

國立交通大學

電機與控制工程學系

博士論文

以二維影像與漸進式相似度外觀圖解法
為基礎之穩健三維物體辨識



**Robust 3D Object Recognition using 2D Views
via an Incremental Similarity-Based
Aspect-Graph Approach**

研究生：蘇宗敏

指導教授：胡竹生 教授

中華民國九十六年九月

以二維影像與漸進式相似度外觀圖解法為基礎
之穩健三維物體辨識

**Robust 3D Object Recognition using 2D Views via an
Incremental Similarity-Based Aspect-Graph
Approach**

研究生：蘇宗敏

Student : Tzung-Min Su

指導教授：胡竹生

Advisor : Jwu-Sheng Hu

國立交通大學

電機與控制工程學系



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

Doctor of Philosophy

in

Electrical and Control Engineering

September 2006

Hsinchu, Taiwan, Republic of China

中華民國九十六年九月

以二維影像與漸進式相似度外觀圖解法 為基礎之穩健三維物體辨識

研究生：蘇宗敏

指導教授：胡竹生 博士

國立交通大學電機與控制工程學系(研究所)博士班

摘要

本論文提出了一套使用二維影像的穩健三維物體辨識架構。在此架構中包含了兩個主要部份，第一部份是前處理的部份，用來抽取出二維影像中的前景物體，以作為後續的學習與辨識之用。第二部份是一套漸進式資料庫建立方法，利用從不同角度所拍攝到的三維物體之二維影像來建構出該三維物體資料庫，並且能夠利用新拍攝到的二維影像來更新已建構好之三維物體資料庫。

在前處理的部份，我們提出了一套包含強光與陰影濾除的背景濾除架構(BSHSR)，使得前景物體在在光影變化與動態背景的影響下，依然能夠精確的被萃取出來。BSHSR中包含了三個模型，分別是以色彩為基礎的機率背景模型(CBM)、以CBM為基礎的梯度機率背景模型(GBM)，以及一個圓錐形的光影模型(CSIM)。CBM是利用高斯混合模型(GMM)針對每個像素的像素值作統計所建構出來的模型。而根據CBM，又可以建構出短期背景模型(STCBM)與長期背景模型(LTCBM)，接著再利用STCBM與LTCBM建構出GBM。而為了區別前景、強光與陰影的不同，本研究中提出了一建構在RGB色彩空間中且具有動態錐形邊界的CSIM。在漸進式資料庫建立方法的部份，我們提出了一套以相似度外觀圖解法為基礎的學習架構(ISAG)。利用相似度外觀圖解法，每個三維物體在資料庫中均可用一組外觀(aspect)來表示，而每一個外觀則包含了數目不一的二維影像，並且用一個特徵面(characteristic view)來代表。本研究所提出的漸進式資料庫建立方法，目的在於提高屬於同一外觀的二維影像彼此之間的相似度，並且降低各個特徵面彼此之間的相似度。此外，為了模擬人類認知物體的能力，我們採用隨機取樣之角度所拍攝的三維物體之二維影像來做為訓練影像，隨著所收集到的二維影像數目增加，該三維物體的資料庫也會隨之更新。最終，本論文先以實際複雜環境中所拍攝的數段影片之實驗結果來說明所提出的BSHSR之可行性，接著將BSHSR應用於三維物體辨識架構中，以抽取出二維影像中之前景物體。而為了驗證所提出三維物體辨識架構之優越性，我們利用ISAG搭配物體的形狀與色彩特徵，將之應用於三種不同的三維物體之問題，分別是剛體辨識、人形姿態辨識與場景辨識，並根據辨識率結果來說明所提出的三維物體辨識架構之可行性。

Robust 3D Object Recognition using 2D Views via an Incremental Similarity-Based Aspect-Graph Approach

Graduate Student: Tzung-Min Su

Advisor: Dr. Jwu-Sheng Hu

Department of Electrical and Control Engineering
National Chiao-Tung University

Abstract

This work presents a framework for robust recognizing 3D objects from 2D views. The proposed framework comprises of two stages: the pre-processing stage and the incremental database construction stage. In the pre-processing stage, foreground objects is extracted from 2D views and applied for building 3D database and recognizing. In the incremental database construction stage, a 3D object database is built and updated using 2D views randomly sampled from a viewing sphere.

A background subtraction scheme involving highlight and shadow removal (BSHSR) is proposed as the pre-processing stage of the framework. Foreground regions can be precisely extracted from 2D views using the BSHSR despite illumination variations and dynamic background. The BSHSR comprises three models, called the color-based probabilistic background model (CBM), the gradient-based version of the color-based probabilistic background model (GBM) and a cone-shape illumination model (CSIM). The Gaussian mixture model (GMM) is applied to construct the CBM using pixel statistics. Based on the CBM, the short-term color-based background model (STCBM) and the long-term color-based background model (LTCBM) can be extracted and applied to build the GBM. Furthermore, a new

dynamic cone-shape boundary in the RGB color space, called the CSIM, is proposed to distinguish pixels among shadow, highlight and foreground.

An incremental database construction method based on similarity-based aspect-graph (ISAG) is proposed for building the 3D object database using 2D views. Similarity-based aspect-graph, which contains a set of aspects and characteristic views for these aspects, is employed to represent the database of 3D objects. An incremental database construction method that maximizes the similarity of views in the same aspect and minimizes the similarity of prototypes is proposed as the core of the framework. To imitate the ability of human cognition, 2D views randomly sampled from a viewing sphere are applied for building and updating a 3D object database. The effectiveness of the BSHSR is demonstrated via experiments with several video clips collected in a complex indoor environment. The BSHSR is applied in the proposed framework to extract foreground object from 2D views. The proposed framework is evaluated on various 3D object recognition problems, including 3D rigid recognition, human posture recognition, and scene recognition. Shape and color features are employed in different applications with the proposed framework to show the efficiency of the proposed method.

致謝

對於本論文的完成，首先要感謝的是我的指導教授 胡竹生 教授，感謝老師在我的碩士班與博士班這七年來，給予我相當多珍貴的意見與指導，往往能夠讓我茅塞頓開，並且能從許多層面去思考問題。這段時間也從老師身上學到了許多研究的態度與寶貴的知識，讓我能夠用更成熟與積極的態度來面對自己的研究。而除了課業之外，老師也給予我關於感情、婚姻與生涯規劃方面的建議，並且在我遇到低潮的時刻給予我鼓勵，真的很感謝老師。此外，也很謝謝老師給予我機會去參與許多比賽與研究計畫，讓我能夠有機會獲得更多的磨鍊，也因而有機會在 2002 年 9 月份跟老師一起到英國參加 CCA 2002 研討會，並且參觀了牛津大學、劍橋大學、IEE 學會、大英博物館、白金漢宮、Glasgow 與愛丁堡等等地點，都是讓我很難忘的回憶。在此，誠摯的致上我最真摯的謝意，謝謝老師。

另外，首先要感謝一起執行數位化居家照護系統三年計畫的學弟妹們，感謝幽默又搞笑的士奇，你是我認識的朋友裡面，唯一一位可以跟打掃工五館的特殊小孩們玩在一起的人，可見你一定很有愛心以及具備特殊的吸引力。感謝住在信義區的小開群棋，你敦厚老實的個性搭配害羞的笑容，總有一天可以找到你的真命天女。感謝認真又細心的佩靜，妳專情與顧家的個性，真的是很值得讚許。感謝搞笑又無厘頭的恆嘉，沒想到你是減肥界的達人，我是從你口中才知道珍珠奶茶的熱量有多驚人。感謝為愛走天涯的弘齡，相信你因為美國之行而成長了許多。

接著要感謝 88 級的學長們，感謝傳承實驗室管理員給我的 DSP 達人鴻志學長，以及幽默的凱學長，感謝你們兩位耐心的教導我 DSP 相關知識，感謝親切的瓊宏學長，謝謝你常常跟我分享影像相關知識以及鼓勵我。感謝 RTOS 達人小陶子學長，實作實力超強的你，讓我非常的敬佩，另外，想不到你對於駭客破解技術頗有研究。感謝話不多但是實力很強的邦正學長，以及感謝溫文儒雅的俊德學長，每次跟你聊天，心情都相當棒。

感謝 89 級同屆進入實驗室的學長與同學，感謝執行力超強的立偉學長，你

對機器人總有說不完的理想，時間對於你來說也總是不夠用，從你身上我學到了許多做事情應有的態度，也很懷念那時候跟你參加 ICM/HIMA 2005 研討會的時候，一起到士林夜市逛街聊天。再來要感謝跟我認識 14 年，高中同校、大學同系、研究所與博士班都同一位指導教授、博士班論文口試同一天，甚至連國防役服役公司都是頗有淵源的兩家公司的維瀚，真的是很有緣分啊！革命情感也不用多說了！一起修課、參加比賽、寫 DSP 程式、寫論文，真的很多很多的回憶啊！接著要感謝桌球很強、又很會吃炸冰淇淋的价呈，記得當年到台南比賽以及到嘉義參加 AMTE 2002 研討會的時候，一起住在飯店準備比賽資料與投影片，在校時間一起修課、選讀博士班、打羽球、玩三國志、CS...，留下許多許多美好的回憶~ 而講到 CS，就讓我想到欣慈，每次玩 CS 的時候，妳的反應總是讓大家記憶深刻啊~ 此外，也相當懷念當年跟妳一起在影像領域研究的日子，跟妳討論之後，總是能有許多不同的體驗。最後，感謝已經改名為峻葦的家銘，你既是我大學同班同學也是我系學會的夥伴，在研究所的那兩年中，除了見識到你小畫家的功力之外，也感染到你在研究上的熱情。

感謝後來常常跑去 DSP 實驗室的德琪，妳最後的論文成果果然是相當豐碩啊~ 感謝每次來參加聚會都讓人感覺越來越瘦的嘉芳，也感謝上次到妳公司面試時，妳在公司陪我聊了那麼久。感謝超有想法的阿鎧，很懷念那時候跟你一起跑環校，可惜你到最後切了西瓜，然後一直被記到現在，為了怕知道這件事情的學弟都畢業了，請原諒我繼去年价呈提了一次之後，今年又再度提了一次。感謝已經改名為昊群的青衛，記得當年跟你與价呈一起討論財經話題，雖然討論的結果跟未來的發展都不盡符合，但還是相當有趣的一段回憶。感謝很像王力宏的 Alan(葉威廷)，帥氣的你事業愛情兩得意，每每都能從你那邊得到許多業界的訊息。感謝重視養生的倉億，我從你那邊學到了許多中藥的知識，還得到了一張你說經過氣功師父加持過的卡片。

感謝思考獨特但總是讓人誤會有在吸毒的春成，感謝作風海派但心思細膩的 Angel(蔡銘謙)，感謝專情又積極上進的家瑋，感謝認真負責的億如，感謝喜歡

說冷笑話然後也很好笑的順智，感謝會煮紅豆湯、會做小卡片的康康，感謝吃素的俊德跟我分享交友經驗談，感謝很 man 的佳興，跟你一起到上海參加 CASE 2006 的那一個禮拜，讓我留下許多美好的回憶，此外，每次吃飯時間跟你討論財經與感情話題，也讓我有許多收穫，感謝減肥界達人鏗元，每次打開實驗室冰箱，總是可以感受到妳的存在，感謝被稱為實驗室一姊的岑思，想不到妳那麼在意被別人踩到鞋子，也希望下次能看到妳攜伴參加聚會喔！感謝很有衝勁的晏榮，感謝具有貴賓狗命的藍蕙，妳總是帶給我們最棒的運氣，感謝也很 man 的耀賢，還記得那天大家一起玩牌的時候，你的演技真的是頗不賴喔！感謝很憨厚又很疼女友的榮煌，感謝信主而不能說謊的永融，你認真的研究態度感染的身邊的所有人，感謝常常熬夜爆肝但肝指數都很正常的 Alphas，感謝具備多種才藝的凱祥，另外，也要感謝還在實驗室的學弟妹們：感謝上次幫我搬家的俊宇、感情快要開學的啟揚、很有研究熱忱的阿吉、很耐操的可以做實驗 12 個小時以上的 PaPa、聽說唱歌很好聽的治宏、愛喝葡萄柚青茶的瓊文、畢業生代表鎮宇、白白淨淨的明堂、有正妹同學的源松以及常常打扮很時髦的育綸。

另外，也要感謝我的室友益生，從你那邊我學到了許多關於論文投稿的知識，跟你一起討論佛教話題也讓我獲益匪淺，感謝常常借我車的室友仁乾，你對於法輪功的熱忱，讓我非常的敬佩！感謝還沒畢業就已經在外面工作很久的室友盟淳，感謝你提供我許多在業界的經驗談，讓我找國防役的時候能夠更有信心，也要感謝我早期的室友豐洲跟與豐洲的女友文真，感謝你們在我低潮的時候給我鼓勵與支持，也要感謝我的高中同學們，每年一次的同學會都讓我充滿了期待。最後，衷心的感謝我的家人，感謝我的爸爸與媽媽，讓我在唸書的這段時間，一直都沒有後顧之憂，可以專心唸書，並且持續的給予我關心與溫暖。感謝我的兩位妹妹，每當我回家，都會給予我最溫暖的關懷。也要感謝認識了 17 年，交往了 9 年的女友雅婷，感謝妳一直扮演我生命中的重要角色，妳持續的關心與鼓勵，一直是我在求學路上重要的力量。謝謝在我生命中的這些貴人，因為有您們的關心，讓我能夠用正面積極的態度來面對人生的旅途，有您們真好！

Contents

Chapter 1	Introduction	1
1.1	Overview of 3D Object Recognition	1
1.1.1	3D Object Recognition	1
1.1.2	Human Posture Recognition	2
1.1.3	Scene Recognition	3
1.2	Overview of Background Subtraction	4
1.3	Outline of Proposed System.....	6
1.3.1	Background Subtraction.....	6
1.3.2	3D Object Recognition	7
1.4	Contribution of this Dissertation.....	8
1.5	Dissertation Organization	9
Chapter 2	Background Subtraction.....	10
2.1	Introduction.....	10
2.2	System Architecture	11
2.3	Background Modeling	12
2.3.1	Color-Based Background Modeling	12
2.3.2	Model Maintenance of the LTCBM and STCBM.....	15
2.3.3	Gradient-Based Background Modeling	20
2.4	Background Subtraction with Shadow Removal	23
2.4.1	Shadow and Highlight Removal	23
2.4.2	Background Subtraction.....	27
Chapter 3	Incremental Similarity-Based Aspect-Graph 3D Object Recognition ..	29
3.1	Introduction.....	29
3.2	System Architecture	32
3.3	Object Representation.....	34
3.3.1	Shape Features	34
3.3.2	Color Features.....	36
3.3.3	Similarity Functions.....	38
3.3.4	Similarity Measures	38
3.4	Flexible 3D Object Recognition Framework.....	39
3.4.1	Generation of Aspects and Characteristic Views	40
3.4.2	Object Recognition using 2D Characteristic Views.....	43
3.4.3	Applications	44
Chapter 4	Experimental Results.....	47
4.1	BSHSR.....	47

4.1.1	Local Illumination Changes	48
4.1.2	Global Illumination Changes	53
4.1.3	Foreground Detection	55
4.1.4	Dynamic Background	55
4.1.5	Short-Term Color-based Background Model (STCBM)	58
4.2	3D Object Recognition	59
4.2.1	Rigid Object Recognition	63
4.2.2	Human Posture Recognition	67
4.2.3	Scene Recognition	69
Chapter 5	Conclusions and Future Researches	74
5.1	Conclusions	74
5.2	Future Researches	77
References	79



Index

Assistant 3D object databases (AOD).....	32
Background subtraction scheme involving highlight and shadow removal (BSHSR) .iv	
Candidate color-based background model (CCBM).....	16
Color-based probabilistic background model (CBM).....	iv
Cone-shape illumination model (CSIM).....	iv
Elected color-based background model (ECBM).....	15
Expectation maximization (EM).....	13
Fourier descriptor (FD).....	34
Gaussian mixture model (GMM).....	iv
Gradient-based version of the color-based background model (GBM).....	iv
Gradient Vector Flow Snake (GVF).....	34
Incremental database construction method based on similarity-based aspect-graph (ISAG).....	v
Long-term color-based background model (LTCBM).....	iv
Main 3D object database (MOD).....	33
Maximum likelihood (ML).....	13
Multicolored Region Descriptor (M-CORN).....	1
Point-to-point length (PPL).....	34
Principal component analysis (PCA).....	3
Short-term color-based background model (STCBM).....	iv

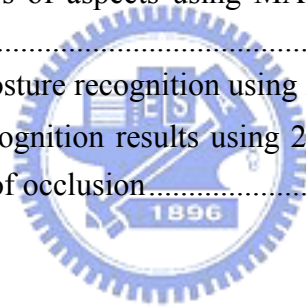
List of Figures

Figure 1-1 Block diagram of proposed 3D object recognition system.	6
Figure 1-2 Block diagram of foreground detection.	7
Figure 1-3 Block diagram of 3D object recognition.	7
Figure 2-1 The block diagram of the BSHSR.	11
Figure 2-2 Block diagram showing the process of building the initial LTCBM, ECBM and CCBM.	16
Figure 2-3 Block diagram showing the process to calculate H_{CG}	20
Figure 2-4 The proposed 3D cone model in the RGB color space.	24
Figure 2-5 2D projection of the 3D cone model from RGB space onto the RG space.	25
Figure 3-1 The system architecture of the proposed framework. A MOD comprises of total AODs.	33
Figure 3-2 The database building procedure, where T_0 is the number of objects in the database and T_1 is the number of sampled views required to build the aspect-graph representation of an object.	33
Figure 3-3 The inner structure of an AOD.	33
Figure 3-4 5D feature vector construction.	36
Figure 3-5 The procedure of the ISAG.	42
Figure 4-1 The results of illumination changes with a yellow desk light, the number below the picture is the index of frame.	50
Figure 4-2 The results of illumination changes with white desk light, the number below the picture is the index of frame.	52
Figure 4-3 The results of global illumination changes with fluorescent lamps, the number below the picture is the index of frame.	54
Figure 4-4 The results of foreground detection.	56
Figure 4-5 The results of background subtraction about dynamic background.	57
Figure 4-6 The results of the advantage of the STCBM, where the red color means the shadow, the green color means the highlight and the blue color means the foreground.	59
Figure 4-7 The first database containing 12 3D rigid objects.	60
Figure 4-8 The second image database containing eight 3D human postures.	60
Figure 4-9 The third image database containing 11 scenes.	61

Figure 4-10 The system architecture of the proposed framework applied on the first experiment (3D rigid object recognition).	64
Figure 4-11 Recognition rates of coarse and fine databases ($D_{18}, D_{36}, D_{54}, D_{72}, D_{90}$ and D_{108}), calculated using 200 results.....	66
Figure 4-12 Standard deviations of recognition rates using coarse to fine databases ($D_{18}, D_{36}, D_{54}, D_{72}, D_{90}$ and D_{108}), calculated using 200 results.....	67
Figure 4-13 The system architecture of the proposed framework applied on the second experiment (human posture recognition).	68
Figure 4-14 The indoor environment from which scenes in the third database are obtained.....	70
Figure 4-15 The sample training image, blob model and conceptual description of each scene captured in the indoor environment (Fig. 4-14).....	70
Figure 4-16 The system architecture of the proposed framework applied on the second experiment (human posture recognition).	71
Figure 4-17 The 11 characteristic views at the 6 th position in the indoor environment.	71
Figure 4-18 The test images captured from the 6 th position in the indoor environment.	71
Figure 5-1 The aspect-graph representation of the first human posture listed in Fig. 4-8 via MAG only.	76
Figure 5-2 3D object recognition system with a combination of a feature predictor and the proposed method illustrated in Fig. 1-3.	77
Figure 5-3 3D object recognition system with an efficient searching algorithm.....	78

List of Tables

Table 2-1 An example to calculate CG histogram	19
Table 4-1 The robustness test between the proposed method and that proposed by Hoprasert [60] via local illumination changes with a yellow desk light.....	49
Table 4-2 The robustness test between the proposed method and that proposed by Hoprasert [60] via local illumination changes with a white desk light.....	51
Table 4-3 The comparison between the proposed method and that proposed by Hoprasert [60] via global illumination changes with fluorescent lamps	53
Table 4-4 The comparison between the proposed method and that proposed by Hoprasert [60] via foreground detection.....	55
Table 4-5 The threshold values for the ISAG	62
Table 4-6 The result of rigid object recognition using 2D views via MAG and PPL..	64
Table 4-7 Results for numbers of aspects using MAG and PPL after updating with additional training views.....	65
Table 4-8 Results of human posture recognition using 2D views via MAG and θ_z ..	68
Table 4-9 Human posture recognition results using 2-D views via BM with position variations and different level of occlusion.....	73



Chapter 1

Introduction

1.1 Overview of 3D Object Recognition

1.1.1 3D Object Recognition



Object recognition is an important topic in computer vision where various approaches have been developed [1-5]. However, numerous technical issues require further investigation, especially for 3D object recognition. Variations in viewing direction and angle [1, 6-7], illumination changes [8-9], and scene clutter and occlusion [10-11] are the main challenges for object recognition. In recent years, many researches were presented for solving these issues. For example, a generic object class detection system [12] that combines the Implicit Shape Model and multi-view specific object recognition is presented to detect object instances from arbitrary viewpoints. A new framework [13] that combines a visual-cortex-like hierarchical structure and an increasingly complex and invariant feature was proposed for robust object recognition. Furthermore, a new object representation, Multicolored

Region Descriptor (M-CORN) [14], was proposed to describe the color and local shape information of objects. Moreover, some low-level visual features, such as object shading, surface texture and an object's contour or binocular disparity, have recently been proposed to describe 3D object representation [15-19]. However, 3D object recognition is primarily influenced by position variations and illumination source type, and the relative positions of an observer and object.

Some advanced theorems of 3D object perception have been investigated to solve these issues and enhance the 3D object recognition task [20]. Existing theorems for high-level 3D object perception can be categorized as object-centered and viewer-centered representations based on a coordinate system [2], and as volume-based (or model-based) and view-based representations based on the constituent elements [21]. Viewer-centered representation describes portions of an object relative to a coordinate system based on an observer. A view-based representation characterizes a 3D object using a set of object views. Both viewer-centered and view-based frameworks conform to the intuition of human perception, during which a person memorizes an object using several primary views without requiring an exhaustive 3D object model. Moreover, S. Kim *et al.* [22] proposed a combined model-based method to recognize 3D objects using a combination of a bottom-up process (model parameter initialization) and a top-down process (model parameter optimization).

1.1.2 Human Posture Recognition

Human posture recognition is an important example of 3D object recognition. A considerable number of studies have been made on this field over the past 10 years [23-24]. Existing approaches [25] for human posture recognition are classified as

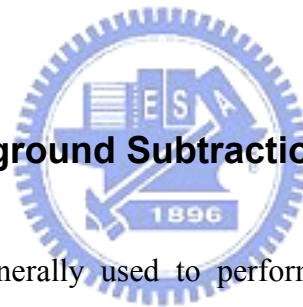
direct and indirect approaches based on the human body model. The model has either a 2D or 3D representation based on the dimensionality of features. The direct approach typically consists of a detailed human body model. For example, Ghost [26] developed a silhouette-based body model, incorporating hierarchical body pose estimation, a convex hull analysis of the silhouette and a partial mapping from body parts to silhouette segments. Furthermore, Pfander [27] utilized color information to develop a multi-class statistical model and identified human body parts using shape detection. However, occlusions and perspective distortion lead to the unreliable results. The indirect approach extracts features about the human body instead of a detailed human body model, and combines classifiers to estimate human posture. For example, Ozer *et al.* [28] utilized the AC-coefficients as the features and adopted principal component analysis (PCA) as the classifier. A recent work [29] used color, edge and shape as the features and the hidden Markov model as the classifier. Furthermore, complex 3D models utilize different equipment to solve problems associated with the angle from which human postures are observed. For instance, Delamarre *et al.* [30] proposed a method for building a 3D human body via three or more cameras, and then calculated the projection of the silhouette for comparison with 2D projections in a database. Additionally, 3D laser scanners [31] or thermal cameras [32] have also been adopted to build a 3D human body model. However, these 3-D solutions require enormous computing time and high device costs.

