 2004, pp. 1347-1352.
- [14] J. Steffens, E. Elagin, and H. Neven, "PersonSpotter-fast and robust system for human detection, tracking and recognition," in *Proc. of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998, pp. 516-521.
- [15] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, Aug. 2000.

- [16] C. J. Li and S. J. Wang, "Detection and tracking of a single deformable object on an active surveillance camera," in *Proc. of the Computer Vision, Graphics, and Image Processing*, Kinmen, Taiwan, Aug. 2003.
- [17] S. Kawato and J. Ohya, "Automatic skin-color distribution extraction for face detection and tracking," in *Proc. of the Signal Processing*, vol. 2, 2000, pp. 1415 – 1418.
- [18] F. J. Huang and T. Chen, "Tracking of multiple faces for human-computer interfaces and virtual environments," *IEEE International Conference on Multimedia and Expo*, vol. 3, 2000, pp. 1563 – 1566.
- [19] E. Loutas, I. Pitas, and C. Nikou, "Probabilistic multiple face detection and tracking using entropy measures," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 14, no. 1, pp. 128-135, Jan. 2004.
- [20] C. J. Li and S. J. Wang, "Detection and Tracking of a Single Deformable Object on an Active Surveillance Camera," in *the 16th IPPR Conference on Computer Vision, Graphics and Image Processing*, 2003, pp. 96-103.
- [21] D. Comaniciu and V. Ramesh, "Robust detection and tracking of human faces with an active camera," in *Proc. of the 3rd IEEE International Workshop on Visual Surveillance*, July 2000, pp. 11-18.
- [22] J. Segen, "A camera-based system for tracking people in real-time," in *Proc. of the 13th International Conference on Pattern Recognition*, vol. 3, 1996, pp. 63-67.
- [23] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1241-1247, Nov.1999.

- [24] P. J. Burt, J. R. Bergen, R. Hingorani, R. Kolczynski, W. A. Lee, A. Leung, J. Lubin, and H. Shvayster, "Object tracking with a moving camera," in *Proc. of the IEEE Workshop on Visual Motion*, March 1989, pp. 2-12.
- [25] P. H. Batavia, D. E. Pomerleau, and C. E. Thorpe, "Overtaking vehicle detection using implicit optical flow," *IEEE Conference on Intelligent Transportation System*, Nov. 1997, pp. 729-734.
- [26] W. J. Gillner, "Motion based vehicle detection on motorways," in *Proc. of the Intelligent Vehicles '95 Symposium*, Sept. 1995, pp.483-487.
- [27] C. E. Smith, C. A. Richards, S. A. Brandt, and N. P. Papanikolopoulos, "Visual tracking for intelligent vehicle-highway systems," *IEEE Transactions on Vehicular Technology*, vol. 45, no. 4, pp. 744-759, Nov. 1996.
- [28] R. Polana and R. Nelson, "Detecting activities," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1993, pp. 2-7.
- [29] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking pedestrian recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, pp.155-163, Sept. 2000.
- [30] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34-58, Jan. 2002.
- [31] T. Sakai, M. Nagano, and T. Kanade, "Computer analysis and classification of photographs of human faces," in *Proc. of the First USA-Japan Computer Conference*, 1972, pp. 2-7.

- [32] Z. F. Liu, Z. S. You, A. K. Jain, and Y. Q. Wang, "Face detection and facial feature extraction in color image," in *Proc. of the 5th International Conference on Computational Intelligence and Multimedia Applications*, Sept. 2003, pp. 126-130.
- [33] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706, 2002.
- [34] S. L. Phung, A. Bouzerdoum, and D. Chai, "A novel skin color model in YCbCr color space and its application to human face detection," in *Proc. of the IEEE International Conference on Image Processing*, vol. 1, Sept. 2002, pp. I-289 - I-292.
- [35] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Transaction on Multimedia*, vol. 1, no. 3, pp. 264-277, Sept. 1999.
- [36] R. Kjeldsen and J. Kender, "Finding skin in color images," in *Proc. of the 2nd International Conference on Automatic Face and Gesture Recognition*, Oct. 1996, pp. 312-317.
- [37] U. Neumann, I. Cohen, S. You, D. Fidaleo, and K. Seo, "Real-time face detection from one camera," in *IMSC and computer science*, USC.
- [38] C. Garcia and M. Delakis, "Convolutional face finder: a neural architecture for fast and robust face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1408-1423, 2004.
- [39] J. Yang and A. Waibel, "A real-time face tracker," in *Proc. of the 3rd IEEE Workshop on Applications of Computer Vision*, Dec. 1996, pp. 142-147.

- [40] K. Schwerdt and J. L. Crowley, "Robust face tracking using color," in *Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 90-95, March 2000.
- [41] C. T. Lin and C. S. George Lee, *Neural fuzzy systems: A neuro-fuzzy synergism to intelligent systems*, Prentice Hall, 1999.
- [42] R. Fitzgerald, "Divergence of the Kalman filter," *IEEE Transaction on Automatic Control*, vol. 16, no. 6, pp. 736-747, Dec. 1971.
- [43] K. V. Ramachandra, *Kalman Filtering Techniques for Radar Tracking*, Marcel Dekker, New York, 2000.
- [44] M. S. Grewal, A. P. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*, Second Edition, John Wiley & Sons., New York, 2001.
- [45] B. Turdoff and D. Murray, *Reactive zoom control while tracking using an affine camera*, Oxford University.
- [46] E. Hayman, *The use of zoom within active vision*, Robotic research group department of engineering science, Oxford University, 2000.
- [47] http://science.nsta.org/enewsletter/2004-10/ss0401_30.pdf
- [48] R. C. Gonzales and R. C. Woods, *Digital image processing*, Prentice Hall, 2002.
- [49] S. Haykin, *Neural Networks: A Comprehensive foundation*, Prentice Hall, 1999.
- [50] <http://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm>
- [51] M. J. Black, *Robust incremental optical flow*, dissertation YALEU/CSD/RR #923, YALE University, 1992.

[52] H. Shah and D. Morrell, "An adaptive zoom algorithm for tracking targets using pan-tilt-zoom cameras," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 2004, pp. 721-724.

