國立臺灣大學電機資訊學院資訊工程學系

博士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Doctoral Dissertation

使用大範圍高解析視訊監控系統從事目標物之偵測與影子去除

Target Detection and Shadow Removal for a Large-Area High-Resolution Visual Surveillance System

林志瑋

Chih-Wei Lin

指導教授：洪一平 博士

Advisor: Yi-Ping Hung, Ph.D.

中華民國 104 年 1 月

January, 2015

國立臺灣大學博士學位論文
口試委員會審定書

使用大範圍高解析視訊監控系統從事目標物之偵測與影子去除

Target Detection and Shadow Removal for A Large-Area High-Resolution Visual Surveillance System

本論文係林志瑋君（學號 D00922004）在國立臺灣大學資訊工程學系完成之博士學位論文，於民國 104 年 1 月 21 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

洪一平
（指導教授）

賴尚宏　　李明穗

陳祝嵩　　徐繼聖

莊永裕　　范國清

洪一平

系　主　任　趙坤茂

致謝

研究所生活即將告一個段落，承蒙指導教授 洪一平博士，在這幾年博士生涯中的敦敦教誨及細心指導，從研究方向的選擇到論文的撰寫與修改，給予我許多寶貴的意見，使我受益良多，論文得以順利完成。在研究上，洪老師讓我了解研究應該以人為中心，做出對人類有貢獻的研究。生活上，洪老師也常提醒我們要反省自身，注重心靈上的成長，在此致上最誠摯的謝意。

於論文口試期間， 蒙口試委員 范國清、陳祝嵩、賴尚宏、徐繼聖、莊永裕及李明穗博士，給予論文上寶貴的建議與指教，使論文得以更臻完善。

在研究期間要特別感謝學長陳冠文，在論文進行過程中，協助解決遇到的許多問題，並提供他寶貴的經驗和建議。同時也要感謝台大同學們秉翰學長、力韋學長、治權學長、朝欽學長、士堯、翰宏、易程、子軒、彥良、宣輯、文慶、大原、勝傑、重儀、政祥、雪翔、偉傑，以及其他不及備載曾經幫助我的同學們，不僅在課業上給予珍貴的建議，同時也豐富了我的博士班研究生活；有學長、同學及學弟妹的幫忙，才能使我的論文得以順利完成，再次感謝他們，此外要感謝我親愛的家人，在求學的路上給我最大的支持與鼓勵，讓我得以全心投入研究而無後顧之憂。 付梓在即，吾以最誠摯的心感謝我的母校台灣大學，謝謝你帶給我快樂且難忘的研究所生活，並感謝所有愛護我、關心我及支持我的人，最後以此論文與他們一起分享。

摘要

隨著國際間恐怖攻擊事件頻傳，世界各國對於反恐意識提升，於近年來相繼於境內各大城市或重要地點架設視訊安全監控系統。在視訊安全監控系統中，大範圍、高解析度的監控畫面以及智慧化的監控系統是不可或缺的。構建此系統必需同時擁有高品質的輸入影像和顯示裝置以及自動化偵測物體的功能。

鑑於高品質的顯示裝置發展迅速，而高解析度監控攝影機發展緩慢，與顯示器相比並沒有廣泛的被使用。在本文章的第一部分中，我們設計了一個創新的攝影機架構，包含固定式廣角攝影機以及高解析度快速球型攝影機，來建構大範圍、多倍數和多重解析度的視訊監控系統，其提供多重解析度的移動物體資訊。

首先，我們發展一套新的攝影機校正方式，計算固定式廣角攝影機以及高解析度快速球型攝影機之間的對應關係、快速球型攝影機自身旋轉校正以及快速球型攝影機多倍數校正。快速球型攝影機多倍數校正，是基於不同放大倍數在不同角度上具有一致性的特性，來加速校正的過程且不影響正確性; 此校正方式為一新穎的校正方式。完成雙攝影機的校正後，我們使用快速球型攝影機合成一個大範圍高解析的背景影像。當前景物在固定式廣角攝影機中被偵測到後，快速球型攝影機就會被驅使並對使用者選擇的物體進行連續追蹤。最後，我們整合了預先建構好的大範圍高解析背景以及分別由固定式廣角及快速球型攝影機所取的低高解析度前景影像，產生大範圍、多倍數以及高解析度監控畫面。