1.1.3 Scene Recognition

Recognizing scene can be addressed as a problem of 3D object recognition, where the scene represents variations due to changing the viewer location or camera pose [33-39]. Scene recognition is a fundamental element in the topological representation

of environment [40-41], where the graph node of the adjacency graph describes the robot's location. Moreover, scene recognition can also be employed to memorize and detect visual landmarks in geometrical representation of environments [42-43]. In [44], a series of experiments were presented to show that only the overall geometry and a few key features are required to perform scene recognition. For capturing the key features, Kröse *et al.* [45] proposed a method for appearance-based modeling of an environment by extracting scene features using PCA. Oliva *et al.* [46] proposed a scene-center-based approach to estimate the structure of a scene image by the mean of global image features. Moreover, a framework combined with a supervised method for recognizing the door and an unsupervised method for learning door-reaching behavior has been proposed in [47].

1.2 Overview of Background Subtraction



A reference image is generally used to perform background subtraction. The simplest means of obtaining a reference image is by averaging a period of frames [48]. However, it is not suitable to apply time averaging on the home-care applications because the foreground objects (especially for the elderly people or children) usually move slowly and the household scene changes constantly due to light variations from day to night, switches of fluorescent lamps and furniture movements etc. In short, the deterministic methods such as the time averaging have been found to have limited success in practice. For indoor environments, a good background model must also handle the effects of illumination variation, and the variation from background and shadow detection. Furthermore, if the background model cannot handle the fast or slow variations from sunlight or fluorescent lamps, the entire image will be regarded as foreground. That is, a single model cannot represent the distribution of pixels with

twinkling values. Therefore, to describe a background pixel by a bi-model instead of a single model is necessary in home-care applications in the real world.

Two approaches were generally adopted to build up a bi-model of background pixel. The first approach is termed the parametric method, and uses single Gaussian distribution [27] or mixtures of Gaussian [49] to model the background image. Attempts were made to improve the GMM methods to effectively design the background model, for example, using an on-line updated algorithm of GMM [50] and the Kalman filter to track the variation of illumination in the background pixel [51]. Furthermore, motion information is used for extracting a set of regions that have coherent motion to improve the efficiency of region classification and computing time [52]. The second approach is called the non-parametric method, and uses the kernel function to estimate the density function of background images [53].

Another important consideration is the shadows and highlights. Numerous recent studies have attempted to detect the shadows and highlights. Stockham [54] proposed that a pixel contains both an intensity value and a reflection factor. If a pixel is termed the shadow, then a decadent factor is implied on that pixel. To remove the shadow, the decadent factor should be estimated to calculate the real pixel value. Rosin [55] proposed that shadow is equivalent to a semi-transparent region, and uses two properties for shadow detection. Moreover, Elgammal *et al.* [53] tried to convert the RGB color space to the rgb color space (chromaticity coordinate). Because illumination change is insensitive in the chromaticity coordinate, shadows are not considered the foreground. However, lightness information is lost in the rgb color space. To overcome this problem, a measure of lightness is used at each pixel [53]. However, the static thresholds are unsuitable for dynamic environment.

1.3 Outline of Proposed System

Figure 1-1 illustrates the block diagram of the proposed framework. In the foreground detection block, image pixels of a 2D view are classified among foreground, shadow, highlight and background. The foreground pixels are applied to extract features for 3D object recognition. In the 3D object recognition block, one or more similarity measures are applied on the extracted features of a testing 2D view and all the objects in a 3D object database. The top three similar objects in the database (Top 3 Matches) are regarded as the recognition results.

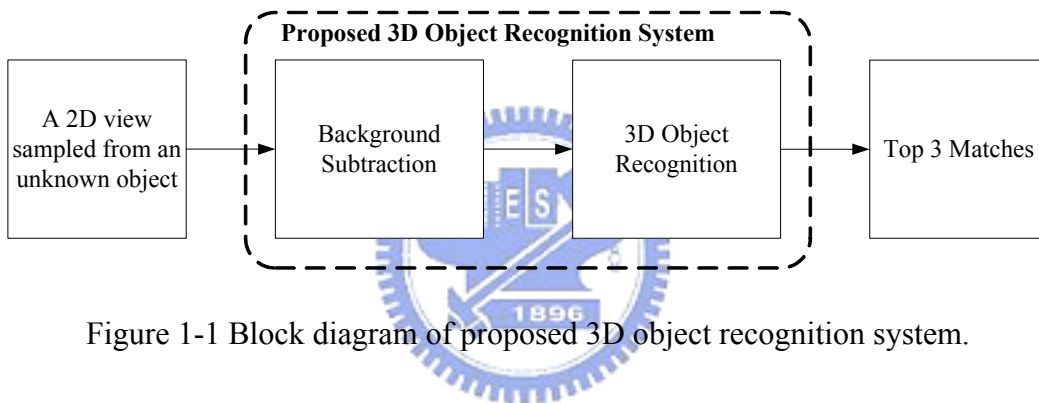


Figure 1-1 Block diagram of proposed 3D object recognition system.

1.3.1 Background Subtraction

In the background subtraction block (Fig. 1-2), a background subtraction scheme with highlight and shadow removal (BSHSR) is proposed to extract foreground regions with three proposed background models. First, the CBM is applied for extracting foreground candidates pixels. After that, the CSIM is applied for classifying foreground candidate pixels among real foreground, shadow and highlight. Finally, the GBM is used for eliminating false foreground pixels from the foreground candidate pixels. Moreover, only those real foreground pixels are reserved for further processing.

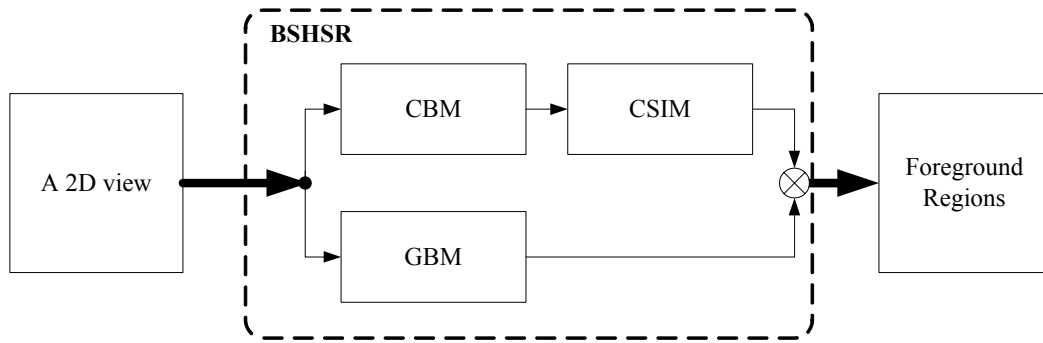


Figure 1-2 Block diagram of foreground detection.

1.3.2 3D Object Recognition

In the 3D object recognition block (Fig.1-3), a main 3D object database (MOD) is built during the building procedure with the proposed incremental similarity-based aspect-graph (ISAG). After that, one or more features are extracted from those foreground region to measure similarity among the objects in the MOD. After a weighted combination of all similarity measures, the top 3 similar objects are regarded as the recognition results.

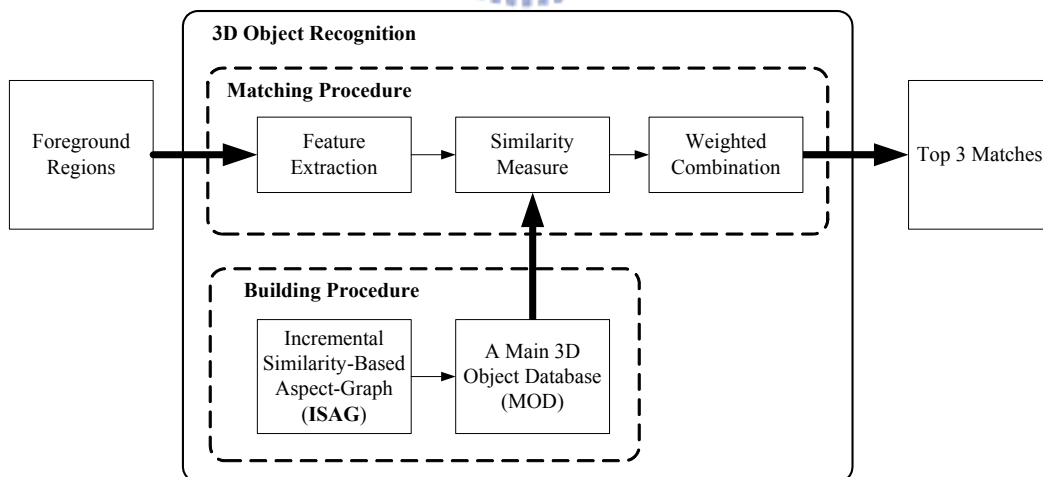


Figure 1-3 Block diagram of 3D object recognition.

1.4 Contribution of this Dissertation

The problem we address in this work is as follows: Given a set of 2D views of 3D objects, such as rigid objects, human postures, and scenes, how do we represent these 3D objects with collected 2D views in an efficient way? In other words, how do we extract the representative views from these 2D views such that a 3D object can be indexed efficiently? For the purpose of this work, we propose a flexible 3D object recognition framework for building the 3D object database and recognizing 3D objects with 2D views. We place emphasis on three issues in practice, which are listed as follows.

1. For extracting the object features from a complex background, a background subtraction scheme involving highlight and shadow removal (**BHSR**) is proposed as the pre-processing stage of the framework. Foreground regions can be precisely extracted from 2D views using the BHSR despite illumination variations and dynamic background. The BHSR comprises three models, called the color-based probabilistic background model (CBM), the gradient-based version of the color-based probabilistic background model (GBM) and a cone-shape illumination model (CSIM).
2. An incremental database construction learning method based on similarity-based aspect-graph (**ISAG**) is proposed for building the 3D object database using 2D views. The accuracy of the object representation increases with minimal growth of search space while collecting additional new object views.
3. For improving the robustness and computing time, a hierarchical matching structure is proposed to decide the final recognition result with a weighted combination of the results from multiple features.

1.5 Dissertation Organization

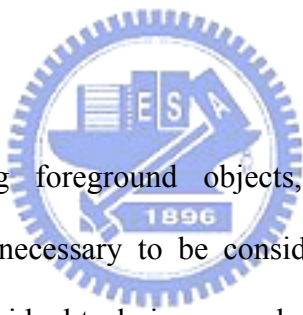
This chapter provides a brief introduction of the background subtraction system and 3D object recognition, including rigid object recognition, human posture recognition and scene recognition. This chapter also briefly discusses two main components in the proposed 3D object recognition framework. The remainder of this dissertation is organized as follows. Chapter 2 presents the proposed background subtraction algorithm (BSHSR), including the descriptions of the CBM, STCBM, LTCBM, GBM, and CSIM. Chapter 3 describes the ISAG and the proposed hierarchal matching structure. Chapter 4 presents experimental results that demonstrate the performance of the proposed method for 3D rigid objects, human postures and scene recognition. Finally, some concluding remarks and future researches are discussed in Chapter 5.



Chapter 2

Background Subtraction

2.1 Introduction



For precisely extracting foreground objects, environmental changes and shadow/highlight effects are necessary to be considered. Despite the existence of abundance of research on individual techniques, as described in Chapter 1, few efforts have been made to investigate the integration of environmental changes and shadow/highlight effects. In this work, we proposed a scheme that combines the color-based background model (CBM), the gradient-based background model (GBM) and the cone-shape illumination model (CSIM) to solve the issue in practice.

The remainder of this chapter is organized as follows. Section 2.2 describes the system architecture and the corresponding dataflow. Section 2.3 describes the statistical learning method used in the probabilistic modeling and defines the STCBM and LTCBM. Section 2.4 then proposes the CSIM using the STCBM and LTCBM to classify shadows and highlights efficiently. A hierarchical background subtraction framework that combined with color-based subtraction, gradient-based subtraction

and shadow and highlight removal was then described to extract the real foreground of an image. Finally, Section 2.6 presents discussions and conclusions.

2.2 System Architecture

Figure 2-1 illustrates the block diagram of the BSHSR. The BSHSR comprises three main models which are called the CBM, GBM and CSIM. The CBM comprises the LTCBM and STCBM, where the LTCBM is defined to record the background changes during a long period and STCBM is defined to record the background changes during a short period. Moreover, the STCBM and LTCBM are used to determine the parameters of the GBM and CSIM with a selection rule. Four stages are involved in the BSHSR. First, color-based background subtraction is performed on the input image for extracting the foreground candidates via the LTCBM. After that, shadow and highlight removal is performed on the foreground candidates via the CSIM for classifying the pixels of foreground candidates among real foreground, shadow and highlight. For eliminating the false foreground regions, gradient-based background subtraction is performed on the input image via the GBM. Finally, a hierarchal background subtraction is performed for combing the results from the CSIM and GBM.

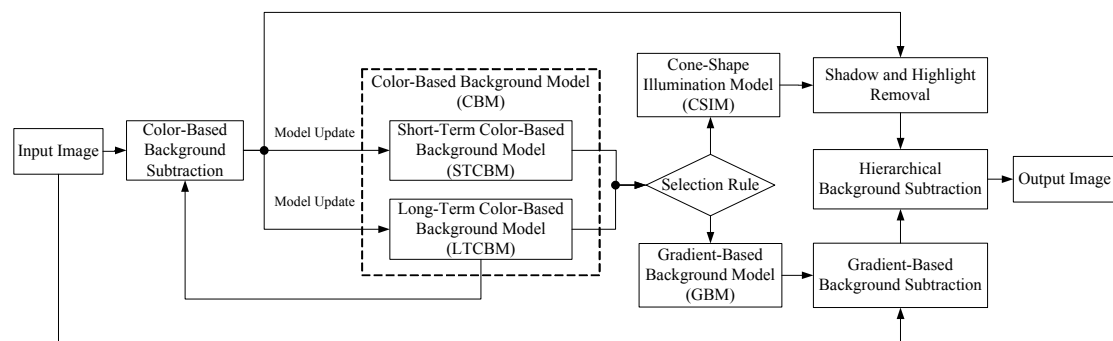


Figure 2-1 The block diagram of the BSHSR.

2.3 Background Modeling

Our previous investigation [56] studied a CBM to record the activity history of a pixel via GMM. However, the foreground regions generally suffer from rapid intensity changes and require a period of time to recover themselves when objects leave the background. In this work, the STCBM and LTCBM are defined and applied to improve the flexibility of the gradient-based subtraction that proposed by Javed *et.al* [57]. The features of images used in this work include pixel color and gradient information. This study assumes that the density functions of the color features and gradient features are both Gaussian distributed.

2.3.1 Color-Based Background Modeling

First, each pixel x is defined as a 3-dimensional vector (R, G, B) at time t . N Gaussian distributions are used to construct the GMM of each pixel, which is described as Eq. (2-1).

$$f(x|\lambda) = \sum_{i=1}^N w_i \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right) \quad (2-1)$$

where λ represents the parameters of GMM,

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, 2, \dots, N \text{ and } \sum_{i=1}^N w_i = 1$$

Suppose $X = \{x_1, x_2, \dots, x_m\}$ is defined as a training feature vector containing m pixel values collected from a pixel among a period of m image frames. The next step is calculating the parameter λ of GMM of each pixel so that the GMM can match the distribution of X with minimal errors. A common method for calculating

λ is the maximum likelihood (ML) estimation. ML estimation aims to find model parameters by maximizing the GMM likelihood function. ML parameters can be obtained iteratively using the expectation maximization (EM) algorithm [58] and the ML estimation of λ is defined as Eq. (2-2).

$$\lambda_{ML} = \arg \max_{\lambda} \sum_{j=1}^m \log f(x_j | \lambda) \quad (2-2)$$

The EM algorithm involves two steps; the parameters of GMM can be derived by iteratively using the Expectation step equation and Maximum step equation, as Eqs. (2-3) and (2-4).

■ Expectation step: (E step)

$$\beta_{ji} = \frac{w_i f(x_j | \mu_i, \Sigma_i)}{\sum_{k=1}^m a_k f(x_j | \mu_k, \Sigma_k)}, i = 1, \dots, N, j = 1, \dots, m \quad (2-3)$$

β_{ji} denotes the posterior probability that the feature x_j belongs to the i th Gaussian component distribution.

■ Maximum step: (M step)

$$\begin{aligned} \hat{w}_i &= \frac{1}{N} \sum_{j=1}^m \beta_{ji} \\ \hat{\mu}_i &= \sum_{j=1}^m \beta_{ji} x_j / \sum_{j=1}^m \beta_{ji} \\ \hat{\Sigma}_i &= \sum_{j=1}^m \beta_{ji} (x_j - \hat{\mu}_i)(x_j - \hat{\mu}_i)^T / \sum_{j=1}^m \beta_{ji} \end{aligned} \quad (2-4)$$

The termination criteria of the EM algorithm are as follows:

1. The increment between the new log-likelihood value and the last log-likelihood value is below a minimum increment threshold.
2. The iterative count exceeds a maximum iterative count threshold.

Suppose an image contains $S = W \times H$ pixels, where W means the image width and H means the image height. There are total S GMMs should be calculated by the EM algorithm with the collected training feature vector of each pixel.

Moreover, this study uses the K-means algorithm [59], which is an unsupervised data clustering used before the EM algorithm iterations to accelerate the convergence. First, N random values are chosen from X and assigned as the center of each class. Then the following steps are applied to cluster the m values of the training feature vector X .

1. To calculate 1-norm distances between the m values and the N center values. Each value of X is classified to the class having the minimum distance with it.
2. After clustering all the values of X , re-calculate each class center by calculating the mean of the values among each class.
3. Calculate the 1-norm distances between the m values and the N new center values. Each value of X is classified to the class which has the minimum distance with it. If the new clustering result is the same as the clustering result before re-calculating each class center, then stop, otherwise return to previous step to calculate the N new center values.
4. After applying K-means algorithm to cluster the values of X , the mean of each class is assigned as the initial value of μ_i , the maximum distance among the points of each class is assigned as the initial value of Σ_i , and the value of w_i is initialized as $1/N$.

2.3.2 Model Maintenance of the LTCBM and STCBM

According to the above sections, an initial color-based probabilistic background model is created using the training feature vector set X with N Gaussian distributions and N is usually defined as 3 to 5 based on the observation over a short period of time m . However, when the background changes are recorded over time, it is possible that more different distributions from the original N distributions are observed. If the GMM of each pixel contains only N Gaussian distributions, only N background distributions are reserved and other collected background information is lost and it is not flexible to model the background with only N Gaussian distributions.

To maintain the representative background model and improve the flexibility of the background model simultaneously, an initial LTCBM is defined as the combination of the initial color-based probabilistic background model and extra N new Gaussian distributions (total $2N$ distributions), an arrangement inspired by the work of [60]. Kaew *et al.* [49] proposed a method of sorting the Gaussian distributions based on the fitness value w_i/σ_i ($\sum_i = \sigma_i^2 I$), and extracted a representative model with a threshold value B_0 .

After sorting the first N Gaussian distributions with fitness value, b ($b \leq N$) Gaussian distributions are extracted with Eq. (2-5).

$$B = \arg \min_b \sum_{j=1}^b w_j > B_0 \quad (2-5)$$

The first b Gaussian distributions are defined as the elected color-based background model (ECBM) to be the criterion to determine the background. Meanwhile, the remainders ($2N-b$) of the Gaussian distributions are defined as the

candidate color-based background model (CCBM) for dealing with the background changes. Finally, the LTCBM is defined using the combination of the ECBM and CCBM. Figure 2-2 shows the block diagram to illustrate the process of building the initial LTCBM, ECBM and CCBM.

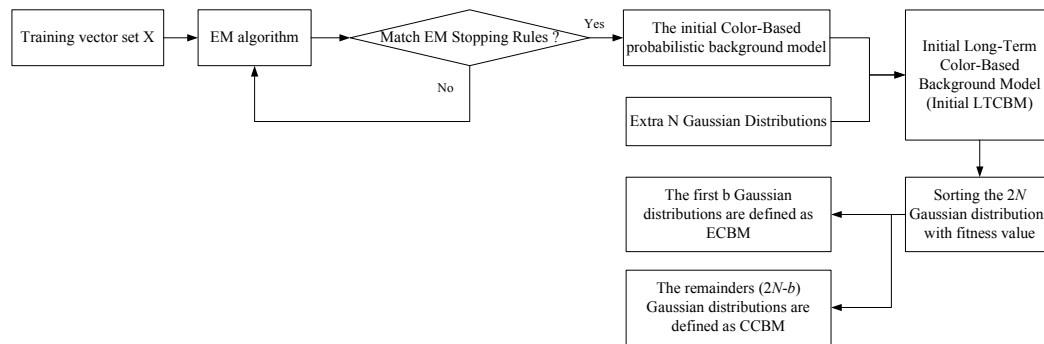


Figure 2-2 Block diagram showing the process of building the initial LTCBM, ECBM and CCBM.

The Gaussian distributions of the ECBM mean the characteristic distributions of “background”. Therefore, if a new pixel value belongs to any of the Gaussian distributions of the ECBM, the new pixel is regarded as “a pixel contains the property of background” and the new pixel is classified as “background”. In this work, a new pixel value is considered as background when it belongs to any Gaussian distribution in the ECBM and has a probability not exceeding 2.5 standard deviations away from the corresponding distribution. If none of the b Gaussian distributions match the new pixel value, a new test is conducted by checking the new pixel value against the Gaussian distributions in the CCBM. The parameters of the Gaussian distributions are updated via Eq. (2-6). ρ and α are termed the learning rates, and determine the update speed of the LTCBM. Moreover, $\hat{p}(w_i^t | X_i^{t+1})$ results from background subtraction which is set to 1 if a new pixel value belongs to the i^{th} Gaussian distribution. If a new incoming pixel value does not belong to any of the Gaussian

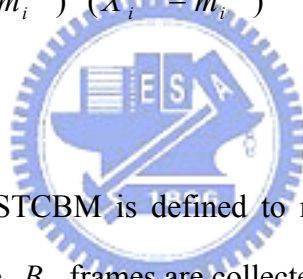
distributions in the CBM and the number of Gaussian components in the CCBM is below $(2\tilde{N}b)$, a new Gaussian distribution is added to reserve the new background information with three parameters: the current pixel value as the mean, a large predefined value as the initial variance, and a low predefined value as the weight. Otherwise, the $(2N - b)^{th}$ Gaussian distribution in the CCBM is replaced by the new one. After updating the parameters of the Gaussian components, all Gaussian distributions in the CBM are resorted by recalculating the fitness values.

$$w_i^{t+1} = (1 - \alpha)w_i^t + \alpha \hat{p}(w_i^t | X_i^{t+1}) \quad (2-6)$$

$$m_i^{t+1} = (1 - \rho)m_i^t + \rho X_i^{t+1}$$

$$\Sigma_i^{t+1} = (1 - \rho)\Sigma_i^t + \rho(X_i^{t+1} - m_i^{t+1})^T (X_i^{t+1} - m_i^{t+1})$$

$$\rho = \alpha g(X_i^{t+1} | m_i^t, \Sigma_i^t)$$



Unlike the LTCBM, the STCBM is defined to record the background changes during a short period. Suppose B_1 frames are collected during a short period B_1 and then B_1 new incoming pixels for each pixel are collected and defined as a test pixel set $P = \{p_1, p_2, \dots, p_q, \dots, p_{B_1}\}$, where p_q means the new incoming pixel at time q . A test pixel set P is defined and used for calculating the STCBM and a result set S is then defined and calculated by comparing P with the LTCBM and is described as Eq.(2-7), where I_q means the result after background subtraction, which means the index of Gaussian distribution of the initial LTCBM, R_q means the index of resorting result for each Gaussian distribution after each update, and F_q means the reset flag of each Gaussian distribution.

$$S = \{S_1, S_2, \dots, S_q, \dots, S_{B_1}, \text{and } S_q = (I_q, R_q(i), F_q(i))\} \quad (2-7)$$

where $1 \leq I_q \leq 2N$, $1 \leq R_q(i) \leq 2N$, $F_q(i) \in \{0, 1\}$, $1 \leq i \leq 2N$

The histogram of CG is then given using Eq. (2-8).

$$H_{CG}(k) = \sum_k [\delta(k - (I_q + R_q(I_q))) + \bar{F}_q \cdot \sum_{q'} \delta(k - (I_{q'} + R_{q'}(I_{q'})))] / B_1 \quad (2-8)$$

where $1 \leq k \leq 2N$, $1 \leq q \leq B_1$, $1 \leq q' < q$

In brief, four Gaussian distributions are used to explain how Eqs. (2-7) and (2-8) work and the corresponding example is listed in Table 2-1. At first, the original CBM contains four Gaussian distributions ($2N = 4$), and the index of Gaussian distribution in the initial CBM is fixed (1,2,3,4). At the first time, a new incoming pixel which belongs to the second Gaussian distribution compares with the CBM, so the result of background subtraction is $I_q = 2$. Moreover, the CBM is updated with Eq. (2-6) and the index of Gaussian distribution in the CBM is changed. When the order of the first and second Gaussian distributions is changed, $R_q(i)$ records the change states; for example, $R_q(1) = 1$ means the first Gaussian distribution has moved forward to the second one, and $R_q(2) = -1$ means the second Gaussian distribution has moved backward to the first one. At the second time, a new incoming pixel which belongs to the second Gaussian distribution based on the initial CBM is classified as the first Gaussian distribution ($I_q = 1$) based on the latest order of the CBM. However, the CG histogram can be calculated according to the original index of the initial CBM with the latest order of the CBM and $R_q(i)$, such that $H_{CG}(I_q + F_q = 2)$ will be accumulated with one. Moreover, $R_q(i)$ changes while the order of Gaussian

distributions changes. For example, at the fifth time in Table 2-1, the order of CBM changes from (2,1,3,4) to (1,2,3,4), and then $R_q(1) = 1 - 1 = 0$ means the first Gaussian distribution of the initial CBM has moved back to the first one of the latest CBM, and $R_q(2) = -1 + 1 = 0$ means the second Gaussian distribution has moved back to the second one of the latest CBM.

Table 2-1 An example to calculate CG histogram

TIME (q)	INDEX OF INITIAL CBM	1	2	3	4	TIME (q)	INDEX OF INITIAL CBM	1	2	3	4
1	Index of CCBM at time q	1	2	3	4	4	Index of CCBM at time q	2	1	3	4
	p_q		*				p_q		*		
	I_q	2					I_q	2			
	R_q	0	0	0	0		R_q	1	-1	0	0
	F_q	0	0	0	0		F_q	0	0	0	0
	CG	0	1	0	0		CG	2	2	0	0
2	Index of CCBM at time q	2	1	3	4	5	Index of CCBM at time q	1	2	3	4
	p_q	*					p_q	*			
	I_q	1					I_q	1			
	R_q	1	-1	0	0		R_q	0	0	0	0
	F_q	0	0	0	0		F_q	0	0	0	0
	CG	0	2	0	0		CG	3	2	0	0
3	Index of CCBM at time q	2	1	3	4	6	Index of CCBM at time q	1	2	3	4
	p_q		*				p_q			*	
	I_q	2					I_q	3			
	R_q	1	-1	0	0		R_q	0	0	0	0
	F_q	0	0	0	0		F_q	0	0	0	0
	CG	1	2	0	0		CG	3	2	1	0

If a new incoming pixel p_q matches the i^{th} Gaussian distribution that has the least fitness value, the i^{th} Gaussian distribution is replaced with a new one and the flag F_q will be set to 1 to reset the accumulated value of $H_{CG}(i)$. Figure 2-3 shows the block diagram about the process of calculating H_{CG} .

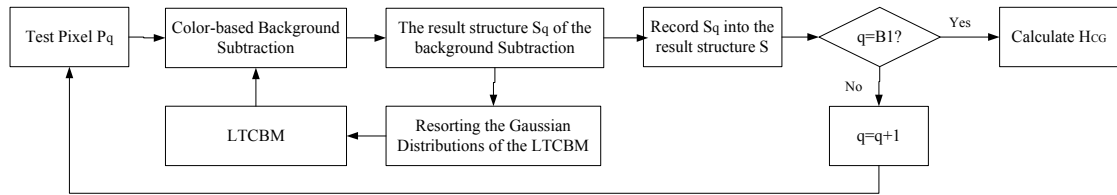


Figure 2-3 Block diagram showing the process to calculate H_{CG} .

After matching all test pixels to the corresponding Gaussian distribution, the result set S can be used to calculating H_{CG} using I_q and F_q . With the reset flag F_q , the STCBM can be built up rapidly based on a simple idea, threshold on the occurring frequency of Gaussian distribution. That is to say, the short-term tendency of background changes is apparent if an element of $H_{CG}(k)$ is above a threshold value B_2 during a period of frames B_1 . In this work, B_1 is assigned a value of 300 frames and B_2 is set to be 0.8. Therefore, the representative background component in the short-term tendency can be determined to be k if the value of $H_{CG}(k)$ exceeds 0.8, otherwise, the STCBM provides no further information on background model selection.

2.3.3 Gradient-Based Background Modeling

Javed *et.al* [57] developed a hierarchical approach that combines color and gradient information to solve the problem about rapid intensity changes. Javed *et.al* [57] adopted the k^{th} , highest weighted Gaussian component of GMM at each pixel to

obtain the gradient information to build the gradient-based background model. The choice of k in [57] is similar to select k based only on the ECBM defined in this work. However, choosing the highest weighted Gaussian component of GMM leads to the loss of the short term tendencies of background changes. Whenever a new Gaussian distribution is added into the background model, it is not selected owing to its low weighting value for a long period of time. Consequently, the accuracy of the gradient-based background model is reduced for that the gradient information is not suitable for representing the current gradient information.

To solve this problem, both STCBM and LTCBM are considered in selecting the value of k for developing a more robust gradient-based background model and maintaining the sensitivity to short-term changes. When the STCBM provides a representative background component (says the k_s^{th} bin in the STCBM), k is set to k_s rather than the highest weighted Gaussian distribution.

Let $x_{i,j}^t = [R, G, B]$ be the latest color value that matched the k_s^{th} distribution of the LTCBM at pixel location (i, j) , then the gray value of $x_{i,j}^t$ is applied to calculate the gradient-based background subtraction. Suppose the gray value of $x_{i,j}^t$ is calculated as Eq. (2-9), then $g_{i,j}^t$ will be distributed as Eq. (2-10) based on independence among RGB color channels,

$$g_{i,j}^t = \alpha R + \beta G + \gamma B \quad (2-9)$$

$$g_{i,j}^t \sim N(m_{i,j}^t, (\sigma_{i,j}^t)^2) \quad (2-10)$$

where

$$m_{i,j}^t = \alpha \mu_{i,j}^{t,k_s,R} + \beta \mu_{i,j}^{t,k_s,G} + \gamma \mu_{i,j}^{t,k_s,B}$$

$$\sigma_{i,j}^t = \sqrt{\alpha^2 (\sigma_{i,j}^{t,k_s,R})^2 + \beta^2 (\sigma_{i,j}^{t,k_s,G})^2 + \gamma^2 (\sigma_{i,j}^{t,k_s,B})^2}$$

After that, the gradient along the x axis and y axis can be defined as $f_x = g_{i+1,j}^t - g_{i,j}^t$ and $f_y = g_{i,j+1}^t - g_{i,j}^t$. From the work of [57], f_x and f_y have the distributions defined in Eqs. (2-11) and (2-12).

$$f_x \sim N(m_{f_x}, (\sigma_{f_x})^2) \quad (2-11)$$

$$f_y \sim N(m_{f_y}, (\sigma_{f_y})^2) \quad (2-12)$$

where

$$m_{f_x} = m_{i+1,j}^t - m_{i,j}^t$$

$$m_{f_y} = m_{i,j+1}^t - m_{i,j}^t$$

$$\sigma_{f_x} = \sqrt{(\sigma_{i+1,j}^t)^2 + (\sigma_{i,j}^t)^2}$$

$$\sigma_{f_y} = \sqrt{(\sigma_{i,j+1}^t)^2 + (\sigma_{i,j}^t)^2}$$



Suppose $\Delta_m = \sqrt{f_x^2 + f_y^2}$ is defined as the magnitude of the gradient for a pixel, $\Delta_d = \sqrt{\tan^{-1}(f_x / f_y)}$ is defined as its direction (the angle with respect to the horizontal axis), and $\Delta = [\Delta_m, \Delta_d]$ is defined as the feature vector for modeling the gradient-based background model. The gradient-based background model based on feature vector $\Delta = [\Delta_m, \Delta_d]$ then can be defined as Eq. (2-13).

$$F^k(\Delta_m, \Delta_d) = \frac{\Delta_m}{2 \Pi \sigma_{f_x}^k \sigma_{f_y}^k \sqrt{1 - \rho^2}} \exp\left(-\frac{z}{2(1 - \rho^2)}\right) > T_g \quad (2-13)$$

where

$$z = \left(\frac{\Delta_m \cos \Delta_d - \mu_{f_x}}{\sigma_{f_x}} \right)^2 - 2\rho \left(\frac{\Delta_m \cos \Delta_d - \mu_{f_x}}{\sigma_{f_x}} \right) \left(\frac{\Delta_m \sin \Delta_d - \mu_{f_y}}{\sigma_{f_y}} \right) + \left(\frac{\Delta_m \sin \Delta_d - \mu_{f_y}}{\sigma_{f_y}} \right)^2$$

$$\rho = \frac{(\sigma_{i,j}^t)^2}{\sigma_{f_x} \sigma_{f_y}}$$

2.4 Background Subtraction with Shadow Removal

2.4.1 Shadow and Highlight Removal

Besides foreground and background, shadows and highlights are two important phenomena that should be considered in most cases. Shadows and highlights result from changes in illumination. Compared with the original pixel value, shadow has similar chromaticity but lower brightness, and highlight has similar chromaticity but higher brightness. The regions influenced by illumination changes are classified as the foreground if shadow and highlight removal is not performed after background subtraction.

Hoprasert *et al.* [60] proposed a method of detecting highlight and shadow by gathering statistics from N color background images. Brightness and chromaticity distortion are used with four threshold values to classify pixels into four classes. The method that used the mean value as the reference image in [60] is not suitable for dynamic background. Furthermore, the threshold values are estimated based on the histogram of brightness distortion and chromaticity distortion with a given detection rate, and are applied to all pixels regardless of the pixel values. Therefore, it is possible to classify the darker pixel value as shadow. Furthermore, it cannot record the history of background information.

This work proposes a 3D cone model that is similar to the pillar model proposed by Hoprasert [60], and combines the LTCBM and STCBM to solve the above problems. A cone model is proposed with the efficiency in deciding the parameters of 3D cone model according to the proposed LTCBM and STCBM. In the RGB space, a Gaussian distribution of the LTCBM becomes an ellipsoid whose center is the mean of the Gaussian component, and the length of each principle axis equals 2.5 standard deviations of the Gaussian component. A new pixel $I(R, G, B)$ is considered to belong to background if it is located inside the ellipsoid. The chromaticities of the pixels located outside the ellipsoid but inside the cone (formed by the ellipsoid and the origin) resemble the chromaticity of the background. The brightness difference is then applied to classify the pixel as either highlight or shadow. Figure 2-4 illustrates the 3D cone model in the RGB color space.

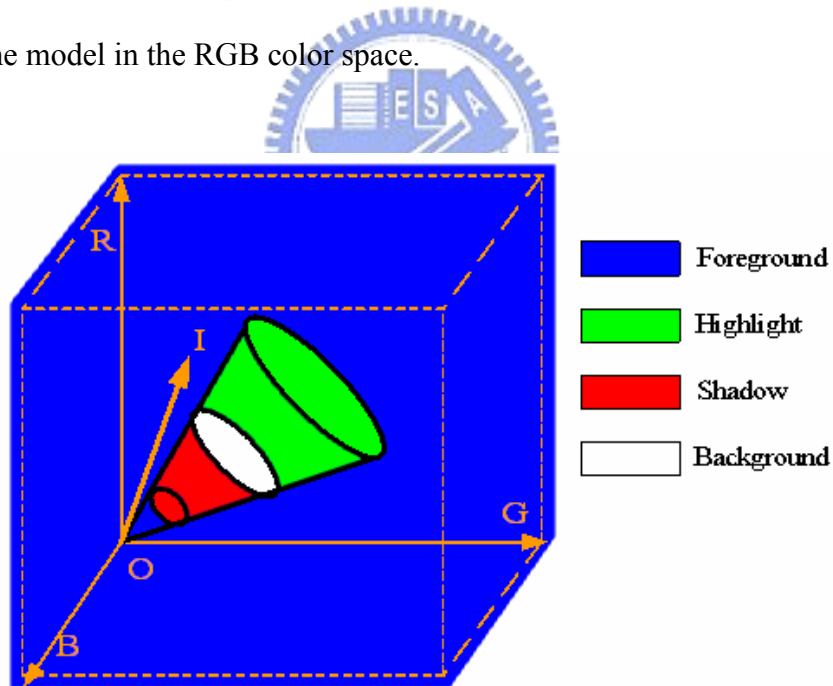


Figure 2-4 The proposed 3D cone model in the RGB color space.

The threshold values τ_{low} and τ_{high} are applied to avoid classifying the darker pixel value as shadow or the brighter value as highlight, and can be selected based on the standard deviation of the corresponding Gaussian distribution in the CBM.

Because the standard deviations of the R, G and B color axes are different, the angles between the curved surface and the ellipsoid center are also different. It is difficult to classify the pixel using the angles in the 3D space. The 3D cone is projected onto the 2D space to classify a pixel using the slope and the point of tangency. Figure 2-5 illustrates the projection of the 3D cone model onto the RG 2D space.

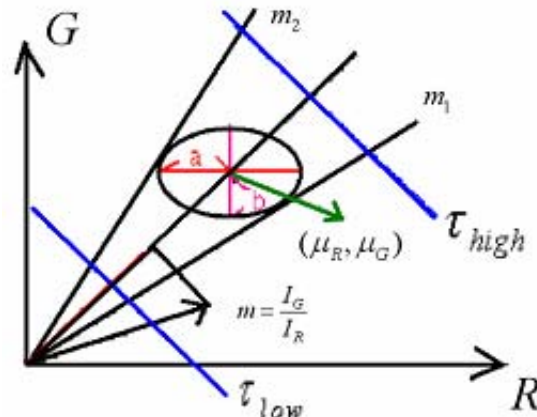


Figure 2-5 2D projection of the 3D cone model from RGB space onto the RG space.

Let a and b denote the lengths of major and minor axis of the ellipse, where $a = 2.5 * \sigma_R$ and $b = 2.5 * \sigma_G$. The center of the ellipse is (μ_R, μ_G) , and the elliptical equation is described as Eq. (2-14).

$$\frac{(R - \mu_R)^2}{a^2} + \frac{(G - \mu_G)^2}{b^2} = 1 \quad (2-14)$$

The line $G = mR$ is assumed to be the tangent line of the ellipse with the slope m . Equation (2-10) can then be solved using the line equation $G = mR$ with Eq.(2-15).

$$m_{1,2} = \frac{-(2\mu_R\mu_G) \pm \sqrt{(a^2 - \mu_R^2)^2 - 4(2\mu_R\mu_G)(b^2 - \mu_G^2)}}{2(a^2 - \mu_R^2)} \quad (2-15)$$

A matching result set is given by $F_b = \{f_{bi}, i = 1, 2, 3\}$, where f_{bi} is the matching result of a specific 2D space. A pixel vector $I = [I_R, I_G, I_B]$ is then projected onto the 2D spaces of R-G, G-B, and B-R. The pixel matching result is set to 1 when the slope of the projected pixel vector is between m_1 and m_2 . Meanwhile, if the background mean vector is $E = [\mu_R, \mu_G, \mu_B]$, the brightness distortion α_b can be calculated via Eq. (2-16)

$$\alpha_b = \frac{\|I\| \cos(\theta)}{\|E\|} \quad (2-16)$$

where

$$\theta = |\theta_I - \theta_E| = \left| \cos^{-1}\left(\frac{I_G}{\sqrt{I_R^2 + I_G^2}}\right) - \cos^{-1}\left(\frac{\mu_G}{\sqrt{\mu_R^2 + \mu_G^2}}\right) \right|$$

The image pixel is classified as highlight, shadow or foreground using the matching result set F_b , the brightness distortion α_b and Eq. (2-17).

$$C(i) = \begin{cases} \text{Shadow} & : \sum F_b = 3 \text{ and } \tau_{\text{low}} < \alpha_b < 1, \text{ else} \\ \text{Highlight} & : \sum F_b = 3 \text{ and } 1 < \alpha_b < \tau_{\text{high}}, \text{ else} \\ \text{Foreground} & : \text{otherwise} \end{cases} \quad (2-17)$$

When a pixel is a large standard deviation away from a Gaussian distribution, the Gaussian distribution probability of the pixel approximately equals to zero. It also means the pixel does not belong to the Gaussian distribution. By using the simple concept, τ_{high} and τ_{low} can be chosen using N_G standard deviation of the corresponding Gaussian distribution in the CBM and are described as Eq. (2-18).

$$\begin{aligned}\tau_{\text{high}} &= 1 + \frac{\|S\| \cdot \cos \theta_\tau}{\|E\|} \\ \tau_{\text{low}} &= 1 - \frac{\|S\| \cdot \cos \theta_\tau}{\|E\|}\end{aligned}\tag{2-18}$$

where

$$\|E\| = \sqrt{(\mu_R)^2 + (\mu_G)^2}$$

$$\|S\| = \sqrt{(\mathbf{N}_G \cdot \sigma_R)^2 + (\mathbf{N}_G \cdot \sigma_G)^2}$$

$$\theta_\tau = |\theta_E - \theta_S| = \left| \cos^{-1}\left(\frac{\mu_G}{\sqrt{\mu_R^2 + \mu_G^2}}\right) - \cos^{-1}\left(\frac{\sigma_G}{\sqrt{\sigma_R^2 + \sigma_G^2}}\right) \right|$$

2.4.2 Background Subtraction

A hierarchical approach combining color-based background subtraction and gradient-based background subtraction has been proposed by Javed *et.al* [57]. This work proposes a similar method for extracting the foreground pixels. Given a new image frame I , the color-based background model is set to the LTCBM and STCBM, and gradient-based model is $F^k(\Delta_m, \Delta_d)$. $C(I)$ is defined as the result of color-based background subtraction using the CBM. $G(I)$ is defined as the result of gradient-based background subtraction. $C(I)$ and $G(I)$ can be extracted by testing every pixel of frame I using the LTCBM and $F^k(\Delta_m, \Delta_d)$. Moreover, $C(I)$ and $G(I)$ are both defined as a binary image, where 1 represents the foreground pixel and 0 represents the background pixel. The foreground pixels labeled in $C(I)$ are further classified as shadow, highlight and foreground by using the proposed 3D cone model. $C'(I)$ can then be obtained from $C(I)$ after transferring the the foreground pixels which have been labeled as shadow and highlight in $C(I)$ into the

background pixel. The difference between Javed *et.al* [57] and the proposed method is that a pixel classifying procedure using the CSIM is applied before using the connected component algorithm to group all the foreground pixels in $C(I)$. The robustness of background subtraction is enhanced due to the better accuracy in $|\partial R_a|$. Moreover, the foreground pixels can be extracted using Eq. (2-19).

$$\frac{\sum_{(i,j) \in \partial R_a} (\nabla I(i,j)G(i,j))}{|\partial R_a|} \geq P_B \quad (2-19)$$

where ∇I denotes the edges of image I and ∂R_a represents the number of boundary pixels of region R_a .



Chapter 3

Incremental Similarity-Based

Aspect-Graph 3D Object Recognition



3.1 Introduction

The common challenge in 3D object recognition, human posture recognition, and scene recognition is the variation in orientations. The simplest method for solving this problem is to characterize an object with a densely sampled collection of independent views. The object can be described in detail by constructing an object model with numerous 2D views; however, this approach significantly increases computing time due to the expansive search space. Thus, several approaches have been developed to extract a minimal set of object views. Appearance-based methods focus on changes in intensity of each view. However, changes to object lighting, rotation, deformation and occlusion affect object recognition results when using the appearance-based method. Aspect-graph representations focus on shape changes to an object's projection [61-62]. Koenderink *et al.* [63] developed the underlying theory that describes 3D objects

using aspect-graph representation. Moreover, the traditional aspect-graph method [64] assumes that an object belongs to a limited class of shapes, and that characteristic views can be extracted using prior knowledge of the object. Aspect-graph vertices represent the characteristic views extracted from points on a transparent viewing sphere with an object in the object center. These characteristic views are extracted as prototypes of an object from a densely sampled collection of object views.

Cyr and Kimia [1] presented a similarity-based aspect-graph method to extract the characteristic views using shape similarity between views. The viewing sphere is sampled at regular (5 degree) intervals and two similarity metrics, which one based on curving matching and the other based on shocking matching, are applied to combine views into aspects. Let there be N objects $\{O_1, O_2, \dots, O_n, \dots, O_{N-1}, O_N\}$, which comprise an object database. Each object is composed of M views sampling the viewing sphere giving rise to a set of views $\{V_1^1, \dots, V_m^n, \dots, V_M^N\}$ where V_m^n denotes the m^{th} view of object O_n . The aspect p of object n is defined as A_p^n , which is a collection of views ranging from V_{m-k}^n to V_{m+k}^n and represented by the characteristic view V_m^n . Moreover, the dis-similarity of two views is represented as $d(V_m^n, V_j^i)$, which is the distance between the m^{th} view of object n and the i^{th} view of object j . The goal is to minimize the set of views required to represent each object O_n . Two criteria are imposed to maintain successful object recognition while forming aspects representation by characteristic views. The first criterion (local monotonicity) supposes that the dis-similarity of two views increases as their relative viewing angle between them increases. The second criterion describes that the distance of each view V_i^n in an aspect A_m^n and the characteristic view of that aspect V_m^n is smaller than

the distance between any non-aspect view V_j^n and the characteristic view V_m^n .

The training views of an object in [1], which are sampled at 5-degree increments and sorted by order, are collected in advanced. When additional views of an object are collected to improve object representation in the work of [1], the total views of an object must be resorted in order of view angles. The first criterion, local monotonicity, is not suitable when an object is symmetrical in the feature space. It is inconvenient to update the aspect-graph representation while collecting more new 2D views. To improve the flexibility of an update mechanism, this work presents an incremental database construction method for building and updating the aspect-graph with object views sampled at random intervals. Object representation becomes increasingly detailed using additional captured and characteristic views without re-calculating similarity measures by re-sorting total views. Moreover, the first criterion in the work of [1], local monotonicity, is not utilized, thereby improving flexibility of extracting aspects of symmetrical objects. Although the proposed approach cannot confirm the view angle of a test view using a specific object view, it improves the flexibility for building an aspect-graph representation, and reduces computing time when updating object aspects. Additionally, the accuracy of the object representation increases with minimal growth of search space while collecting additional new object views.