對於智慧化視訊安全監控系統，自動偵測動物體是一個重要的議題，使用背景相減法來偵測前景物，是研究多年卻仍然很重要的部分。一個好的背景相減演算法可以忍受環境的變化，例如：動態背景和光照的突然變化。在本文章的第二部分中，我們設計了一個空間背景模型(spatial background model, SBM)的新架構。包含兩個主要成分，背景模型(background model, BM)和背景梯度提取器(background gradient extractor, BGE)，來提取前景物體。對於每一張影像，我們都透過傳遞鄰居的資訊來建構背景模型，用於處理動態背景和突然的光線變化。背景梯度提取器與背景模型為同時建構和更新。為保持前景物形狀的完整性，我們利用背景梯度資訊設計禁止傳遞的策略。該方法可以有效地擷取前景和消除背景噪音。

此外，在視訊安全監控的應用中，物體的影子偵測和移除是物體偵測不可或缺的部分。在本文章的第三部分中，我們提出一個新的物體影子去除架構。整合兩個主要偵測器，移動物體偵測器和影子偵測器。對於移動物體，我們利用空間背景模型來偵測移動物體。對於影子去除，我們首先抽取出影子的特徵，包含色度、物理以及紋理性質。然後，使用隨機森林演算法學習影子特徵並產生隨機森林影子偵測模組。接著，我們對時空背景模型產生的結果使用隨機森林影子偵測模組進行影子去除。我們所提出的方法可以有效地檢測出移動物體並移除陰影的影響。此外，透過與其他技術做比較來展示我們所提出的方法，物體偵測與影子去除，的性能。

採用上述方法，使用雙攝影機架構建構出大範圍、高解析度影像，並利用時空背景模型有效的偵測出前景物資訊，進一步利用隨機森林演算法去除移動物體影子的影像，更精準的擷取出移動物體範圍，有效提高智慧化視訊監控的正確性。

關鍵詞：大範圍和高解析度視訊監控系統、雙攝影機系統、空間背景模型、物體偵測、影子去除、支援向量機、隨機森林。

ABSTRACT

Due to the terrorist attacks occur frequently, the anti-terrorism awareness of each country is raising. Therefore, the visual surveillance monitoring systems are setting up at important sites or in major cities in recent years. In visual surveillance monitoring system, the large-area high-resolution visual monitoring view and intelligence monitoring systems are indispensable in surveillance applications. To construct such systems, high-quality image capture, high-resolution display devices and automated detection of objects are required.

Whereas high-quality displays have rapidly developed, the high-resolution surveillance cameras have progressed slowly and remain not widely used compared with displays. In the first part of this study, we designed an innovative framework, using a dual-camera system comprising a wide-angle fixed camera and a high-resolution pan-tilt-zoom (PTZ) camera to construct a large-area high-resolution visual-monitoring system that features multiresolution monitoring of moving objects. First, we developed a novel calibration approach to estimate the relationship between the two cameras and calibrate the PTZ camera. The PTZ camera was calibrated based on the consistent property of distinct pan-tilt angle at various zooming factors, accelerating the calibration process without affecting accuracy; this calibration process has not been reported previously. After calibrating the dual-camera system, we used the PTZ camera and synthesized a large-area high-resolution background image. When foreground targets were detected in the images captured by the wide-angle camera, the PTZ camera was controlled to continuously track the user-selected target. Last, we integrated preconstructed high-resolution background and low-resolution foreground images captured using the wide-angle camera and the high-resolution foreground image captured using the PTZ camera to generate a large-area high-resolution view of the scene.