The remainder of this chapter is organized as follows. Section 3.2 presents the system architecture and the corresponding dataflow. Section 3.3 describes the procedure for extracting features and the similarity measures for building database and object matching. Section 3.4 describes the ISAG for extracting the aspects and characteristic views of objects. Furthermore, the object matching procedure is described with a weighting combination between different similarity measures. Conclusions are discussed in Section 3.5.

3.2 System Architecture

The proposed framework (Fig. 3-1) contains two parts, which are called the database building procedure and the matching procedure. Suppose an object database contains T_0 objects, and T_1 2D views of each object are randomly sampled from a viewing sphere. In the database building procedure, the ISAG (Fig. 3-2) is applied to extract the aspects of each object using T_1 2D views. The main 3D database contains a set of assistant 3D object databases (AOD). Furthermore, an AOD comprises the aspects of each object, where the aspects are represented by their characteristic views. Figure 3-3 illustrates the inner structure of an AOD. In Fig. 3-3, a set of aspects is employed to represent the database of a 3D object in the aspect level. The prototypes for these aspects, called the characteristic views, are utilized to represent an object for object matching. The passage from one characteristic view to another is defined with only the similarity measure. The proposed similarity-based aspect-graph focuses on an efficient learning method with associated features and similarity functions. While object features are sufficient to discriminate the similarity between each two 2D training views, the aspects and characteristic views can be extracted using associated similarity measures. Even if the objects are complex, the characteristic views can be extracted in the feature space.

In the matching procedure, a similarity measure is applied between a 2D view sampled from an unknown object and all the characteristic views of the 3D object database. After the weighted combination of all similarity measures, the first three characteristic views that have the highest similarity with the testing 2D view are regarded as the recognition results (the top three matches).

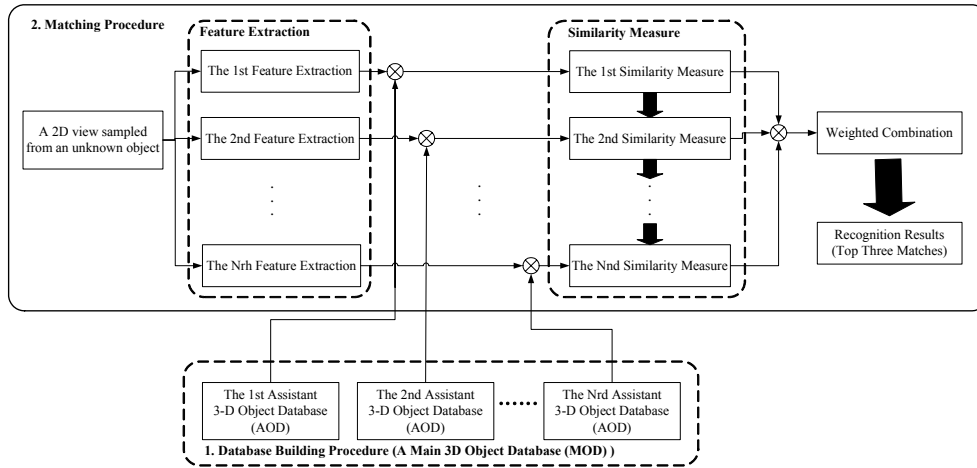


Figure 3-1 The system architecture of the proposed framework. A MOD comprises of total AODs.

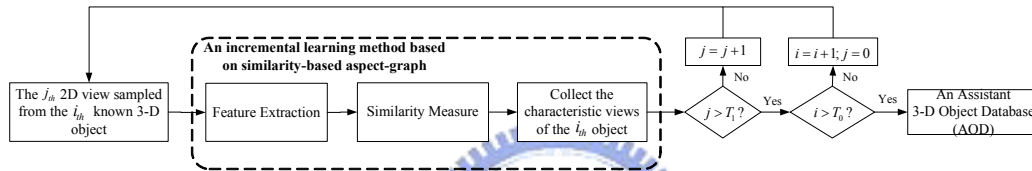


Figure 3-2 The database building procedure, where T_0 is the number of objects in the database and T_1 is the number of sampled views required to build the aspect-graph representation of an object.

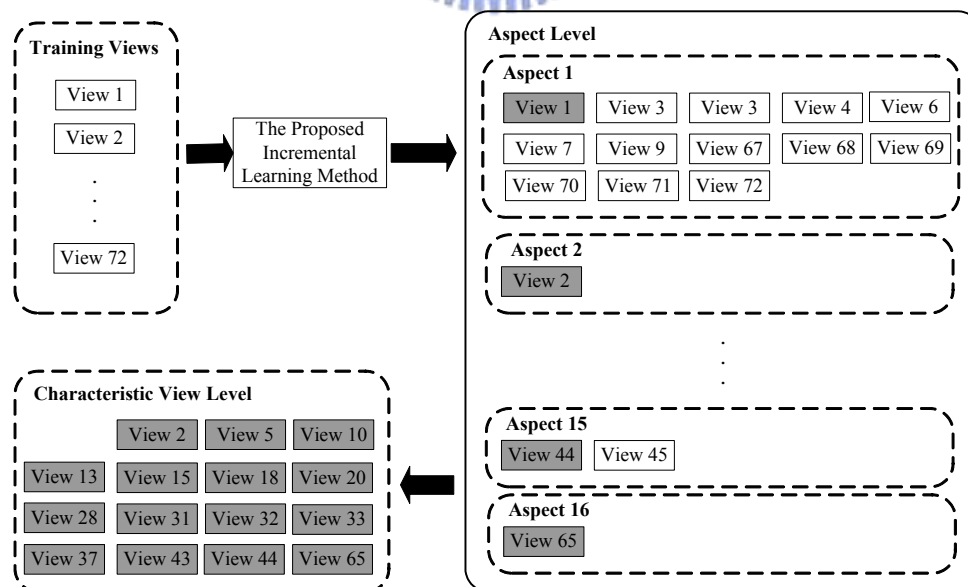
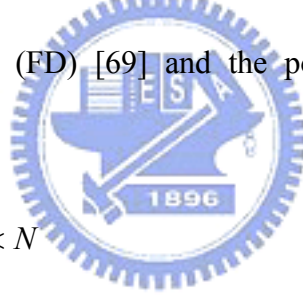


Figure 3-3 The inner structure of an AOD.

3.3 Object Representation

In this work, shape and color features are utilized to measure similarity between two object views. To extract shape information, a robust background subtraction framework from previous works [65-66] is utilized to extract foreground regions while considering shadows and highlights. Foreground detection provides flexibility when constructing the object database, even in an out-of-control environment. Canny edge detection [67] is then applied to extract shape edge, and the Gradient Vector Flow Snake (GVF) [68] is applied to extract the contour information. Assume that the contour information is included in a set \mathbf{Z} , which is composed of N points z_i , where z_i is a complex form given by Eq. (3-1). Two kinds of shape features, which are called the Fourier descriptor (FD) [69] and the point-to-point length (PPL), are extracted from \mathbf{Z} .

$$\mathbf{Z} = \{z(i)\} = \{x_i + jy_i\}, \quad 0 \leq i < N \quad (3-1)$$



3.3.1 Shape Features

The points inside the set \mathbf{Z} are re-sampled using Eq. (3-2) to eliminate variations in shift and scale.

$$\tilde{\mathbf{Z}} = \{\tilde{z}(i)\} = \{L_c[(x_i - x_c) + j(y_i - y_c)]/L\} \quad (3-2)$$

where $0 \leq i < N$; L denotes contour length of \mathbf{Z} , L_c is expected contour length, and (x_c, y_c) is the location of the contour center of \mathbf{Z} . Then, the Fourier transform is applied to $\tilde{\mathbf{Z}}$ to compute FD using Eq. (3-3).

$$FD(k) = \sum_{n=0}^{N-1} \tilde{z}(n) \exp(-j2\pi kn / N), 0 \leq k < N \quad (3-3)$$

The low-frequency parts of FD are extracted with the consideration of decreasing the variations of high-frequency noises, and are defined as MAG. Notably, MAG is composed of $2T_2$ magnitude values of frequency information selected among $2N$ frequencies. The method for extracting MAG is given by Eq. (3-4).

$$MAG = \{|FD(k)|, |FD(N-k)|, 1 \leq k \leq T_2\} \quad (3-4)$$

Intuitively speaking, MAG only characterizes the shape and not the orientation of human posture. Therefore, MAG cannot discriminate between similar shapes oriented differently. To solve this problem, phase information for FD must be used for memorizing an object. The work in [70] proposes that memorizing the phase value at low frequency is sufficient. Suppose the phase information is θ_z , then θ_z can be calculated using $FD(1)$ and $FD(N-1)$, as described in Eqs. (3-5) and (3-6).

$$FD(1) = |FD(1)| \cdot \exp(j\theta_1) = R_1 + jI_1 \quad (3-5)$$

$$FD(N-1) = |FD(N-1)| \cdot \exp(j\theta_{N-1}) = R_{N-1} + jI_{N-1} \quad (3-6)$$

Furthermore, θ_z can be calculated using Eq. (3-7).

$$\theta_z = (\theta_1 + \theta_{N-1}) / 2 = (\arctan(I_1 / R_1) + \arctan(I_{N-1} / R_{N-1})) / 2 \quad (3-7)$$

where R_1 and R_{N-1} denote the real parts of $FD(1)$ and $FD(N-1)$, I_1 and I_{N-1} denote the imaginary parts of $FD(1)$ and $FD(N-1)$, and θ_1 and θ_{N-1} are the phases of $FD(1)$ and $FD(N-1)$.

Moreover, the lengths between each pair of points in \mathbf{Z} are defined as PPL. *PPL* is suitable for describing shape details. To calculate PPL is time consuming due

to that each point is considered as a start point. Equations (3-8) and (3-9) describe the calculating processes of PPL.

$$PPL(k) = \{l_i\} = \{\|\tilde{z}(i) - \tilde{z}(i-1)\|\} = \{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}\} \quad (3-8)$$

$$\tilde{z}(k) = \tilde{z}(N+k) \quad (3-9)$$

where

$$1 \leq k \leq N, k \leq i \leq N+k$$

3.3.2 Color Features

Numerous features, such as edge, corner, texture, color and shape, have been utilized to extract useful information from an image. Among these features, color involves the intuitive information to represent the conceptual idea of an image. Therefore, pixel color and pixel position are utilized in this work to extract the conceptual idea of an image. The color space used in this work is RGB color space, a format common to most video devices. To enhance the regional information of an image, the position (x, y) feature is combined with RGB color information as the feature vector. That is, each pixel contains a 5D feature vector (R, G, B, x, y) , which is shown in Fig. 3-4.

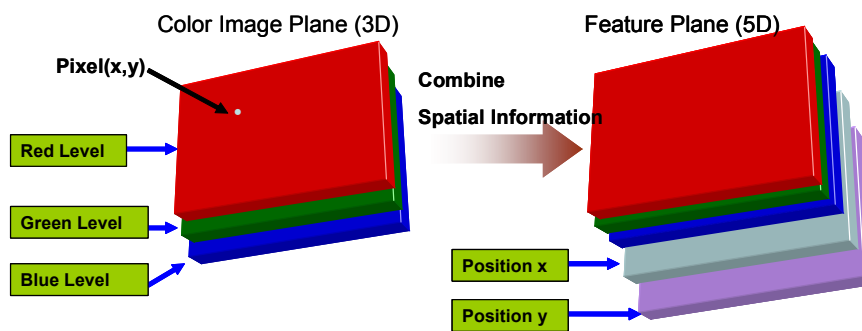


Figure 3-4 5D feature vector construction.

This work applies Gaussian mixture model (GMM) to model region information in a scene image as a blob model, which is defined as BM, using 5D feature vectors (R, G, B, x , y). We assume that the density function of color and position features have Gaussian distributions. First, each pixel x is defined as a 5-dimensional vector at time t . Moreover, N Gaussian distributions are used to construct the GMM, which is described in Eq. (3-10).

$$f(x | \lambda) = \sum_{i=1}^N w_i \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right) \quad (3-10)$$

λ represents the parameters of GMM,

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, 2, \dots, N \text{ and } \sum_{i=1}^N w_i = 1$$

Next, parameters λ of GMM are calculated to enable the GMM to match the feature vector distribution with least errors. The most common method for calculating parameters λ is ML estimation. The objective of ML estimation is to identify model parameters by maximizing the likelihood function of GMM obtained from training feature vectors X . The ML parameters are derived iteratively using the EM algorithm. Supposing there are s feature vectors x_1, x_2, \dots, x_s (In this work, s is defined as image size, $320 \times 240 = 76,800$), then the ML estimation of λ can be calculated using Eq. (3-11).

$$\lambda_{ML} = \arg \max_{\lambda} \sum_{j=1}^s \log f(x_j | \lambda) \quad (3-11)$$

Furthermore, unsupervised data clustering is used before the EM algorithm iterations to accelerate convergence. This study uses the K-means algorithm [59] for clustering. The number of clusters is defined, and then the initial center of each cluster is obtained randomly. The appropriate center and variance of each cluster can be

estimated iteratively using the K-means algorithm and applied as the initial mean and variance of each Gaussian component of the GMM.

3.3.3 Similarity Functions

To determine the similarity between two objects when building databases and recognizing objects, a similarity measurement $D(U, V)$ is applied to extract features.

We assume that the features extracted from two contours are $U = \{u_0, \dots, u_i, \dots, u_{I-1}\}$

and $V = \{v_0, \dots, v_i, \dots, v_{I-1}\}$, respectively, where I denotes the feature size. Two

similarity measures are applied using 1-norm distance (Eq. (3-12)) and K-L distance

[71] (Eq. (3-13)), where c denotes the number of points on an extracted contour and

s denotes image size. In this work, c is defined as 256 and s is defined as 76800.

$$D_{1-norm}(U, V) = \sum_{i=0}^{I-1} |u_i - v_i|, I = c \quad (3-12)$$

$$D_{KL}(U, V) \approx \sum_{t=0}^{I-1} \left(p_1(t) \cdot \log\left(\frac{p_1(t)}{p(t)}\right) + p_0(t) \cdot \log\left(\frac{p_0(t)}{p(t)}\right) \right), I = s \quad (3-13)$$

$$\text{where } p_0(t) = \frac{u_t}{u_{sum}}, \quad p_1(t) = \frac{v_t}{v_{sum}}, \quad u_{sum} = \sum_{i=0}^{s-1} u_i, \quad v_{sum} = \sum_{i=0}^{s-1} v_i, \quad p(t) = \frac{p_0(t) + p_1(t)}{2}$$

3.3.4 Similarity Measures

Suppose V_{new}^n represents a new sampled view of the n^{th} object and C_m^n represents the m^{th} characteristic view of the n^{th} object. Moreover, $A_{m^{min}}$ denotes the aspects that have the minimum distance from V_{new}^n and $C_{m^{min}}^n$ represents the minimal distance, where m^{min} is the index of $A_{m^{min}}$. $C_{m^{min}-1}^n$ and $C_{m^{min}+1}^n$ denote the

neighboring views of $C_{m_{\min}}^n$. Let $d_M^1(V_{new}^n, C_m^n)$ (Eq. (3-14)) denotes the similarity measure using MAG and 1-norm distance, $d_M^2(V_{new}^n, C_m^n)$ (Eq. (3-15)) denotes the similarity measure using PPL and 1-norm distance, $d_M^3(V_{new}^n, C_m^n)$ (Eq. (3-16)) denotes the similarity measure using BM and K-L distance, and $d_a^1(V_{new}^n, C_m^n)$ (Eq.(3-17)) denotes the similarity measure using θ_z and 1-norm distance.

$$d_M^1(V_{new}^n, C_m^n) = \sum_{k=1}^{T_2} |MAG^{V_{new}^n}(k) - MAG^{C_m^n}(k)| + |MAG^{V_{new}^n}(N-k) - MAG^{C_m^n}(N-k)| \quad (3-14)$$

$$d_M^2(V_{new}^n, C_m^n) = \sum_{k=1}^N |PPL^{V_{new}^n}(k) - PPL^{C_m^n}(k)| \quad (3-15)$$

$$d_M^3(V_{new}^n, C_m^n) = \sum_{t=0}^{s-1} \left(p_1(t) \cdot \log\left(\frac{p_1(t)}{p(t)}\right) + p_0(t) \cdot \log\left(\frac{p_0(t)}{p(t)}\right) \right) \quad (3-16)$$

where $p_0(t) = \frac{u_t}{u_{sum}}$, $p_1(t) = \frac{v_t}{v_{sum}}$, $u_{sum} = \sum_{i=0}^{s-1} u_i$, $v_{sum} = \sum_{i=0}^{s-1} v_i$, $p(t) = \frac{p_0(t) + p_1(t)}{2}$

$$d_a^1(V_{new}^n, C_m^n) = |\theta_z^T(k) - \theta_z^D(k)| \quad (3-17)$$

3.4 Flexible 3D Object Recognition Framework

A flexible framework using the ISAG is described in this section. In the framework, a MOD is composed of one or more AODs. Each AOD is built using one main feature or using one main feature with one assistant feature. Moreover, each feature has its similarity function, such as Eqs. (3-14)-(3-17).

3.4.1 Generation of Aspects and Characteristic Views

The ISAG is a four-step procedure and is illustrated as Fig. 3-5. Step A-1 to A-4 is applied to extract aspects and characteristic views. Those aspects comprise an object database and the characteristic views are used for object matching with a new view

V_{new}^n .

Step A-1:

Initialize the number of aspects be zero. 2D views of the n^{th} object are randomly sampled from a viewing sphere and each 2D view is regarded as V_{new}^n .

Step A-2:

When the number of existing aspects of the n^{th} object equals zero, V_{new}^n is regarded as a characteristic view of a new aspect.

Step A-3:

When the number of existing aspects of the n^{th} object equals one or two,

(A-3.1) When Eqs. (3-18) and (3-19) are both satisfied, V_{new}^n is combined into the m^{\min} aspect, and the characteristic view of the m^{\min} aspect remains the same.

$$\min_{all C_m^n \in A_{m^{\min}}} d_M(V_{new}^n, C_m^n) < T_3 \quad (3-18)$$

$$\min_{all C_m^n \in A_{m^{\min}}} d_a(V_{new}^n, C_m^n) < T_5 \quad (3-19)$$

where T_3 and T_5 are both predefined threshold values.

(A-3.2) Otherwise, if Eq. (3-18) is satisfied and Eq. (3-19) is not, V_{new}^n is combined into the m^{\min} aspect, and is regarded as a new characteristic view of the m^{\min} aspect.

(A-3.3) Otherwise, if Eqs. (3-18) and (3-19) are both unsatisfied, a new aspect of the n^{th} object is established, and V_{new}^n is regarded as the new characteristic view of the new aspect.

Step A-4:

When the number of existing aspects of the n^{th} object is ≥ 3 ,

(A-4.1) If either Eq. (3-20) or Eq. (3-21) is true, a new aspect is constructed and

V_{new}^n is considered the characteristic view of the new aspect. When a new aspect is established, the aspect order can be determined to let similar aspects be close to each other using Eq. (3-22). If Eq. (3-22) is true, the similarity distance between V_{new}^n and $C_{m^{\min}+1}^n$ exceeds that between V_{new}^n and $C_{m^{\min}-1}^n$, then the new aspect is inserted between aspect m^{\min} and aspect $m^{\min}-1$; otherwise, the new aspect is inserted between aspects m^{\min} and $m^{\min}+1$.

$$\min_{all C_m^n \in A_{m^{\min}}} d_M(V_{new}^n, C_m^n) > T_4 \quad (3-20)$$

$$T_3 < \min_{all C_m^n \in A_{m^{\min}}} d_M(V_{new}^n, C_m^n) < T_4 \quad and \quad d_M(V_{new}^n, C_{m^{\min} \pm 1}^n) > T_4 \quad (3-21)$$

$$d_M(V_{new}^n, C_{m^{\min}+1}^n) > d_M(V_{new}^n, C_{m^{\min}-1}^n) \quad (3-22)$$

(A-4.2) Otherwise, if Eqs. (3-20) and (3-21) are both unsatisfied and Eq. (3-19) is

true, V_{new}^n is combined into the m^{\min} aspect and the characteristic view of the m^{\min} aspect remains the same.

(A-4.3) Otherwise, if Eqs. (3-20) and (3-21) are both unsatisfied and Eq. (3-19) is

not true, V_{new}^n is combined into the m^{\min} aspect and is regarded as a new characteristic view of the m^{\min} aspect.

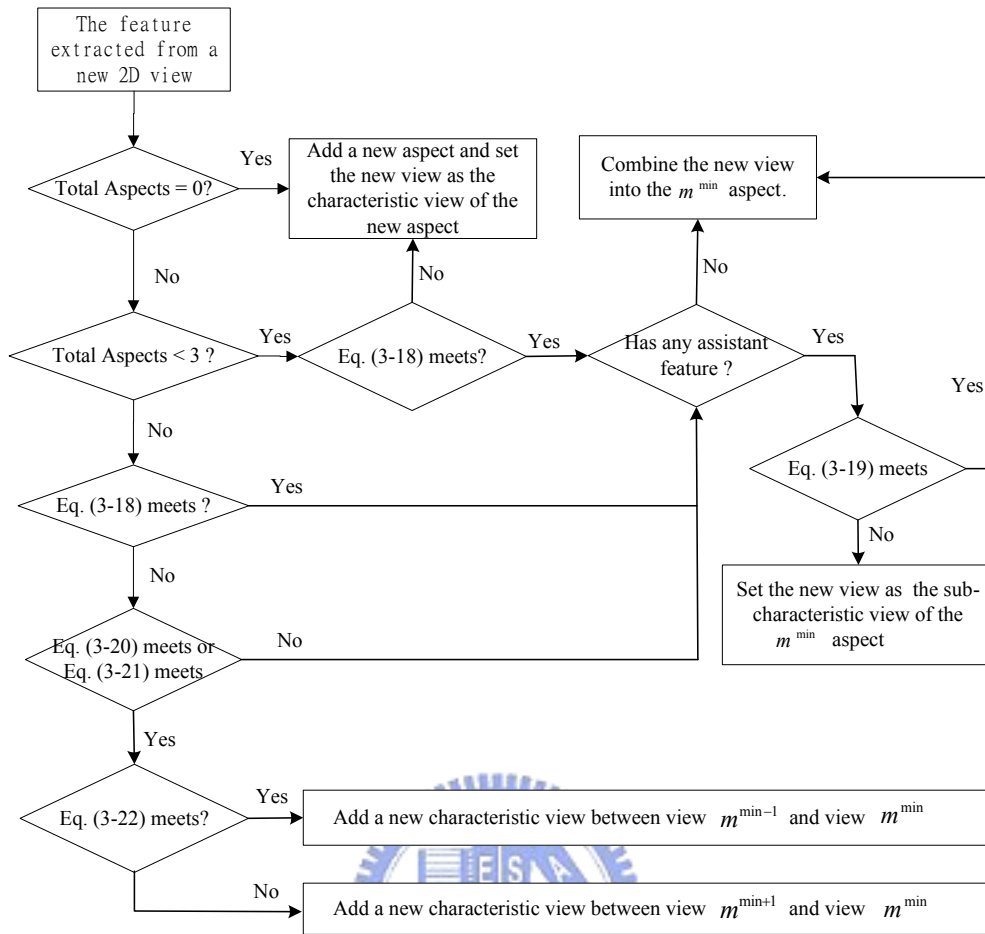


Figure 3-5 The procedure of the ISAG

Terms T_3 and T_4 are two predefined threshold values, where $T_4 \geq T_3$. The criterion for selecting T_3 and T_4 depends on the precise level for describing the object. If T_3 and T_4 are both small, then the criterion of combining 2D views becomes strict and, thus, the number of aspects increases. Furthermore, if the difference between T_3 and T_4 decreases, the tolerance for the difference between 2D views inside an aspect decreases, thereby the number of aspects decreases. Additionally, T_3 and T_4 should be initialized manually and modified iteratively until the final number of aspect reaches an acceptable number, which is determined

based on the degree of object symmetry. In this work, $T_3^{d_1^M}$ and $T_4^{d_1^M}$ are defined as T_3 and T_4 while adopting MAG as the feature; $T_3^{d_2^M}$ and $T_4^{d_2^M}$ are defined as T_3 and T_4 while adopting PPL as the feature; $T_3^{d_3^M}$ and $T_4^{d_3^M}$ are defined as the T_3 and T_4 while adopting BM as the feature, and $T_5^{d_4^M}$ is defined the T_5 whiling adopting θ_z as the feature. Section 4.2 presents the values of $T_3^{d_1^M}$, $T_3^{d_2^M}$, $T_3^{d_3^M}$, $T_4^{d_1^M}$, $T_4^{d_2^M}$, $T_4^{d_3^M}$, and $T_5^{d_4^M}$.