For intelligence visual surveillance monitoring system, the background subtraction is a crucial component, which has been studied over years. However, an efficient algorithm that can tolerate the environment changes such as dynamic backgrounds and sudden changes of illumination is still demanding. In the second part of this study, we design an innovative framework called the spatial background model (SBM) from a single-layer codebook model. Two main components, the background model (BM) and the background gradient extractor (BGE), are constructed to extract the foreground objects. The background model is built for each single frame with spatial information propagated from the neighbor locations, which is useful for handling dynamic background and sudden lighting changes. The background gradient extractor is also constructed and updated, and we design a propagation forbidden policy for background updating, so as to keep the completeness of foreground shape via the background gradient information. The proposed method can efficiently capture the foreground and eliminates the noise of background.

Moreover, Cast shadows detection and removal is indispensable in the object detection to many surveillance applications. In the third part of this study, we present a novel framework for removing cast shadow of moving objects. Two main components, moving foreground detector and shadow detector, are integrated. For moving objects, we utilize the spatial background model (SBM) to detect the moving objects. For shadow removal, we first extract the shadow features which are chromaticity, physical property, and texture. Then, using the classifier, Random Forest, to learn the shadow model with shadow features. After that, removing the shadow from the result of SBM with the Random Forest shadow detector (RFSD). The proposed method can effectively detect the moving objects and remove the effect of shadow. Furthermore, we demonstrate the performance of our method compared with some techniques of object detection and shadow removal.

Using these methodologies, constructing a large-area and high-resolution view using a dual-camera system, detecting the moving objects with spatial background model, and using random forest shadow detector to remove the shadow effect, to improve the accuracy of intelligence of monitoring surveillance.

Keywords: large-area and high-resolution visual monitoring system; dual-camera system; spatial background model; object detection; shadow removal; support vector machine; random forest.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Traditionally, the visual surveillance system is as a passive system which is used for recording and for reviewing after the occurrence of events. Recently, due to the terrorist attacks, there have been much interest in visual surveillance for security that is developing toward intelligent. The research in visual surveillance spans for several disciplines, such as monitoring system and image analysis, and various visual surveillance systems have been employed for diverse purposes, for example monitoring traffic and ensuring the security of communities, businesses, schools, hospitals, homes, and the public.

In various visual surveillance applications, a large-area high-resolution monitoring system is required. For example, when monitoring the parking area, a wide area must be monitored and details (license plate) must be concurrently captured. A more challenging class of constructing a large-area high-resolution monitoring system is the high-quality image capturing and display devices. The size and resolution of monitors have rapidly increased in the recently year, and monitor prices have decreased. Compared with the enhancement in the quality of displays, the image quality of surveillance cameras has improved slowly. A single camera, such as a wide-angle fixed camera, a PTZ camera, and a fish-eye camera, is used to capture images and can be used for different applications, such as detection [1] [2], tracking [3] [4], and recognition [5]. The field of view of a wide-angle fixed camera or a fish-eye camera is wide and that can be used to monitor the activities of targets over a large area. However, the limitation of these cameras is the low resolution input and that cannot obtain the detail information for security, such as face and license plate. Although a PTZ camera can provide the high-quality image, the field of view is narrow and lots of information of the monitoring environment is lost. Following, the dual-camera system is used, various combination of dual-camera is proposed, such as two PTZ cameras, one PTZ camera with one static cameras, and one PTZ camera with one fish-eye camera. These systems provide the detail information from PTZ camera with zoom in function and obtain the global information from other cameras. However, the information of detail and that of global are not integrated. To achieve the large-area and high-resolution qualities required, an alternative to high-end fixed camera is the pan-tilt zoom (PTZ) camera, which is an inexpensive and suitable compromise, although most of the monitored area is not covered when the PTZ camera is zoomed in. Another solution is constructing a camera network comprising multiple cameras to cover a large area; however, integrating multiple views is difficult. Therefore, in this thesis, we first integrate a PTZ and a wide-angle fixed cameras as a dual-camera monitoring system to construct a large-area high-resolution monitoring view.