3.4.2 Object Recognition using 2D Characteristic Views

After constructing the aspect-graph representation of each object, a test view of an unknown object is recognized by matching itself with all the characteristic views of each AOD. If multiple AODs are utilized in the framework, a hierarchical matching process is applied calculate the final recognition results with a weighting combination of all similarity measures. Suppose the candidate objects in the k^{th} AOD are included in a set of N^k , and $n(N^k)$ denotes the number of candidate object in N^k . The number of candidate objects reduces after each object matching procedure, which is described in Eq. (3-23).

$$n(N^{k+1}) \leq n(N^k), k \geq 1 \quad (3-23)$$

In the k^{th} AOD, the main feature and the assistant feature of the test view are extracted to match with all the characteristic views. Suppose V_j^i denotes a test view of an unknown object, the object matching in the proposed framework is described as Eqs. (3-24.1) and (3-24.2).

$$d^k(V_j^i, C_m^n) = d_m^k(V_j^i, C_m^n) + \omega_1^k \cdot d_a^k(V_j^i, C_m^n), n \in N^k, k = 1 \quad (3-24.1)$$

$$d^k(V_j^i, C_m^n) = d_{\min}^{k-1}(n) + \omega_2^k \cdot (d_m^k(V_j^i, C_m^n) + \omega_1^k \cdot d_a^k(V_j^i, C_m^n)), n \in N^k, k \geq 2 \quad (3-24.2)$$

Let $d_m^k(V_j^i, C_m^n)$ and $d_a^k(V_j^i, C_m^{n^{(k)}})$ denotes the main and assistant similarity distances between the unknown object and C_m^n , where n denotes the n^{th} object in the set of candidate objects N^k . If the framework comprises only one AOD, the characteristic views of the first three smallest similarity distances in $d^1(V_j^i, C_m^n)$ (Eq.(3-24.1)) are regarded as the top-three matches. In Eq. (3-24.1), ω_1^k is a weighting parameter for combing different similarity measures. When no assistant feature is utilized, ω_1^k is set to zero. Otherwise, the objects included in the first half smallest similarity distances of $d^k(V_j^i, C_m^n)$ are defined as a set N^{k+1} . The objects in N^{k+1} are preserved for further recognition in the $(k+1)^{th}$ AOD.

If the framework comprises two or more AODs, the characteristic views of the first three smallest similarity distances in $d^k(V_j^i, C_m^n)$ (Eq. (3-24.2)) are regarded as the top-three matches. In Eq. (3-24.2), $d_{\min}^{k-1}(n)$ denotes the minimum similarity distance between the unknown object and the n^{th} candidate object in the $(k-1)^{th}$ database. Moreover, ω_2^k is a weighting parameter for combing the similarity measure between the k^{th} and $(k-1)^{th}$ AODs.

3.4.3 Applications

The proposed framework is evaluated on various object recognition problems, including 3D object recognition, human posture recognition, and scene recognition. Three assumptions are made for applying the proposed framework to the above three applications. First, different features are used in different applications with the

proposed framework. In this dissertation, the features described in Section 3.3 are employed in the above three applications to perform the efficiency of the proposed framework. Second, training images for extracting the aspect-graph of objects in different applications are randomly sampled from a viewing sphere of a robot platform. A Pan-Tilt-Zoom camera is set up in a fixed position in the robot platform. Next, the efficiency of the proposed framework is performed with the object database and testing images that belong to the same category with the object database. For example, 2D rigid-object testing images are tested with the rigid object database, and etc.

The similarity measures described in Section 3.3 are employed in the three applications. Three kinds of combing structures are performed with these similarity measures. In the 3D rigid object recognition, two AODs are utilized with two main features MAG and PPL. The weighting combination of the similarity measures is described in Eqs. (3-25) and (3-26).

$$d^1(V_j^i, C_m^n) = d_m^1(V_j^i, C_m^n), n \in N^1 \quad (3-25)$$

$$d^2(V_j^i, C_m^n) = d_{\min}^1(n) + \omega_2^2 \cdot (d_m^2(V_j^i, C_m^n)), n \in N^2 \quad (3-26)$$

Moreover, $d_m^1(V_j^i, C_m^n)$ is calculated with MAG using Eq. (3-14), and $d_m^2(V_j^i, C_m^n)$ is calculated with PPL using Eq. (3-15). Furthermore, the weighting parameters ω_1^1 and ω_1^2 are both set to zero and the weighting parameters ω_2^2 is defined as the Eq. (3-27). $T_4^{d_1^1}$ and $T_4^{d_1^2}$ are the threshold values applied on the ISAG, and are defined in Section 4.1.

$$\omega_2^2 = T_4^{d_1^2} / T_4^{d_1^1} \quad (3-27)$$

In the human posture recognition, only one AOD is utilized with one main feature MAG and one assistant feature θ_z . The weighting combination of the similarity measures is described in Eq (3-28).

$$d^1(V_j^i, C_m^n) = d_m^1(V_j^i, C_m^n) + \omega_1^1 \cdot d_a^1(V_j^i, C_m^n), n \in N^1 \quad (3-28)$$

In Eq. (3-28), $d_m^1(V_j^i, C_m^n)$ is calculated using Eq. (3-14) and $d_a^1(V_j^i, C_m^n)$ is calculated using Eq. (3-17). Furthermore, the weighting parameter ω_1^1 is defined as Eq. (3-29), where $T_5^{d_a^1}$ is the threshold values applied on the ISAG, and is defined in Section 4.1.

$$\omega = 1/T_5^{d_a^1} \quad (3-29)$$

In the scene recognition, only one AOD is utilized with one main feature BM. The weighting combination of the similarity measures is described in Eq (3-30).

$$d^1(V_j^i, C_m^n) = d_m^1(V_j^i, C_m^n), n \in N^1 \quad (3-30)$$

In Eq. (3-30), $d_m^1(V_j^i, C_m^n)$ is calculated using Eq. (3-16). Furthermore, the weighting parameter ω_1^1 is defined as zero.

Chapter 4

Experimental Results

The chapter provides experimental results to assess the efficiency of the proposed 3D object recognition system. In Section 4.1, five experiments are performed to test the robustness of the BSHSR with a complex background in an indoor environment. After that, the BSHSR is applied to extract foreground regions for building a 3D object database using the ISAG and testing the performance of the proposed 3D object recognition system. Three object recognition problems, namely rigid object recognition, human posture recognition, and scene recognition, are performed with the proposed method in Section 4.2.

4.1 BSHSR

The video data for experiments was obtained using a SONY DVI-D30 PTZ camera in an indoor environment. Morphological filter was applied to remove noise and the camera controls were set to automatic mode. The same threshold values were used for all experiments. The values of the important threshold values were $N_G = 15$,

$\alpha = 0.002$, $P_B = 0.1$, $B_0 = 0.7$, $B_1 = 300$ and $B_2 = 0.8$. Meanwhile, the computational speed was around five frames per second on a P4 2.8GHz PC, while the video had a frame size of 320 x 240.

4.1.1 Local Illumination Changes

The first experiment was performed to test the robustness of the proposed method about the local illumination changes. Local illumination changes resulting from desk lights occur constantly in indoor environments. Desk lights are usually white or yellow. Two video clips containing several changes of desk light are collected to simulate local illumination changes. Figure 4-1(a) shows 15 representative samples of the first one video clip. Meanwhile, Fig. 4-1(b) shows the classified result of the foreground pixel using the proposed method, the CBM and CSIM, where red indicates shadow, green indicates highlight and blue indicates foreground. Figure 4-1(c) displays the result of the result of final background subtraction to demonstrate the robustness of the proposed method, where the white and black color represents the foreground and background pixels respectively. The image sequences comprise different levels of illumination changes. The desk light was turned on at the 476th frame and its brightness increased until the 1000th frame. The overall picture becomes the foreground regions of the corresponding frames in Fig. 4-1(b) owing to the lack of such information in the CBM. However, the final result of background subtraction of the corresponding frames in Fig. 4-1(c) is still good owing to the proposed scheme combining the CBM, CSIM and GBM. The desk light was then turned off at the 1030th frame, and became darker until the 1300th frame. The original Gaussian distribution in the ECBM became the component in the CCBM, and a new representative Gaussian distribution in the ECBM is constructed for that a new

background information is involved from the new collected frames between the 476th and the 1000th frame are more than the initial collected 300 frames. Consequently, the 1300th frame in Fig. 4-1(b) has many foreground regions. However the final result of the 1300th frame is still good. The illumination changes are all modeled into the LTCBM when the background model records the background changes. The area of the red, blue and green regions reduces after the 1300th frame.

Table 4-1 compares the proposed scheme with the method proposed by Hoprasert [60]. Comparison criteria are identified by labeling the foreground regions of a frame manually. The CSIM can be constructed based on the appropriate representative Gaussian distribution chosen from the LTCBM and STCBM. The ability to handle illumination variation and the accuracy of the background subtraction are improved and the results are shown in Table 4-1.

Table 4-1 The robustness test between the proposed method and that proposed by Hoprasert [60] via local illumination changes with a yellow desk light

FRAME		476		480		500		580		650	
PROPOSED (%)	HOPRASERT [60] (%)	100.00	94.05	99.84	36.40	99.93	22.50	99.91	15.38	83.96	23.42
FRAME		750		900		1000		1030		1120	
PROPOSED (%)	HOPRASERT [60] (%)	91.50	31.51	93.10	30.91	95.44	34.26	97.75	38.28	99.15	32.90
FRAME		1150		1300		1330		1400		1600	
PROPOSED (%)	HOPRASERT [60] (%)	93.79	50.72	99.95	99.84	93.31	92.40	96.22	13.03	99.30	34.66

*: The value in the table means the recognition rate that correct background pixels in a frame divide total pixels in a frame(%)

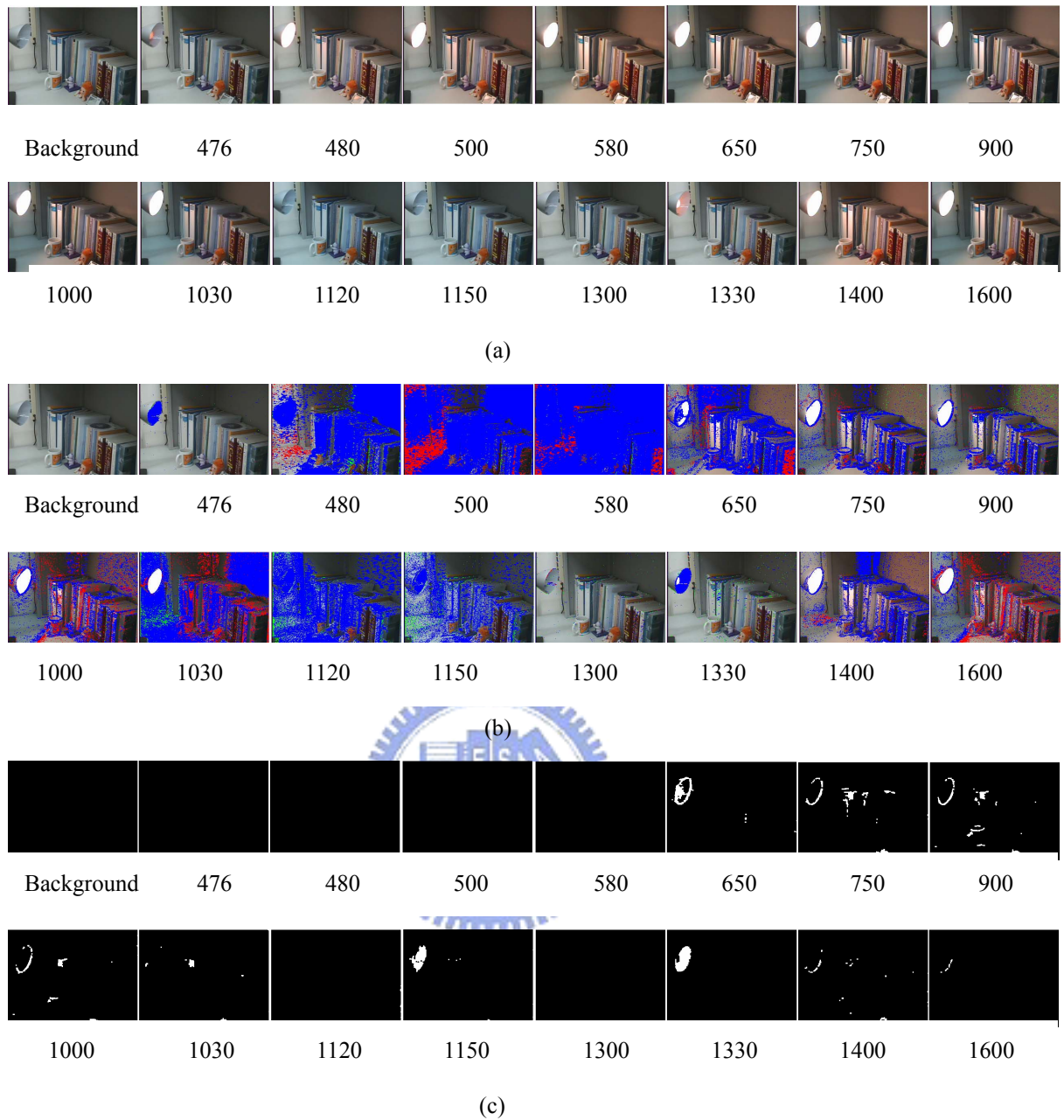


Figure 4-1 The results of illumination changes with a yellow desk light, the number below the picture is the index of frame. (a) Original images. (b) The results of pixel classification, where red indicates the shadow, green indicates the highlight and blue indicates the foreground. (c) The results of background subtraction with shadow removal using the proposed method, where dark indicates the background and white indicates the foreground.

Figure 4-2(a) shows a similar image sequence to that on Fig. 4-1(a). The two sequences differ only in the color of the desk light. The desk light was turned on at the 660th frame and the same brightness was maintained until the 950th frame. The desk light was then turned off at the 1006th frame and turned on again at the 1180th frame. The results of shadows and highlights removal are shown in Fig. 4-2(b) and the results of final background subtraction are shown in Fig. 4-2(c). The results of background subtraction in Fig. 4-2 and the comparison result in Table 4-2 are shown to demonstrate the robustness of the proposed scheme.

Table 4-2 The robustness test between the proposed method and that proposed by Hoprasert [60] via local illumination changes with a white desk light

Frame		660		665		670		860		950	
Proposed (%)	Hoprasert[60] (%)	99.02	99.48	97.93	79.81	95.92	92.22	96.73	93.81	97.44	94.46
Frame		1006		1020		1150		1180		1250	
Proposed (%)	Hoprasert[60] (%)	98.12	95.65	99.94	98.85	99.78	99.68	98.94	99.08	97.28	93.81
Frame		1300		1375		1377		1380		1445	
Proposed (%)	Hoprasert[60] (%)	97.49	95.26	97.73	87.50	98.83	98.92	99.73	99.32	100.00	99.71

*: The value in the table means the recognition rate that correct background pixels in a frame divide total pixels in a frame(%).

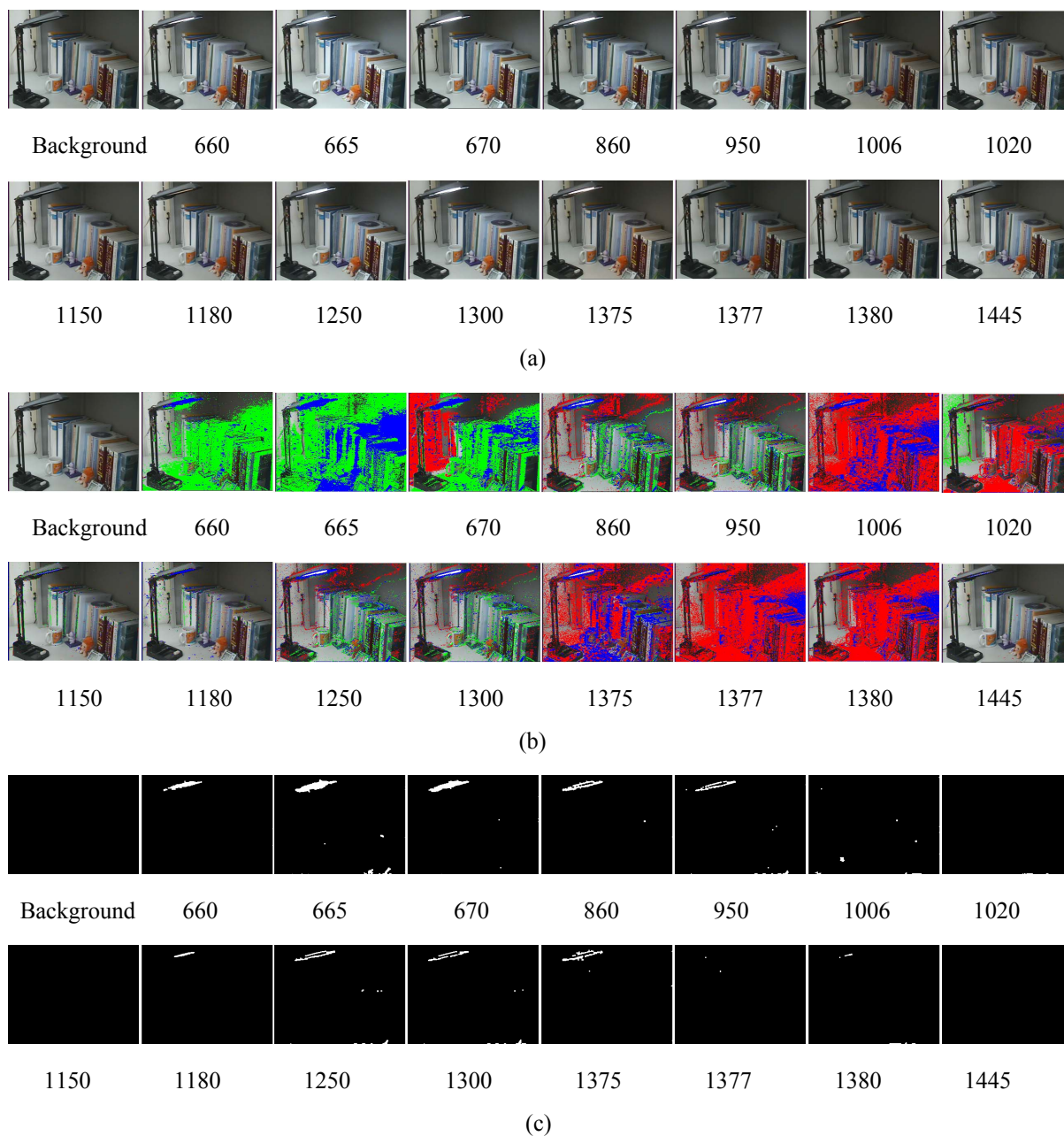


Figure 4-2 The results of illumination changes with white desk light, the number below the picture is the index of frame. (a) Original images, where red indicates the shadow, green indicates the highlight and blue indicates the foreground. (b) The results of pixel classification. (c) The results of background subtraction with shadow removal using our proposed method, where dark indicates the background and white indicates the foreground.

4.1.2 Global Illumination Changes

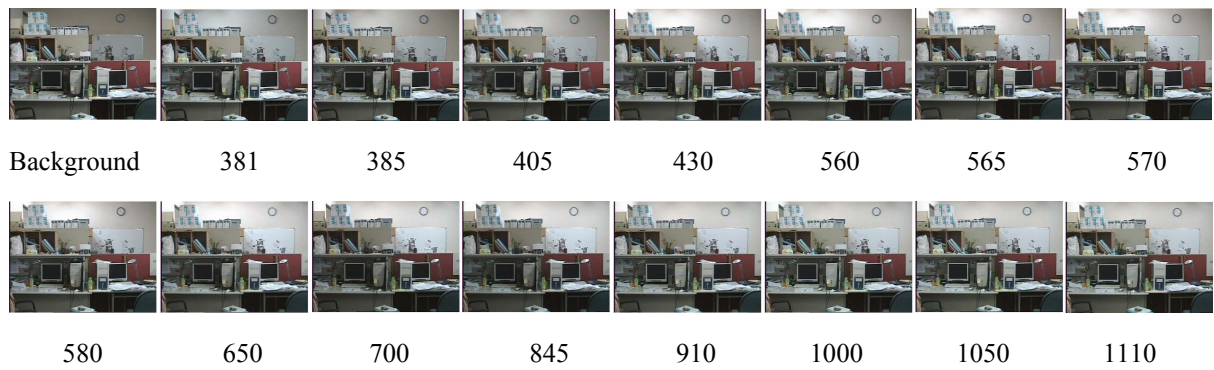
The second experiment was performed to test the robustness of the proposed method in terms of global illumination changes. The image sequences consist of illumination changes where a fluorescent lamp was turned on at the 381th frame and more lamps were turned on at the 430th frame. The illumination changes are then modeled into the LTCBM when the proposed background model recorded the background changes. Notably the area of the red, blue and green regions decreases at the 580th frame. When the third daylight lamp is switched on in the 650th frame, it is clear that fewer blue regions appear at the 845th frame owing to illumination changes having been modeled in the LTCBM. However, the final results of background subtraction shown in Fig. 4-3(c) are all better than those of pure color-based background subtraction shown in Fig. 4-3(b). Table 4-3 shows the comparison results between the proposed scheme and that proposed by Hoprasert [60]. The comparison demonstrates that the proposed scheme is robust to global illumination changes.