Although the large-area high-resolution monitoring system can monitor a wide area and details concurrently, the intelligent is lacked. An intelligent visual monitoring system must automatically detect and track the moving target in which the performance of the system is depend on the accuracy of object detection. The background subtraction is often one of the most crucial components for object detection. However, the problem becomes difficult when dynamic background and illumination changes occur. The dynamic-background problem is caused by periodic-like motion or vibration such as water rippling and waving trees. Illumination changes could be caused by human activity or by natural. For example, people could turn on/off the lights in an indoor environment, and sunlight could also be changed due to the occlusion of clouds in an outdoor environment. Many methods have been proposed for background subtraction. Among these methods, the most intuitive one is the temporal differencing mechanism [6] [7] that subtracts the current image with the previous one(s). The method is adaptive to different environments but easily extracts incomplete foreground objects. Stauffer *et al.*

[1] [8] proposed the Mixture-of-Gaussian (MoG) model which is probably the most popular parametric technique for background modeling. Several Gaussian distributions are used to generate the background candidates for each pixel, which attempts to solve the problem of dynamic background. However, a drawback of MoG is that its parameters are sensitive to the background. Kim *et al*. [2] [9] modeled each pixel as a codebook with one or more codewords depending upon the background variation. The model performs well under illumination changes and also shows some advantages in dealing with dynamic backgrounds. However, it needs a period of time to construct the background model and lacks of the mechanism of re-modeling. In this thesis, we propose a framework that addresses the limitations of these methods, which integrates the information of spatial to construct a spatial background model (SBM). Our method can effectively cope with the problems of dynamic background and sudden changes of illumination.

Due to the ambient light, the cast shadow is induced by moving objects and also detected as the foreground. However, shadows make the object detection incorrect, some applications which are based on object detection become less reliable and reduce the performance of visual surveillance. Therefore, the effective method for removing cast shadows is needed. In this thesis, several useful shadow features, such as chromaticity, physical properties, and texture, are selected and that are used for training shadow classifier based on Random Forest algorithm. Besides, we integrate the spatial background model with Random Forest classifier to produce the effective foreground detector. The proposed method can effectively detect the moving objects and remove the shadow effect. Furthermore, we demonstrate the performance of the proposed method compared with some state-of-the-art techniques of object detection and shadow removal.

1.2 Outline of this Research

In this dissertation, we investigate both object detection and monitoring problems in dual-camera surveillance system. A large-area high-resolution visual monitoring system is proposed. The dual-camera system architecture is adopted, and a wide area and details concurrently captured. Then, we apply the spatial background model into large-area high-resolution monitoring system to detect foreground which overcomes the problems of the dynamic background and the sudden changes of illumination. Finally, the shadow removal technique is adopted to remove the effect of shadow and enhance the accuracy of object detection

1.2.1 Large-Area High-Resolution Visual Monitoring Using a Dual-Camera System

Large-area high-resolution visual monitoring systems are indispensable in surveillance applications. To construct such systems, high-quality image capture and display devices are required. Whereas high-quality displays have rapidly developed, as exemplified by the announcement of the 85-inch 4K ultrahigh-definition TV by Samsung at the 2013 Consumer Electronics Show (CES), the high-resolution surveillance cameras have progressed slowly and remain not widely used compared with displays. In this chapter, we designed an innovative framework, using a dual-camera system comprising a wide-angle fixed camera and a high-resolution pan-tilt-zoom (PTZ) camera to construct a large-area high-resolution visual-monitoring system that features multiresolution monitoring of moving objects. First, we developed a novel calibration approach to estimate the relationship between the two cameras and calibrate the PTZ camera. The PTZ camera was calibrated based on the consistent property of distinct pan-tilt angle at various zooming factors, accelerating the calibration process without affecting accuracy; this calibration process has not been reported previously. After calibrating the dual-camera system, we used the PTZ camera and synthesized a large-area high-resolution background image. When foreground targets were detected in the images captured by the wide-angle camera, the PTZ camera was controlled to continuously track the user-selected target. Last, we integrated preconstructed high-resolution background and low-resolution foreground images captured using the wide-angle camera and the high-resolution foreground image captured using the PTZ camera to generate a large-area high-resolution view of the scene.