Table 4-3 The comparison between the proposed method and that proposed by Hoprasert [60] via global illumination changes with fluorescent lamps

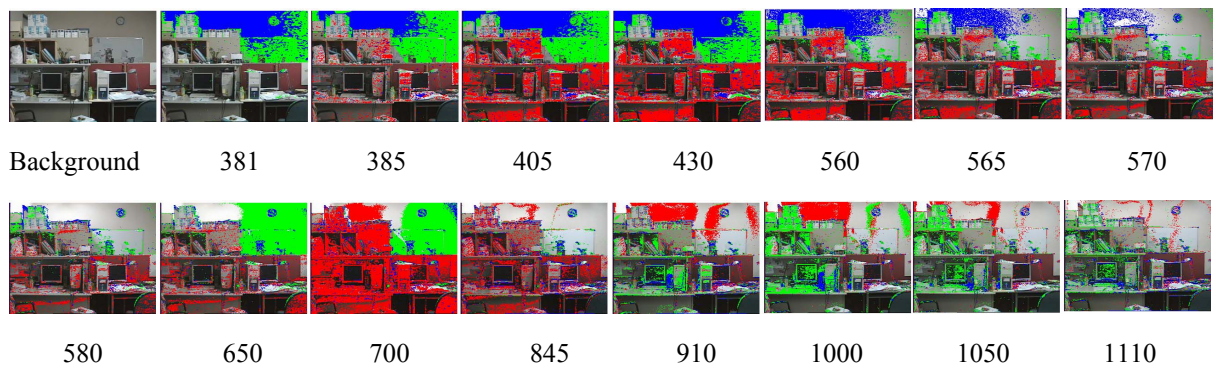
Frame		381 (1 ^{**})		385 (1 ^{**})		405 (1 ^{**})		430 (2 ^{**})		560 (2 ^{**})	
Proposed (%)	Hoprasert[60] (%)	98.24	93.54	88.35	82.14	83.85	78.24	56.50	68.42	66.85	69.82
Frame		565 (2 ^{**})		570 (2 ^{**})		580 (2 ^{**})		650 (3 ^{**})		700 (3 ^{**})	
Proposed (%)	Hoprasert[60] (%)	79.87	69.30	96.88	69.69	99.08	69.55	99.23	45.62	99.49	46.22
Frame		845 (3 ^{**})		910 (3 ^{**})		1000 (3 ^{**})		1050 (3 ^{**})		1110 (3 ^{**})	
Proposed (%)	Hoprasert[60] (%)	99.56	46.18	99.39	53.58	99.85	57.87	99.93	60.83	99.64	60.32

*: The value in the table means the recognition rate that correct background pixels in a frame divide total pixels in a frame(%).

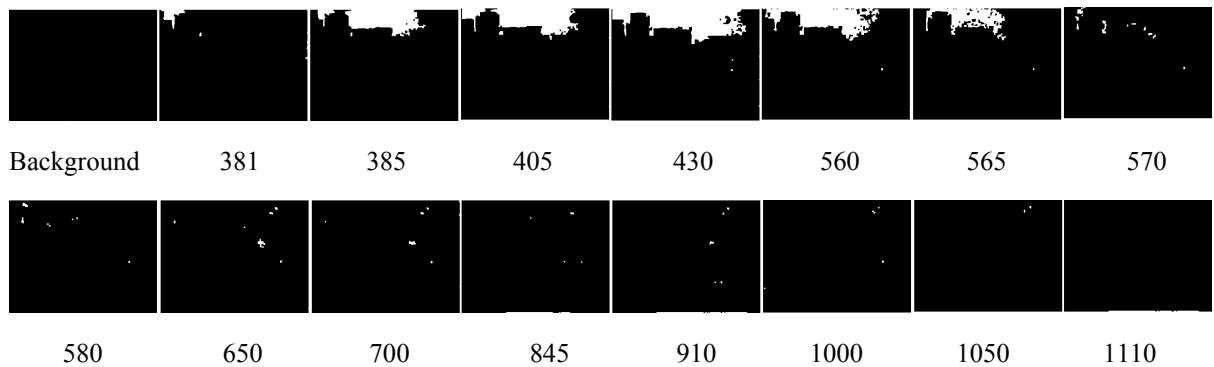
** : The number inside the parentheses indicates the number of fluorescent lamps that have turned on.



(a)



(b)



(c)

Figure 4-3 The results of global illumination changes with fluorescent lamps, the number below the picture is the index of frame. (a) Original images. (b) The results of pixel classification, where red indicates the shadow, green indicates the highlight and blue indicates the foreground. (c) The results of background subtraction with shadow removal using our proposed method, where dark indicates the background and white indicates the foreground.

4.1.3 Foreground Detection

In the third experiment (Fig. 4-4), a person goes into the monitoring area, and the foreground region can be effectively extracted regardless of the influence of shadow and highlight in the indoor environment. Owing to the captured video clip having little illumination variation and dynamic background variation, the comparison of the recognition rate of final background subtraction between the proposed method and that of Hoprasert [60] reveals that both methods are about the same, as listed in Table 4-4.

Table 4-4 The comparison between the proposed method and that proposed by Hoprasert [60] via foreground detection

Frame		380		450		530		590		620	
Proposed (%)	Hoprasert[60] (%)	90.45	89.18	86.50	85.80	89.38	88.87	88.45	87.72	88.67	88.76
Frame		680		700		735		755		840	
Proposed (%)	Hoprasert[60] (%)	91.07	90.62	85.63	85.15	82.76	80.71	92.44	92.46	100.00	99.61

*: The value in the table means the recognition rate that correct background pixels in a frame divide total pixels in a frame(%).

4.1.4 Dynamic Background

In the fourth experiment (Fig. 4-5), image sequences consist of swaying clothes hung on a frame. The proposed method gradually recognizes the clothes as background owing to the ability of LTCBM to record the history of background changes. In situations involving large variation of dynamic background, a representative initial color-based background model can be established by using more training frames to handle the variations.



Background 380 450 530 590 620

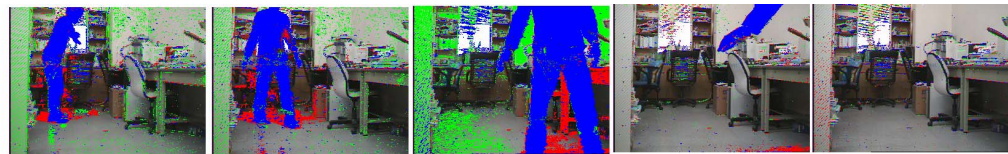


680 700 735 755 840

(a)



Background 380 450 530 590 620



680 700 735 755 840

(b)



Background 380 450 530 590 620



680 700 735 755 840

(c)

Figure 4-4 The results of foreground detection. (a) Original images. (b) The results of pixel classification, where the red color means the shadow, where red indicates the shadow, green indicates the highlight and blue indicates the foreground. (c) The results of background subtraction with shadow removal using our proposed method, where dark indicates the background and white indicates the foreground.



Background 500 540 580 620 660 700 740



780 820 860 900 940 980 1020 1060

(a)



Background 500 540 580 620 660 700 740



780 820 860 900 940 980 1020 1060

(b)



Background 500 540 580 620 660 700 740



780 820 860 900 940 980 1020 1060

(c)

Figure 4-5 The results of background subtraction about dynamic background. (a) Original images. (b) The results of pixel classification, where the red color means the shadow, where red indicates the shadow, green indicates the highlight and blue indicates the foreground. (c) The results of background subtraction with shadow removal using our proposed method, where dark indicates the background and white indicates the foreground.

4.1.5 Short-Term Color-based Background Model (STCBM)

The final experiment (Fig. 4-6) shows the advantage of adding the STCBM. A doll is placed on the desk at the 360th frame. Initially, it is regarded as foreground, and at the 560th frame, the foreground region becomes background owing to the LTCBM. However, the Gaussian component belonging to the doll still does not have the highest weighting. Without adding the STCBM, when a hand is placed above the doll at the 590th frame, the foreground regions at the 670th frame remain the same as those at the 590th frame, as shown in Fig. 4-6(b). The foreground regions under our hand become shadows at the 670th frame in Fig. 4-6(c) for that shadows and highlights removal works well using a representative Gaussian component based on the STCBM. This experiment demonstrates the efficiency of the STCBM that a representative Gaussian component of the CBM can be selected by giving consideration to long-term tendency and short-term tendency. Besides, the advantage of the STCBM helps to reduce the computing time used in the GBM and increase the recognition rate of foreground detection.

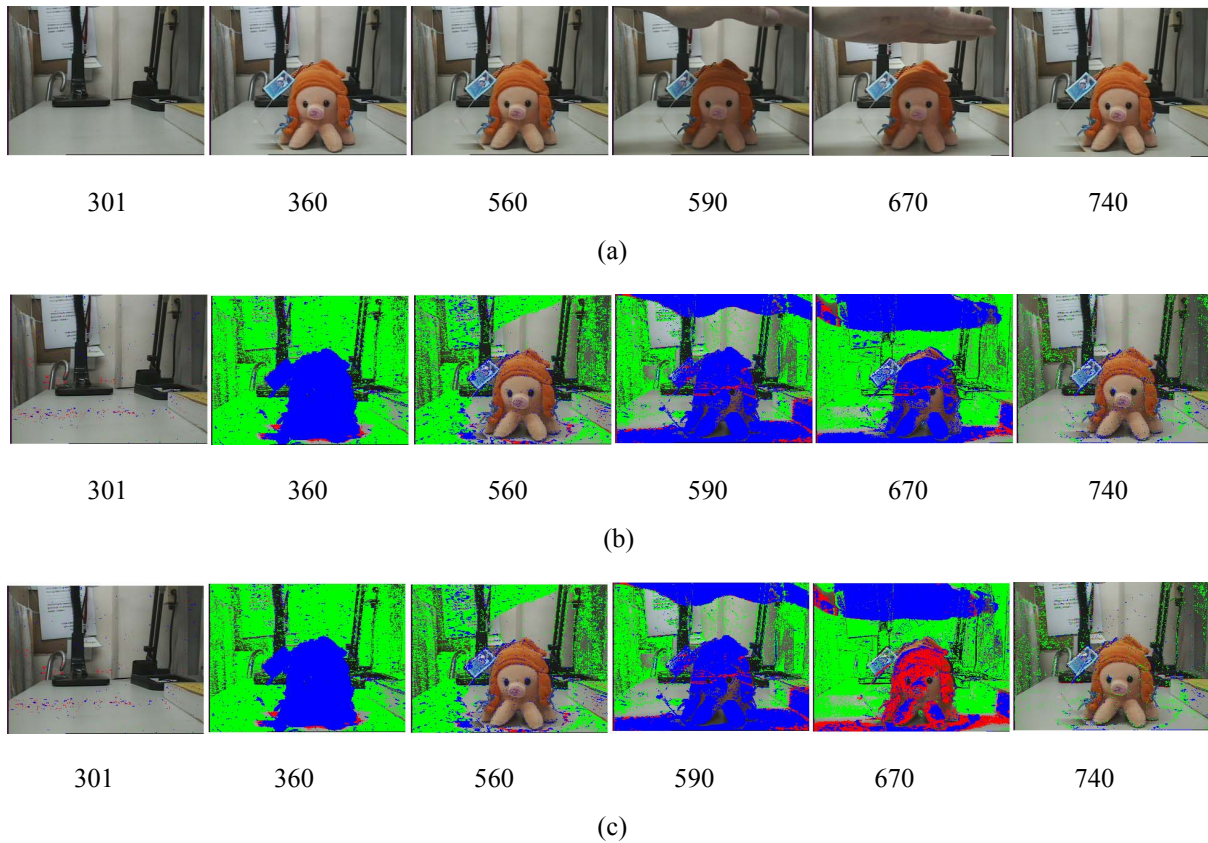


Figure 4-6 The results of the advantage of the STCBM, where the red color means the shadow, the green color means the highlight and the blue color means the foreground. (a) Original images. (b) The results of background subtraction without the STCBM. (c) The results of background subtraction with the STCBM, where red indicates the shadow, green indicates the highlight and blue indicates the foreground.

4.2 3D Object Recognition

This section describes several experiments demonstrating the effectiveness of the proposed 3D object recognition. A SONY EVI-D30 PTZ camera was employed to capture object views. The following three databases were built to test the proposed method: Fig. 4-6 contains 12 3D rigid objects, Fig. 4-7 contains six 3D human postures, and Fig. 4-8 contains 11 scenes.



Figure 4-7 The first database containing 12 3D rigid objects.

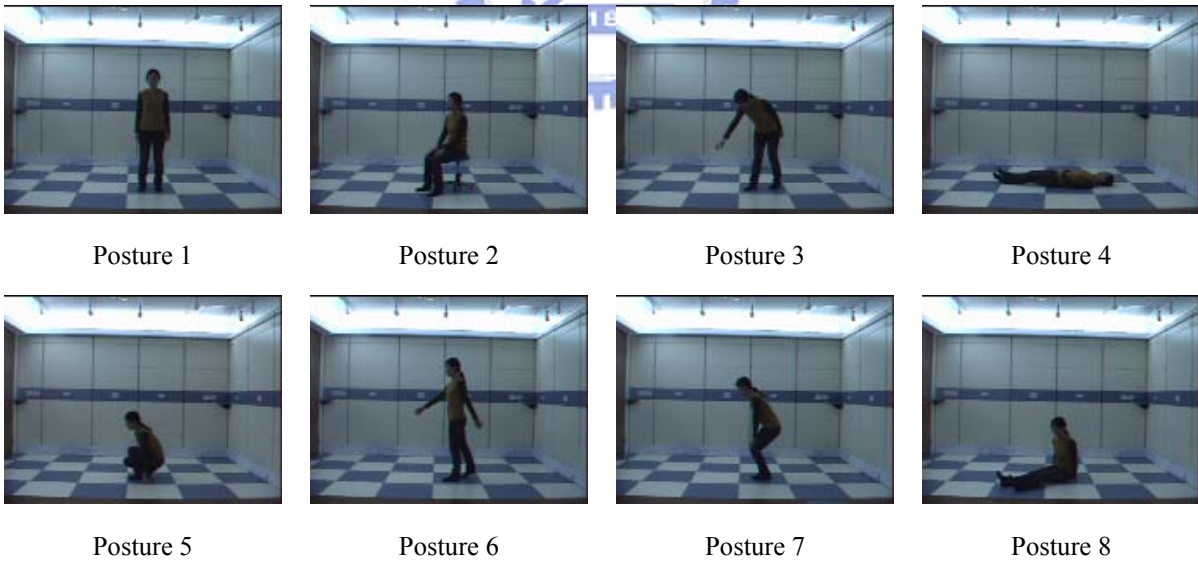


Figure 4-8 The second image database containing eight 3D human postures.



Figure 4-9 The third image database containing 11 scenes.

The notation $V_d^{1,j}$ and $V_d^{2,j}$ denote the sets of training views captured at 5° intervals, where $V_d^{1,j}$ is employed during rigid object recognition, and $V_d^{2,j}$ is employed during human posture recognition. The notation $V_d^{3,j}$, which denotes the set of training views captured at each location at a 1° increment, is utilized during scene recognition. Moreover, $V_t^{1,j}$ and $V_t^{2,j}$ denote the set of testing views captured from trisection points between each pair of points separated by 5° , where $V_t^{1,j}$ is utilized during rigid object recognition, and $V_t^{2,j}$ is utilized during human posture recognition. Moreover, $V_t^{3,j}$ denotes the set of testing views captured at locations away from the original locations in four directions (forward, backward, left and right), five distances (5cm, 10cm, 15cm, 20cm and 50cm) and five covering rates (5%, 10%, 15%, 20% and 50%). $V_t^{3,j}$ is utilized during scene recognition. The descriptions of the captured views are given by Eqs. (4-1)-(4-6).

$$\mathbf{V}_d^{1,j} = \{V_d^{1,j}(i)\}, \text{ where } 1 \leq j \leq 12, 1 \leq i \leq 72 \quad (4-1)$$

$$\mathbf{V}_d^{2,j} = \{V_d^{2,j}(i)\}, \text{ where } 1 \leq j \leq 8, 1 \leq i \leq 72 \quad (4-2)$$

$$\mathbf{V}_d^{3,j} = \{V_d^{3,j}(i)\}, \text{ where } 1 \leq j \leq 11, 1 \leq i \leq 61 \quad (4-3)$$

$$\mathbf{V}_t^{1,j} = \{V_t^{1,j}(i)\}, \text{ where } 1 \leq j \leq 12, 1 \leq i \leq 216 \quad (4-4)$$

$$\mathbf{V}_t^{2,j} = \{V_t^{2,j}(i)\}, \text{ where } 1 \leq j \leq 8, 1 \leq i \leq 216 \quad (4-5)$$

$$\mathbf{V}_t^{3,j} = \{V_t^{3,j}(i)\}, \text{ where } 1 \leq j \leq 11, 1 \leq i \leq 6100 \quad (4-6)$$

In the following experiments, T_0 denotes the number of objects, and is 12 for the rigid object recognition, 8 during human posture recognition, and 11 during scene recognition; T_1 denotes the number of training views, and is 72 during rigid object recognition and human posture recognition, and 61 during scene recognition. T_2 denotes the number of low frequency information in FD, and is 40 in the following experiment. Moreover, the threshold values used in the ISAG is listed as Table 4-5.

Table 4-5 The threshold values for the ISAG

Experiment	The first AOD			The second AOD		
	Main Feature		Assistant Feature	Main Feature		Assistant Feature
	T_3	T_4	T_5	T_3	T_4	T_5
3D object recognition	$T_3^{d_M^1} = 640$	$T_4^{d_M^1} = 1.25 * T_3^{d_M^1}$	N/A	$T_3^{d_M^2} = 336$	$T_4^{d_M^2} = 1.25 * T_3^{d_M^2}$	N/A
Human posture recognition	$T_3^{d_M^1} = 1450$	$T_4^{d_M^1} = 1.25 * T_3^{d_M^1}$	$T_5^{d_a^1} = 10$	N/A		
Scene recognition	$T_3^{d_M^3} = 1100$	$T_4^{d_M^3} = 1.25 * T_3^{d_M^3}$	N/A	N/A		

Computing time for calculating similarity between a test view and a view in the database was approximately 0.006 seconds for rigid object recognition, 0.004 seconds for human posture recognition and 0.01 seconds for scene recognition on a P4 3.2G CPU with 1GB RAM.

4.2.1 Rigid Object Recognition

In the first experiment, the efficiency of the proposed framework was assessed using 2-D views captured at random intervals with the first database (Fig. 4-7). To determine average performance of the proposed method, training views were generated by sampling views in $V_d^{1,j}$ in 200 different random orders. Background subtraction was first performed on training 2D views to extract foreground objects. After that, Canny edge detection and GVF were performed on the extracted foreground objects to extract the object contour. Two features, called the MAG and PPL, are then extracted from the object contour and be used for building the AODs with the ISAG (Fig. 3-2). The characteristic views of aspects in each AOD are utilized for object matching. A recognition result is calculated with a weighted combination of the similarity measures from both AODs. Figure 4-10 illustrates the system architecture of the proposed framework for the 3D rigid object recognition.

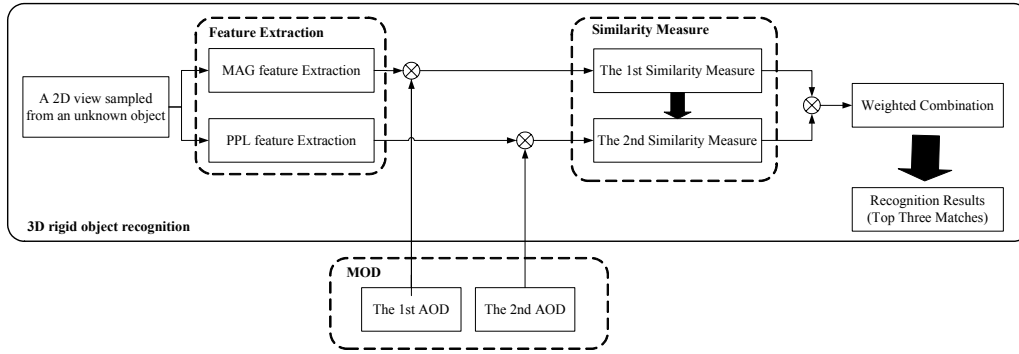


Figure 4-10 The system architecture of the proposed framework applied on the first experiment (3D rigid object recognition).

Table 4-6 The result of rigid object recognition using 2D views via MAG and PPL

Recognition Results	The index of the objects in the first database listed in Fig. 4-7												Avg.
	1	2	3	4	5	6	7	8	9	10	11	12	
Numbers of aspect of MAG	34.66	3.84	27.83	24.75	6.87	9.47	2.04	25.62	17.14	16.16	16.62	28.75	17.81
Numbers of aspect PPL	38.72	14.08	14.32	22.84	10.98	20.12	8.41	31.07	25.79	17.68	23.61	19.88	20.63
Top 1 Match (%)	98.25	99.97	97.71	97.39	100	99.81	99.79	99.35	99.90	97.97	98.44	96.83	98.78
Top 2 Match (%)	99.21	100	98.96	98.73	100	99.96	99.86	99.67	99.97	98.68	99.47	98.17	99.39
Top 3 Match (%)	99.61	100	99.39	99.34	100	99.98	99.89	99.78	99.99	98.98	99.77	98.64	99.62

Table 4-6 presents statistical information for the means of aspect numbers using MAG and PPL. Furthermore, symmetrical objects, such as objects 2, 5, 6 and 7, had few aspects, thereby reducing computing time for recognizing objects. The views in $V_r^{1,j}$ were adopted as unknowns, and tested whenever aspect-graph representations were built each time (200 times). The proposed aspect-graph generation is efficient due to its high recognition rate in the Top 1 to Top 3 matches in the Table 4-7.

The proposed method, which constructs an aspect-graph representation using sampled views at random intervals, generates a practicable updating mechanism that integrates the database using new collected views. In this experiment, 18 random views sampled from $V_d^{1,j}$ are first utilized to construct a coarse aspect-graph

representation of each object, called D_{18} . Eighteen additional random views are then adopted from the remaining views in $V_d^{1,j}$ to increase the accuracy of the database D_{18} , called D_{36} . Similarly, D_{54} and D_{72} are constructed using views in remaining $V_d^{1,j}$. Additionally, D_{90} and D_{108} are further constructed with extra random views sampled from $V_t^{1,j}$. Table 4-7 presents the average aspect numbers for each rigid object from 200 iterations. Although the aspect numbers increase when new views are employed to update the coarse database, the number of stored views remains significantly smaller than the number of original views. Figure 4-11 presents the recognition rate results obtained when using coarse to fine databases. Figure 4-12 presents the standard deviations for recognition rates. The recognition rate increases when aspect-graph representations are trained using additional object views. Moreover, stability increases based on decreasing standard deviation. Therefore, the proposed method is demonstrated as effective for updating aspect-graph representations without re-sorting the overall collected views, or re-calculating overall similarity measures.

Table 4-7 Results for numbers of aspects using MAG and PPL after updating with additional training views

Numbers of aspect	The index of the objects in the first database listed in Figure 4-7											
	1	2	3	4	5	6	7	8	9	10	11	12
D_{18}	14.11	3.40	11.98	10.13	5.32	6.36	1.58	12.64	8.80	8.43	8.73	11.10
D_{36}	22.86	3.60	18.83	16.15	6.23	8.01	1.80	19.16	12.53	12.17	12.48	18.29
D_{54}	29.52	3.74	23.95	20.96	6.65	8.94	1.92	23.24	15.20	14.53	15.06	24.05
D_{72}	34.66	3.84	27.83	24.75	6.87	9.47	2.04	25.62	17.14	16.16	16.62	28.75
D_{90}	39.28	3.99	28.68	25.99	7.14	9.83	2.14	27.32	18.04	17.37	17.86	30.99
D_{108}	43.28	4.07	29.50	27.14	7.36	10.12	2.26	28.67	18.90	18.50	19.06	33.12

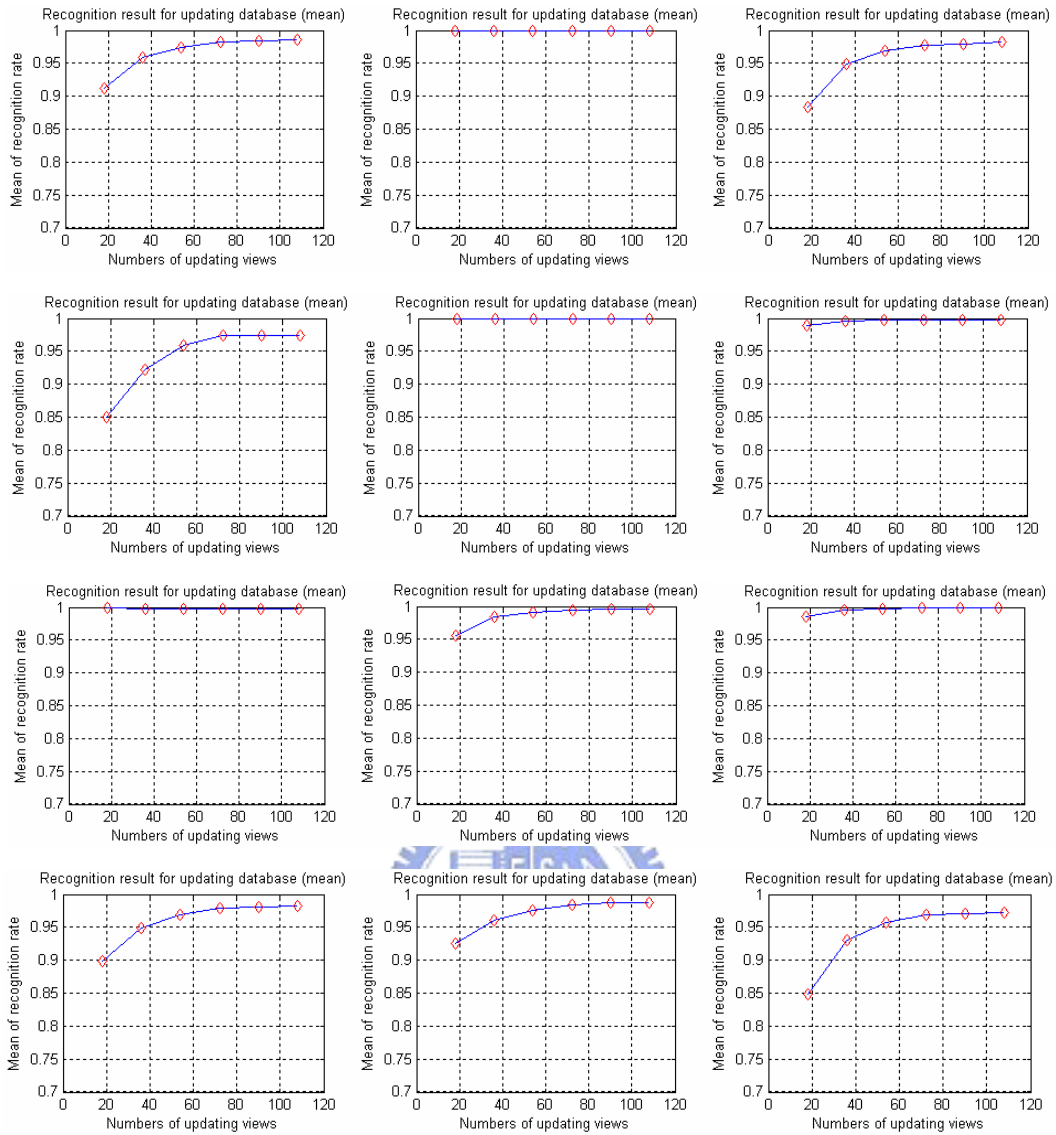


Figure 4-11 Recognition rates of coarse and fine databases ($D_{18}, D_{36}, D_{54}, D_{72}, D_{90}$ and D_{108}), calculated using 200 results.

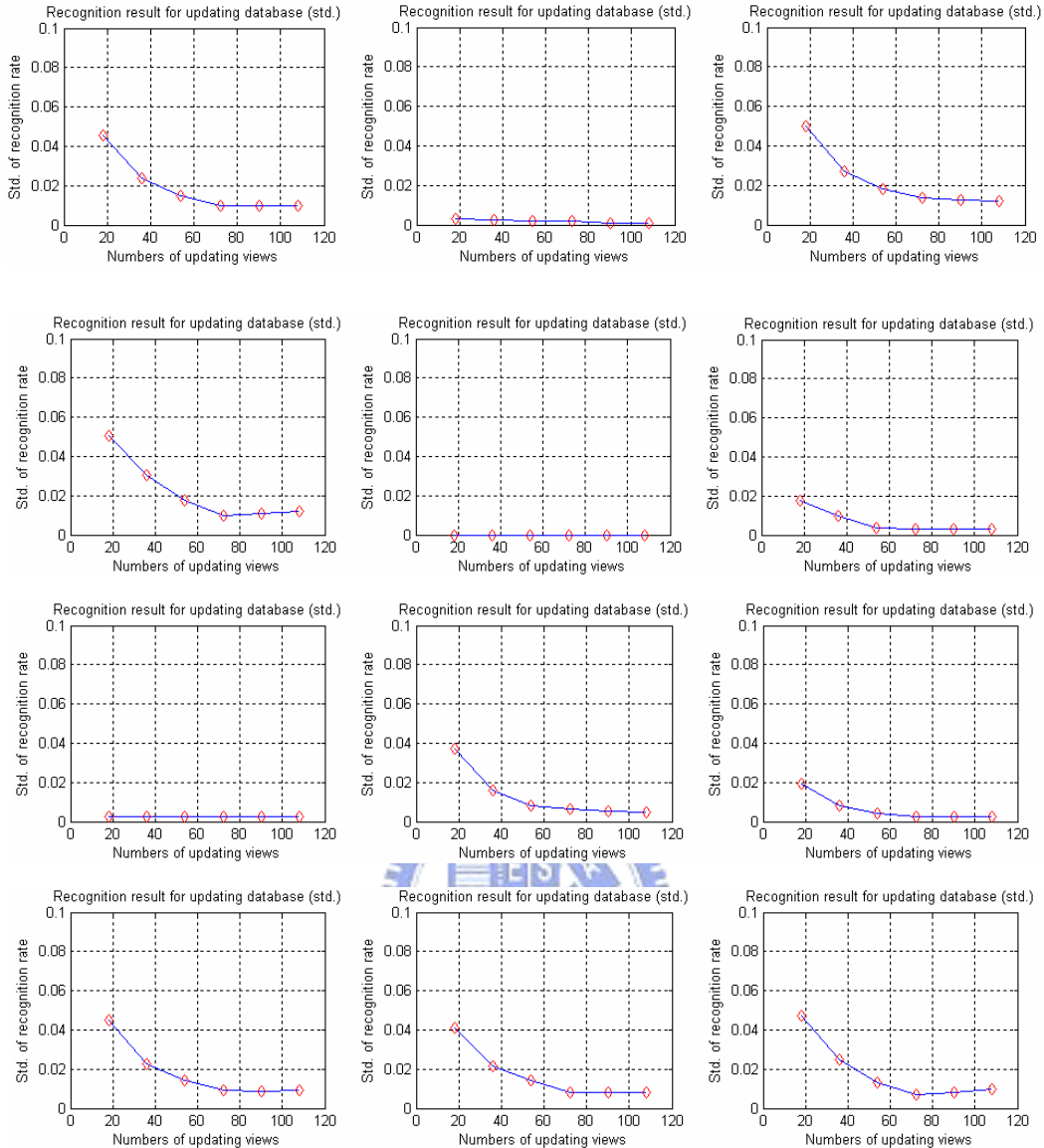


Figure 4-12 Standard deviations of recognition rates using coarse to fine databases (D_{18} , D_{36} , D_{54} , D_{72} , D_{90} and D_{108}), calculated using 200 results

4.2.2 Human Posture Recognition

The efficiency of the proposed method is demonstrated using the second image database (Fig. 4-8). As the same pre-processing in the first experiment, object contour (the contour of human posture) was extracted for further utilization. In the second experiment, two features, called the MAG and θ_z , are extracted from the object contour and be used for building the AODs. The characteristic views of aspects in each AOD are utilized for human posture recognition. Figure 4-13 illustrates the

system architecture of the proposed framework for the human posture recognition. Table 4-8 shows the efficiency of the proposed method with a high recognition rate.

The proposed method decreases the number of aspects for each human posture, and, thus, computing time for recognizing objects is decreased. Furthermore, adopting θ_z instead of PPL reduces the computing time. The similarity measure, which is based on posture contour with N points between an unknown posture and the posture in the database, requires computing N similarity distances while adopting PPL as the feature, but the similarity is computed only once while adopting θ_z .

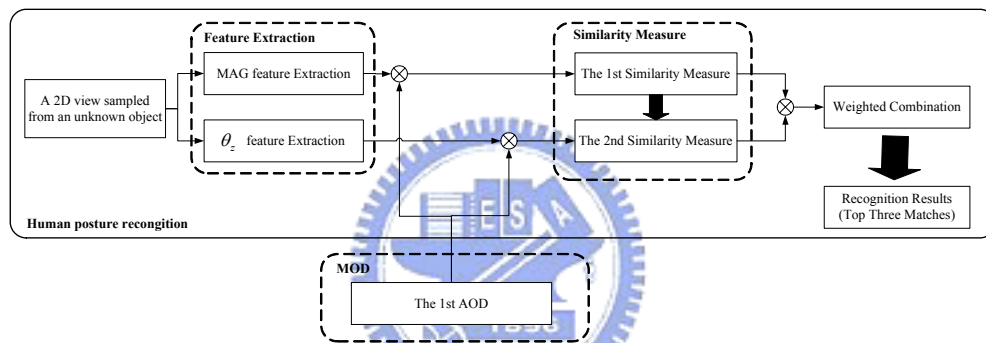


Figure 4-13 The system architecture of the proposed framework applied on the second experiment (human posture recognition).

Table 4-8 Results of human posture recognition using 2D views via MAG and θ_z

Recognition Results	The index of the postures in the second database listed in Fig. 4-8								
	1	2	3	4	5	6	7	8	Avg.
Number of Aspects	8	25	37	41	42	38	8	38	29.63
Top 1 Match (%)	94.91	99.07	98.15	100	99.07	96.30	99.54	100	98.38
Top 2 Match (%)	99.07	99.54	100	100	100	99.54	100	100	99.77
Top 3 Match (%)	100	99.54	100	100	100	99.54	100	100	99.88

4.2.3 Scene Recognition

In the third experiment, training images of 11 locations (Fig. 4-9) in an environment (Fig. 4-14) are obtained by rotating the PTZ camera from -30° to 30° using 1° increments at each location, thereby generating 61 images for each position. Furthermore, 12 Gaussian distributions are adopted in this work to build the blob model (BM feature). The number of aspects of each scene is below 13 after the combination processes. Figure 4-15 presents the sample training images, blob models and conceptual descriptions for each scene. Figure 4-16 illustrates the system architecture of the proposed framework for the scene recognition. Additionally, for the sake of illustration, the set of characteristic views at the 6th position in the indoor environment is cited as an instance of scene (Fig. 4-17).

To test the efficiency of the proposed method, test images are captured by a mobile robot moving in four directions (forward, backward, left and right) at five different distances (5cm, 10cm, 15cm, 20cm and 50cm) and five different levels of occlusion (5%, 10%, 15%, 20% and 50%). Sixty-one images are captured at each position by rotating the camera from -30° to 30° at 1° increments with no occlusion. Figure 4-18 presents the test image samples captured at the 6th position. Figure 4-18(a) shows the test images captured in the forward and backward directions; Figure 4-18(b) presents the test images captured in the left and right directions.

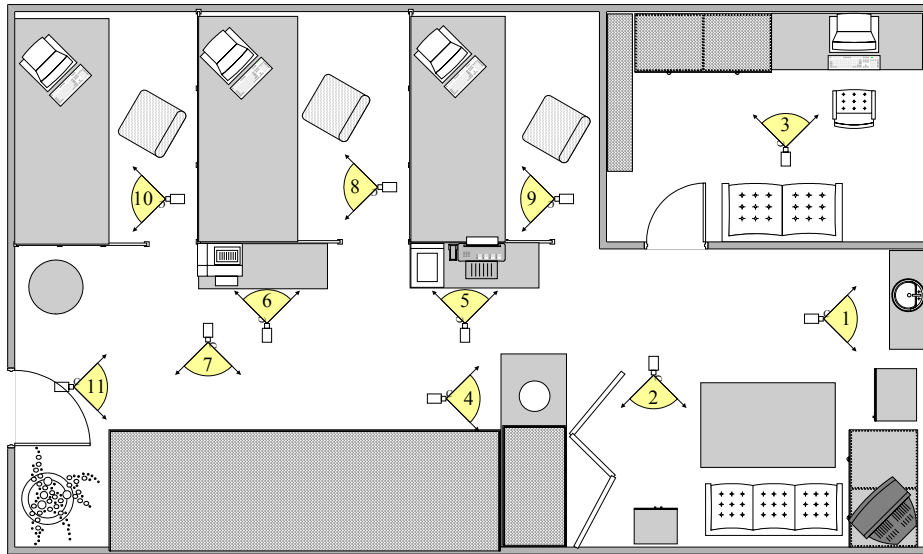


Figure 4-14 The indoor environment from which scenes in the third database are obtained.

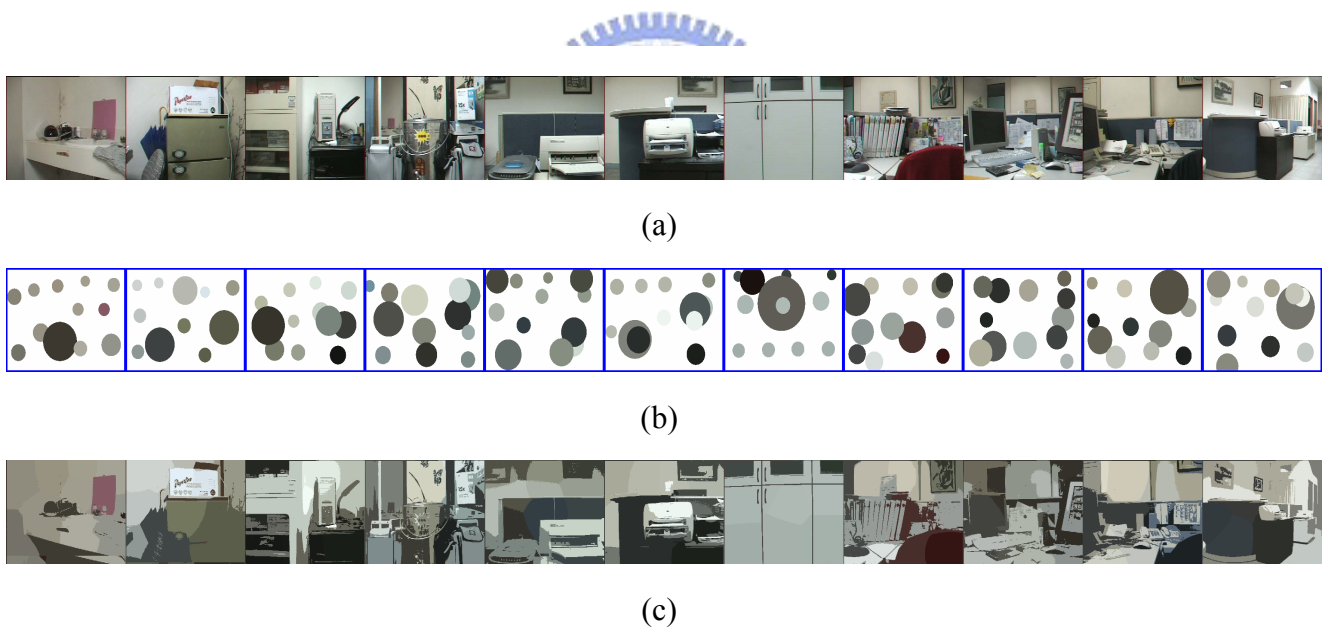


Figure 4-15 The sample training image, blob model and conceptual description of each scene captured in the indoor environment (Fig. 4-14). (a) The sample image captured at each location in the indoor environment (from left to right is positions 1,2,...,11). (b) The blob model of each sample captured image in (a) with 12 Gaussian distributions. (c) The conceptual description of each sample captured image in (a), which are calculated by comparing the original pixel values of each captured image with its blob model.

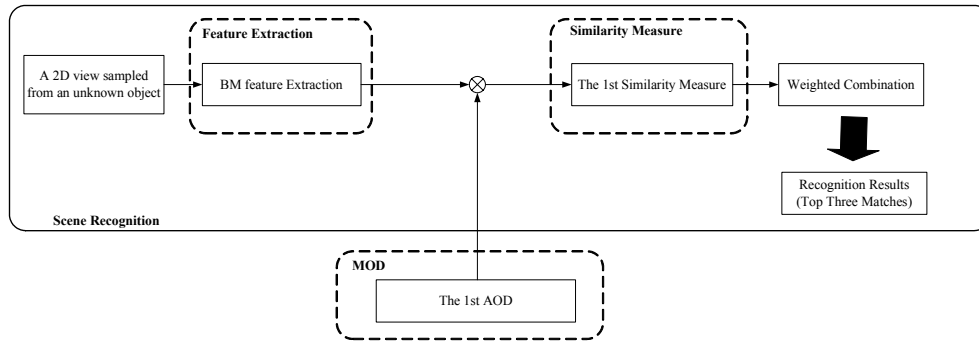


Figure 4-16 The system architecture of the proposed framework applied on the second experiment (human posture recognition).



Figure 4-17 The 11 characteristic views at the 6th position in the indoor environment.



(a)



Figure 4-18 The test images captured from the 6th position in the indoor environment. (a) The test images captured in the forward and backward directions; the shifted distances are as follows: backward 50 *cm*, backward 20 *cm*, backward 15 *cm*, backward 10 *cm*, backward 5 *cm*, 0 *cm*, forward 5 *cm*, forward 10 *cm*, forward 15 *cm*, forward 20 *cm* and forward 50 *cm*. (b) The test images captured in the left and right directions; the shifted distances are left 50 *cm*, left 20 *cm*, left 15 *cm*, left 10 *cm*, left 5 *cm*, 0 *cm*, right 5 *cm*, right 10 *cm*, right 15 *cm*, right 20 *cm* and right 50 *cm*.

To increase the robustness of scene cognition, multiple-view recognition is appropriate for testing. In this experiment, three arbitrary images $I_i (1 \leq i \leq 3)$ obtained with different rotating angles of the PTZ camera are utilized for scene recognition. The first three recognized results that have the first three minimum similarity measures are adopted as candidates for further processing. Suppose O is

the set of recognized result defined as follows:

$$O = \{o_{ij}\}, 1 \leq i \leq 3, 1 \leq j \leq 3, 1 \leq o_{ij} \leq 11,$$

where i is the index of the test image, and j is the index of the order of recognition result.

Three methods are proposed for estimating the final scene cognition result. The first result, R_1 , is estimated using only one recognition result with only one captured image. The second result, R_2 , uses the first three recognition result with only one captured image. The third result, R_3 , uses all combinations of the first three recognition results with three captured images. The descriptions of R_1 , R_2 and R_3 are derived by Eq. (4-7).

$$R_k = \begin{cases} o_{11}, k = 1 \\ o_{11} \cdot \bar{D}_1 + r_1 \cdot D_1, k = 2 \\ o_{11} \cdot (\bar{D}_1 \bar{D}_2 \bar{D}_3) + r_1 \cdot (D_1) + \bar{D}_1 [r_2 \cdot (D_2) + \bar{D}_2 (r_3 \cdot D_3)], k = 3 \end{cases} \quad (4-7)$$

where

$$r_p = \arg \max(F_p), 1 \leq p \leq 3$$

$$D_p = \begin{cases} 1, & \text{if } \arg \max(F_p) \text{ exists} \\ 0, & \text{if } \arg \max(F_p) \text{ doesn't exist} \end{cases}, 1 \leq p \leq 3$$

$$F_p = \{f_{pq}, 1 \leq q \leq 11\}, f_{pq} = \sum_{j=1}^3 \delta(q - v_{pj})$$

Based on recognition results (Table 4-9), the recognition rates of the three methods are all above 95% when the level of occlusion is less than 20% and variation positions are below 20cm. Although the level of occlusion is 50% and variation positions are 50cm, recognition rates are still above 50%. Moreover, the third method, R_3 , performs best and is reasonable based on the human vision used for

localization. When a person enters an unknown place, multi-directional views are captured by the eyes to assist recall of past experiences of the unknown place. In this work, the same strategy is adopted to increase scene cognition robustness.


Table 4-9 Human posture recognition results using 2-D views via BM with position variations and different level of occlusion

Shift Distance (<i>cm</i>) and Direction		Covering Rate (1.000=100%)														
		5%			10%			15%			20%			50%		
		R ₁	R ₂	R ₃	R ₁	R ₂	R ₃	R ₁	R ₂	R ₃	R ₁	R ₂	R ₃	R ₁	R ₂	R ₃
0	cm	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	0.769	0.817
5	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.797	0.775	0.809
	Backward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	0.794	0.763	0.809
	Left	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.794	0.779	0.806
	Right	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.791	0.754	0.818
10	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.796	0.784	0.802
	Backward	0.997	0.979	1.000	1.000	0.997	1.000	1.000	0.997	1.000	0.997	0.996	1.000	0.785	0.738	0.802
	Left	1.000	0.993	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.993	1.000	0.796	0.772	0.805
	Right	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	0.999	1.000	0.770	0.747	0.817
15	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.784	0.769	0.797
	Backward	1.000	0.996	1.000	1.000	0.993	1.000	0.997	0.988	1.000	0.994	0.987	1.000	0.763	0.726	0.781
	Left	0.997	0.979	1.000	0.997	0.979	1.000	0.997	0.979	1.000	0.994	0.975	1.000	0.779	0.736	0.794
	Right	0.992	0.982	0.997	0.992	0.977	0.997	0.992	0.979	0.997	0.992	0.977	0.997	0.726	0.705	0.757
20	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.765	0.751	0.782
	Backward	0.999	0.987	1.000	0.994	0.982	1.000	0.991	0.979	1.000	0.988	0.976	1.000	0.733	0.711	0.748
	Left	0.976	0.960	0.987	0.979	0.970	0.993	0.975	0.961	0.979	0.975	0.963	0.979	0.748	0.699	0.776
	Right	0.975	0.955	0.984	0.979	0.970	0.993	0.976	0.951	0.984	0.975	0.951	0.984	0.714	0.694	0.744
50	Forward	0.845	0.838	0.854	0.845	0.844	0.849	0.842	0.829	0.845	0.832	0.815	0.839	0.508	0.503	0.523
	Backward	0.881	0.839	0.925	0.848	0.821	0.896	0.830	0.809	0.872	0.796	0.785	0.833	0.525	0.508	0.553
	Left	0.750	0.741	0.775	0.748	0.742	0.768	0.733	0.729	0.754	0.723	0.711	0.735	0.502	0.501	0.531
	Right	0.811	0.794	0.841	0.799	0.784	0.827	0.781	0.770	0.817	0.753	0.763	0.778	0.532	0.502	0.531

Chapter 5

Conclusions and Future Researches

5.1 Conclusions



This dissertation presents a robust framework for recognizing 3D objects from 2D views. Two main stages, namely the BSHSR and ISAG, are proposed in the framework. In the BSHSR, a background subtraction scheme involving highlight and shadow removal is proposed for extracting foreground regions with the consideration of illumination variations and dynamic background. Three background models, namely the CBM, GBM, and CSIM, are involved in the BSHSR. In the CBM, ECBM and CCBM are defined to increase the ability of recording a long period of background changes. Furthermore, the STCBM and LTCBM are defined to improve the flexibility and robustness of the gradient-based background subtraction. Most important, the CSIM is proposed to extract the shadow and highlight in this work with a 3D cone-shape boundary and combined with the CBM in the RGB color space. The threshold values τ_{high} and τ_{low} of CSIM can be calculated automatically using the standard deviation of the Gaussian distribution selected using the STCBM and

LTCBM. The proposed 3D cone model is compared with the nonparametric model in a complex indoor environment. The experimental results show the effectiveness of the proposed scheme for background subtraction with shadow and highlight removal.