1.2.2 Spatial Background Modeling Using a Single-Layer Codebook Model

Background subtraction is a crucial component in visual surveillance, which has been studied over years. However, an efficient algorithm that can tolerate the environment changes such as dynamic backgrounds and sudden changes of illumination is still demanding. In this chapter, we design an innovative framework called the spatial background model (SBM) from a single-layer codebook model. Two main components, the background model (BM) and the background gradient extractor (BGE), are constructed to extract the foreground objects. The background model is built for each single frame with spatial information propagated from the neighbor locations, which is useful for handling dynamic background and sudden lighting changes. The background gradient extractor is also constructed and updated, and we design a propagation forbidden policy for background updating, so as to keep the completeness of foreground shape via the background gradient information. The proposed method can efficiently capture the foreground and eliminates the noise of background. The performance of the proposed method is compared with MoG [1] [8], Codebook [2] [9] and ViBe [10] on the Wallflower [11] and Perception [12] datasets.

1.2.3 Combining Spatial Background Modeling and Random Forest Classifier for Foreground Segmentation and Shadow Removal

Cast shadows detection and removal is indispensable in the object detection to many surveillance applications. In this chapter, we present a novel framework for removing cast shadow. Two main components, moving foreground detector and shadow detector, are integrated. For moving objects, we utilize the spatial background model (SBM) to detect the moving objects which is comprised of the background mode (BM) and the background gradient extractor (BGE) as describe in Chapter 4.

SBM features the object detection in the dynamic background and the sudden lighting changes environments. For shadow removal, we first extract the shadow features which are chromaticity, physical property, and texture. Then, we use the classifier, Random Forest, to learn the shadow model with shadow features. After that, we remove the shadow from the result of SBM with the shadow classifier. The proposed method can effectively detect the moving objects and remove the effect of shadow. Furthermore, we demonstrate the performance of our method compared with some techniques of shadow removal.

1.3 Organization of the Thesis

The rest of this thesis is organized as follows. First, we describe work related to this study in Chapter 2, after which we focus on the large-area high-resolution visual monitoring using a dual-camera system in Chapter 3. In Chapter 4, we present the spatial background model which is useful to overcome the dynamic background and the sudden changes of illumination and be used in Chapter 5. Next, we present an algorithm of shadow removal which can effectively remove the effect of shadow. We conclude this thesis and provide the directions for future work in Chapter 6.

CHAPTER 2

RELATED WORK

In this chapter, we provide context for this research in the backdrop of previous work. Similar to the classification in the previous chapter, we review the related works in the areas of visual monitoring system, background modeling for objects detection, and cast shadow removal and discuss the advantages and limitations that we address.

2.1 Visual Monitoring System

In the past 20 years, numerous frameworks have been proposed for visual surveillance system. Based on the types of cameras used, visual monitoring systems can be classified into single-camera, hybrid-camera, and camera-network systems. The single-camera system, which is widely used in visual surveillance systems, a wide-angle fixed camera, a PTZ camera, or a fish-eye camera can be used as the input device. The wide-angle fixed cameras, which is also called a static or stationary camera, covers a wide angle of view and provides coarse information that is used for visual surveillance. Zhao and Nevatia [13] segmented images of single humans acquired using a wide-angle fixed camera by using the Markov chain Monte Carlo (MCMC) approach. Zhao and Nevatia [14] combined MCMC with the Bayesian maximum a posteriori probability to segment images of humans in crowds. These methods attained a satisfactory level of performance by using the coarse information obtained from the wide-angle fixed camera. However, these frameworks designed for surveillance system cannot provide minute image details. [15] [16] [17] used a PTZ camera in surveillance systems to automatically track humans and vehicles, employing the zoom feature of the PTZ camera to track targets of interest and zoom in on them. Although detail was captured using the zoom function, global information was lacking. Single cameras have been used in numerous studies such as those of object detection [1] [2] [18], tracking [3] [4] [19] [20], and recognition [5] [21] [22]. For tracking and detecting objects, a single low-resolution camera is typically adequate, but the image is not of sufficiently high quality to extract features required for face recognition even when multiple cameras are used. A PTZ camera can provide image details at a high resolution, but background information is lacking.