Moreover, the ISAG is proposed for building the 3D object database using 2D views sampled at random intervals. A robust database, which called a MOD, is composed of AODs. Each AOD is built using the ISAG with one main feature or one main feature and one assistant feature. The final recognition result can be estimated by combining the results calculated from each AOD. To demonstrate the efficiency of the proposed framework, three various object recognition problems, including 3D object recognition, human posture recognition, and scene recognition are performed in the experiments.

Although the threshold values (T_3 and T_4) applied on the ISAG are determined manually case by case, the criteria for selecting T_3 and T_4 are described in Section 3.4. The selection of T_3 and T_4 , which is a trade-off in this work, affects the number of aspects and thus affects the computing time and the error performance. Moreover, the feature selection plays an important role while applying the proposed method in different applications. Although the recognition rate decreases while the number of objects in the database increases in most applications, the proposed framework provides a hierarchical structure to combine more features to maintain the robustness of the recognition system.

Moreover, the ISAG is practical for extracting aspects when features of an object conflict with the first criterion, namely, local monotonicity, as indicated by Cyr and Kimia [1]. For instance, the combinational algorithm developed by Cyr and Kimia [1] cannot efficiently combine 2D views of a human posture with *MAG*. However, the

ISAG overcomes this problem, and efficiently decreases the aspect number. In Fig. 5-1 (a), the blue circles represents 2D views of a human posture, and the black human postures with the red and green lines connected to the blue circle represent the 2D views belonging to the same aspect. In Fig. 5-1 (b), the black human postures are the characteristic views of the aspects of human postures. The two aspects in Fig. 5-1 (b) clearly contain two clusters of 2D views that are opposites.

The proposed method decreases computing time when updating the aspects with new 2D views. Using the method proposed by Cyr and Kimia [1], an object with N collected views requires a computing time of $N(N+1)/2$ to calculate the mutual similarity distances between the $(N+1)$ 2-D views and to extract the aspects and characteristic views. However, the proposed method requires only a computing time of N times to calculate the similarity distance between new incoming views and N existing views via the proposed method. However, as the proposed method has a high computation requirement, improving its efficiency is a topic for future works.

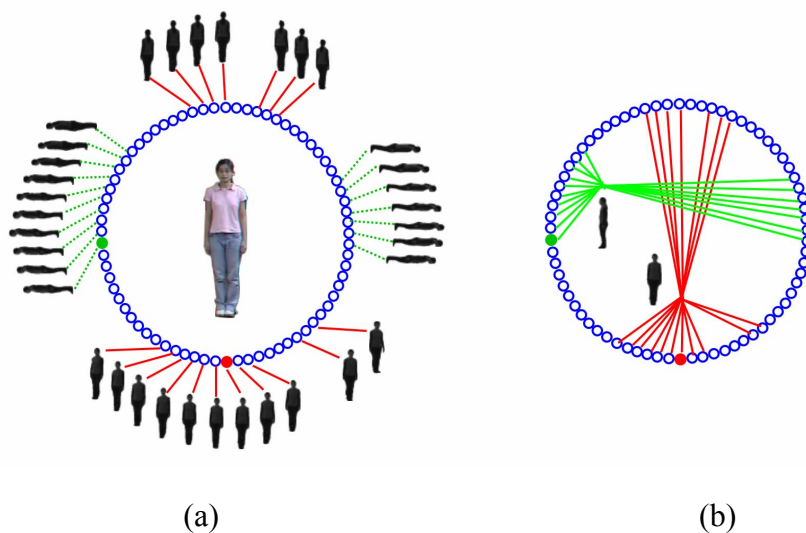


Figure 5-1 The aspect-graph representation of the first human posture listed in Fig. 4-8 via MAG only. (a)The similar 2D views of two aspects. (b)The characteristic views of two aspects.

5.2 Future Researches

To improve the current 3D object recognition system, this dissertation purposes three possible further research directions. The first one is to combine a feature predictor with the original framework (Fig. 1-3). Figure 5-2 illustrates the modified 3D object recognition system. The feature predictor provides supplementary information to improve the robustness of extracted features based on statistics from continuous image frames. After that, even if the extracted features are influenced with noises, the feature predictor can fix errors or compensates the insufficient parts. For example, object contour is adopted as a kind of feature in this work. While the object edges are not determined in good condition, the object contour may converge to a false shape. In such condition, feature predictor should compensate the false part of the extracted contour based on past experience or some prior knowledge.

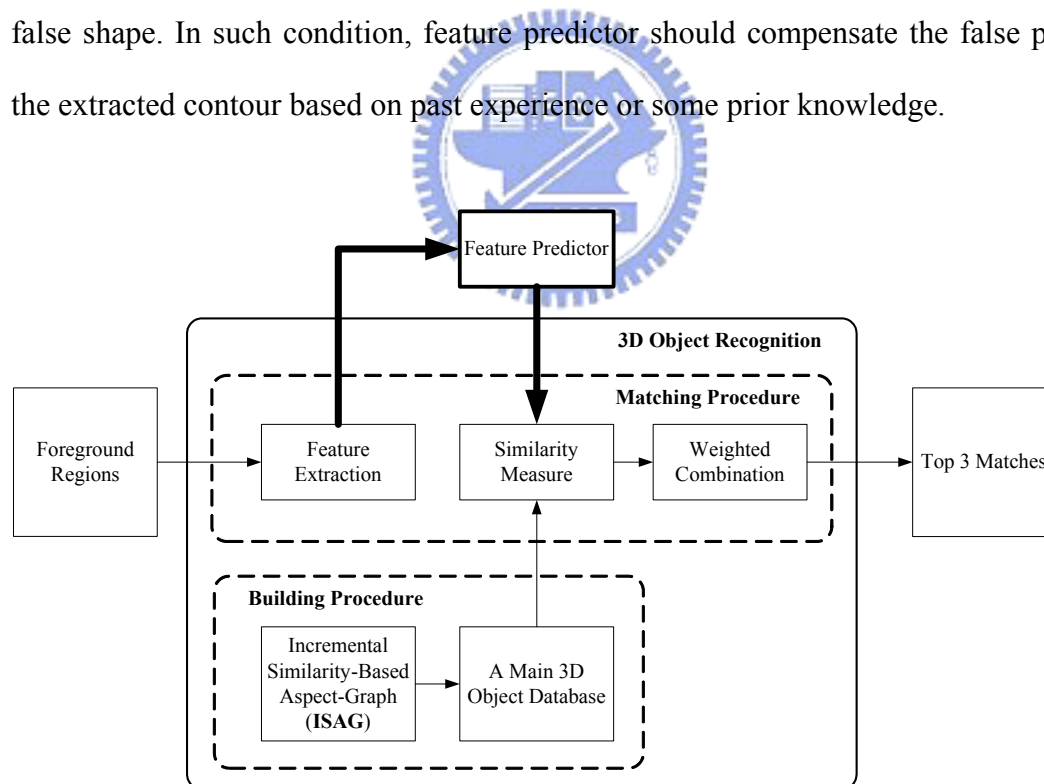


Figure 5-2 3D object recognition system with a combination of a feature predictor and the proposed method illustrated in Fig. 1-3.

The second one is to investigate an efficient searching algorithm for finding top three matches. For example, Dayan *et al.* [72] proposed an approach to transform a multiclass recognition problem into a minimal binary classification problem using Minimal Classification Method (MCM). Unlike prevalent one-versus-one strategy that separates only two classes at each classification, their work has to separate two groups of multiple classes. Figure 5-3 shows the flow chart of the modified framework. As mentioned in Section 3.1, the accuracy of an object representation increases with minimal growth of search space while collecting additional new object views. However, the computing time increases while the number of objects in the database increases. Moreover, only the 2D views concerning the pan motion of the camera are collected for building the database. If we consider the 2D views sampled from the tilt motion of a camera, the characteristic views of an object increase. Therefore, an efficient searching algorithm is necessary while applying the proposed method in practice.

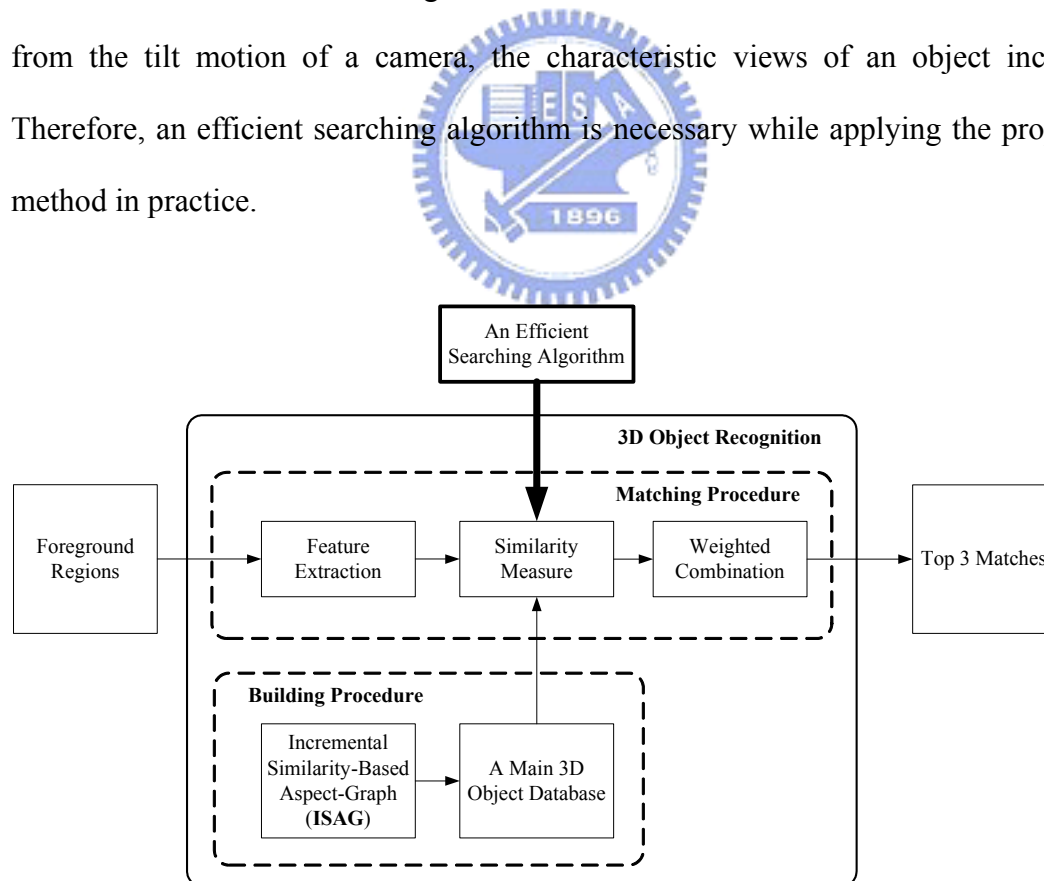


Figure 5-3 3D object recognition system with an efficient searching algorithm.

References

- [1] C.M. Cyr and B. Kimia, "A Similarity-Based Aspect-Graph Approach to 3D Object Recognition," *International Journal of Computer Vision*, vol. 57, no. 1, pp. 5–22, 2004.
- [2] G. Peters, "Theories of Three-Dimensional Object Perception - A Survey," *Recent Research Developments in Pattern Recognition*, Transworld Research Network, 2000.
- [3] G. Mamic and M. Bennamoun, "Representation and recognition of 3D free-form objects," *Digital Signal Processing*, vol. 12, pp. 47-76, 2002.
- [4] R.J. Campbell and P.J. Flynn, "A survey of free-form object representation and recognition techniques," *Computer Vision and Image Understanding*, vol. 81, no. 2, pp.166-210, 2001.
- [5] A.R. Pope, "Model-based Object Recognition. A Survey of Recent Research," Technical Report 94–04, Univ. of British Columbia, Jan. 1994.
- [6] H. Schneiderm and T. Kanade, "Object Detection using the Statistics of Parts," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 151-177, 2004.
- [7] S. Ullman, "Three-Dimensional Object Recognition Based on the Combination of Views," *Cognition*, vol. 67, no. 1, pp. 21-44, July 1998.
- [8] A. Diplaros, T. Gevers, and I. Patras, "Combining Color and Shape Information for Illumination-Viewpoint Invariant Object Recognition," *IEEE Transactions on Image Processing*, vol. 15, no.1, pp. 1-11, 2006.
- [9] T.K. Kim, J. Kittler, and R. Cipolla, "Discriminative Learning and Recognition of Image Set Classes using Canonical Correlations," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005-1018, 2007.
- [10] Y. Shan, H.S. Sawhney, B. Matei, and R. Kumar, "Shapeme Histogram Projection and Matching for Partial Object Recognition," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 28, no.4, pp. 568-577, 2006.
- [11] A.S. Mian, M. Bennamoun, and R.A. Owens, "Three-Dimensional Model-Based Object Recognition and Segmentation in Cluttered Scenes," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1584-1601, 2006.
- [12] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, and B. Schiele, and L. Van Gool, "Towards Multi-View Object Class Detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, USA, June, 2006.

- [13] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411-426, 2007.
- [14] S.K. Naik and C.A. Murthy, "Distinct Multicolored Region Descriptor for Object Recognition," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1291-1296, 2007.
- [15] A. Diplaros, T. Gevers, and I. Patras, "Color-Shape Context for Object Recognition," *IEEE Workshop on Color and Photometric Methods in Computer Vision (in conjunction with ICCV 2003)*, Nice, France, Oct. 2003.
- [16] S. Abbasi and F. Mokhtarian, "Affine-Similar Shape Retrieval: Application to Multiview 3-D Object Recognition," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 131-139, 2001.
- [17] C. de Trazegnies, C. Urdiales, A. Bandera, and F. Sandoval, "3D Object Recognition Based on Curvature Information of Planar Views," *Pattern Recognition*, vol. 36, no. 11, pp. 2571-2584, Nov. 2003.
- [18] C. Dorai and A.K. Jain, "Shape Spectrum Based View Grouping and Matching of 3D Free-Form Objects," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 19, no.10, pp. 1139-1145, 1997.
- [19] J. Zhang, X. Zhang, H. Krim, and G.G. Walter, "Object Representation and Recognition in Shape Spaces," *Pattern Recognition*, vol. 36, no. 5, pp. 1143-1154, 2003.
- [20] V. Blanz, M.J. Tarr, and H.H. Bultho, "What Object Attributes Determine Canonical Views?," *Perception*, vol. 28, pp. 575-599, 1999.
- [21] I. Weiss and M. Ray, "Model-Based Recognition of 3D Objects from Single Images," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol.23, no.2, pp.116-128, 2001.
- [22] S. Kim, G.J. Jang, W.H. Lee, and I.S. Kweon, "Combined Model-Based 3D Object Recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 7, pp. 839-852, 2005.
- [23] K. Akita, "Image Sequence Analysis of Real World Human Motion," *Pattern Recognition*, vol. 17, no.4, pp. 73-83, 1984.
- [24] H. Jiang, Z.N. Li, and M.S. Drew, "Recognizing Posture in Pictures with Successive Convexification and Linear Programming," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 26-37, 2007.
- [25] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani, "Probabilistic Posture Classification for Human-Behavior Analysis," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, no.1, Jan. 2005.

- [26] I. Haritaoglu, D. Harwood, and L.S. Davis, "Ghost: A Human Body Part Labeling System using Silhouettes," in *Proceeding of International Conference on Pattern Recognition*, pp. 77-82, Aug. 1998.
- [27] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 19, no 7, pp. 780-785, 1997.
- [28] L.B. Ozer and W. Wolf, "Real-Time Posture and Activity Recognition," in *Proceeding IEEE Workshop, Motion and Video Computing*, pp.133-138, Dec. 2002.
- [29] L.B. Ozer, T. Lu, and W. Wolf, "Design of A Real-Time Gesture Recognition System: High Performance Through Algorithms and Software," *IEEE Signal Processing Magazine*, vol. 22, pp. 57-64, May 2005.
- [30] Q. Delamarre and O. Faugeras, "3-D Articulated Models and Multi-View Tracking with Silhouettes," *IEEE Conference on Computer Vision*, pp. 716-721, Sept. 1999.
- [31] N. Werghi and Y. Xiao, "Recognition of Human Body Posture from a Cloud of 3-D Data Points using Wavelet Transform Coefficients," in *Proceeding of IEEE International Conference on Automatic Face and. Gesture Recognition*, pp. 70-75, 2002.
- [32] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima, "Real-Time Human Posture Estimation using Monocular Thermal Images," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 492-497, Apr. 1998.
- [33] J. Luo and M. Boutell, "Automatic Image Orientation Detection via Confidence-Based Integration of Low- Level and Semantic Cues," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 715-726, 2005.
- [34] A.M. Martinez and J. Vitria, "Clustering in Image Space for Place Recognition and Visual Annotations for Human-robot Interaction," *IEEE Transactions on System Man and Cybernetics B*, vol. 31, no. 5, pp. 669-682, 2001.
- [35] A. Torralba and A. Oliva, "Statistics of Natural Image Categories," *Network: Computation in Neural Systems*, vol.14, pp. 391-412, 2003.
- [36] R.F. Wang and D.J. Simons, "Active and Passive Scene Recognition across Views," *Cognition*, vol. 70, pp. 191-210, 1999.
- [37] O.M. Mozos, C. Stachniss, and W. Burgard, "Supervised Learning of Places from Range Data using AdaBoost," in *Proceeding of the IEEE International Conference on Robotics and Automation, ICRA*, pp. 1742-1747, Spain, Apr. 2005.

- [38] S. Se, D.G. Lowe, and J.J. Little, "Vision-based Global Localization and Mapping for Mobile Robots," *IEEE Transactions on Robotics*, vol. 21, no. 3, pp. 364-375, 2005.
- [39] L.W. Renninger, and J. Malik, "When is Scene Identification Just Texture Recognition?," *Vision Research*, pp. 2301-2311, 2004.
- [40] I. Ulrich and I. Nourbakhsh, "Appearance Based Place Recognition for Topological Localization," in *IEEE Conference on Robotics and Automation*, pp. 1023-1029, Nov. 2000.
- [41] P. Lamon, A. Tapus, E. Glauser, N. Tomatis, and R. Siegwart, "Environmental Modeling with Fingerprint Sequences for Topological Global Localization," in *IEEE International Conference on Intelligent Robots and Systems*, pp. 3781-3786, Oct. 2003.
- [42] G.N. Desouza and A.C. Kak, "Vision for Mobile Robot Navigation: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 237-267, 2002.
- [43] A. Kosaka and A.C. Kak, "Fast Vision-Guided Mobile Robot Navigation using Model-Based Reasoning and Prediction of Uncertainties," *Computer Vision, Graphic, and Image Processing -- Image Understanding*, vol. 56, no. 3, pp.271-329, Nov. 1992.
- [44] M. Bessa, A. Coelho, J. Bulas Cruz, and A. Chalmers, "Selective Presentation of Perceptually Important Information to Aid Orientation and Navigation in an Urban Environment," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 20, no. 4, pp. 467-482, 2006.
- [45] B.J.A. Kröse, N. Vlassis, R. Bunschoten, and Y. Motomura, "A Probabilistic Model for Appearance-Based Robot Localization," *Image and Vision Computing*, vol. 19, pp. 381-391, 2001.
- [46] A. Oliva and A. Torralba, "Building the Gist of a Scene: The Role of Global Image Features in Recognition," *Progress in Brain Research: Visual perception*, vol. 155, pp. 23-36, 2006.
- [47] G. Cicirelli, T. D'Orazio, and A. Distanti, "Different Learning Methodologies for Vision-Based Navigation Behaviors," *International Journal of Pattern Recognition and Artificial Intelligence*, vol.19, no. 8, pp. 949-975, 2005.
- [48] N. Friedman and S. Russell, "Image Segmentation in Video Sequences: A Probabilistic Approach," in *Proceeding Thirteenth Conference Uncertainty in Artificial Intelligence*, pp. 175-181, Aug. 1997.
- [49] P. Kaew TrakulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proceeding 2nd European Workshop on Advance Video Based Surveillance Systems*, Sept. 2001.
- [50] C. Stauffer and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," in *Proceeding IEEE Conference on Computer Vision and Pattern Recognition*, pp. 246-252, 1999.
- [51] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, "Towards Robust Automatic Traffic Scene Analysis in Real-Time," in *Proceeding*

- of the 33rd IEEE Conference on Decision and Control, pp. 3776 -3781, Dec. 1994.
- [52] S.S. Huang, L.C. Fu, and P.Y. Hsiao, "Region-Level Motion-Based Background Modeling and Subtraction Using MRFs," *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1446-1456, 2007.
- [53] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, "Background and Foreground Modeling using Nonparametric Kernel Density Estimation for Visual Surveillance," in *Proceeding of the IEEE*, vol. 90, pp.1151-1163, July 2002.
- [54] T.G. Stockham, "Image Processing in the Context of A Visual Model," in *Proceeding of the IEEE*, vol. 60 no. 7, pp.828-842, July 1972.
- [55] P.L. Rosin and T. Ellis, "Image Difference Threshold Strategies and Shadow Detection," in *Proceeding of the sixth British Machine Vision Conference*, pp. 347-356, Sept. 1995.
- [56] T.M. Su and J.S. Hu, "Background Removal in Vision Servo System using Gaussian Mixture Model Framework," in *Proceeding of IEEE Conference on Networking, Sensing and Control*, pp. 3776 -3781, March 2004.
- [57] O. Javed, K. Shafique, and M. Shah., "A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information," *IEEE Workshop on Motion and Video Computing*, Orlando, pp. 22-27, Dec. 2002.
- [58] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.
- [59] J.B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proceeding of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, vol. 1, pp. 281-297, 1967.
- [60] T. Hoprasert, D. Harwood, and L.S. Davis, "A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection," in *Proceeding IEEE International Conference Computer Vision, Frame Rate Workshop*, pp. 1-19, Sept. 1999.
- [61] S. Dutta Roy, S. Chaudhury, and S. Banerjee, "Aspect Graph Construction with Noisy Feature Detectors," *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, vol. 33, no. 2, pp.340 -351, 2003.
- [62] I. Chakravarty and H. Freeman, "Characteristic Views as a Basis for Three-Dimensional Object Recognition," in *Proceeding SPIE Conference Robot Vision*, vol. 336, pp. 37-45, 1982.
- [63] J.J. Koenderink and A.J. van Doorn, "The singularities of the visual mapping," *Biological Cybernetics*, vol. 24, pp.51-59, 1976.

- [64] I. Shimshoni and J. Ponce, "Finite-Resolution Aspect Graphs of Polyhedral Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, no 4, pp. 315-327, 1997.
- [65] J.S. Hu, T.M. Su, and S.C. Jen, "Robust Background Subtraction with Shadow Removal for Indoor Environment Surveillance," in *Proceeding of IEEE IROS*, China, Oct. 2006.
- [66] C.C. Lin, "Shape Memorization and Recognition of 3-D Objects Using A Similarity-Based Aspect-Graph Approach," Master Thesis, Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., June 2005.
- [67] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no.6, pp. 679-698, 1986.
- [68] C. Xu and J.L. Prince, "Gradient Vector Flow: A New External Force for Snakes," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 66-71, 1997.
- [69] E. Persoon and K.S. Fu, "Shape Discrimination Using Fourier Descriptors," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 7, no. 3, pp. 170-179, 1977.
- [70] P.C. Lin, "Human Posture Recognition System using 2-D Shape Features," Master Thesis, Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., June 2006.
- [71] J.B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proceeding of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, vol. 1, pp. 281-297, 1967.
- [72] Y. Rubner, C. Tomasi, and L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99-121, Nov. 2000.
- [73] D.M. Sivalingam and N. Pandian, "Minimal Classification Method with Error Correcting Codes for Multi-Class Recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 5, pp. 663-680, 2005.