Sinha *et al.* [23] used a PTZ camera to stitch a panorama view for each scale by using a number of high-resolution images with different scales and the detailed information is obtained after aligning all panoramas together. Although their method provided a multilayered and high-resolution view, the moving objects did not be addressed and the panorama at each scale may contain moving objects that cause false foreground detection. Kang *et al.* [24] [25] used a PTZ camera to construct a panoramic, which is the multilayered high-resolution view, by using distinct zooming factors. During events, they controlled the PTZ camera to monitor the area and update the panorama. However, these methods are inefficient when the panoramic background model is built for each layer by using key frames and the background information cannot be instantly updated when several people are present in the monitored area. A single PTZ camera framework cannot provide multiresolution foreground information; when this framework is used to monitor a target at high-resolution, information regarding other moving objects cannot be captured.

The hybrid-camera monitoring system features combinations such as a pair of PTZ cameras, a fisheye camera plus a PTZ camera, and a wide-angle camera plus a PTZ camera. Marchesotti *et al*. [26]; Zhou *et al*. [27]; Chen *et al.* [28] used a dual-camera system in which the operations of two PTZ cameras were coordinated to track and characterize biometric information. One PTZ captured the wide-angle view, whereas the other PTZ camera provided detail. Singh and Atrey [29] proposed a technique for camera cooperation using Model Predictive Control (MPC) and demonstrated the utility by using two PTZ cameras. Zhang *et al*. [30]; Chen *et al*. [31] used fisheye and PTZ cameras to design a dual-camera system in which a wide field of view was captured for image analysis by the fisheye camera. Alahi *et al*. [32]; Micheloni *et al*. [33] proposed a dual-camera system comprising a wide-angle fixed camera and a PTZ camera, which was used to detect, actively track and recognize people in wide outdoor environments. Reale *et al*. [34] proposed an eye gaze estimation method by using a dual camera comprising wide-angle fixed camera and PTZ cameras which were used to detect face and eyes, respectively. These dual-camera systems concurrently provide wide-angle view at low-resolution and detailed information at high-resolution; however, high-quality background information was not available and in these works, the low- and high-resolution images were separately displayed. Such a display mode is not intuitive to users for monitoring events [35]. Dornaika *et al*. [36] designed a foveated panoramic sensor comprising a fisheye camera and a PTZ camera in which the panoramic view was captured by the fisheye camera and high resolution view is proposed by the PTZ camera to generate the multi-resolution view. Although this system integrated low- and high-resolution information in a single scene, the high-quality background information was also not available. The display mode of dual-camera systems can be classified into overview plus detail (O+D) [37] [38], focus plus context (F+C) [37] [39], and steerable high-resolution display (steerable F+C) [40] [41]. Chen *et al*. [35] demonstrated that the steerable high-resolution display mode exhibits superior performance levels compared with the other displays, providing integrated low- and high-resolution information in a single scene during visual surveillance. Chen *et al*. [35]; Chen *et al*. [42] also used a dual camera comprising static wide-angle and PTZ cameras, developing a multiresolution system that integrated the overview image captured at low resolution by the wide-angle fixed camera with the detailed information captured at high resolution by the PTZ camera and proposing the PTZ camera turning calibration method with difference zoom in factor. Therefore, by using this system, a large field of view could be monitored and the details of targets of interest could be concomitantly observed and zoom in function could be used; however, this system cannot capture the details of the entire scene which means it does not have a high-resolution background image, the PTZ camera tuning calibration method needs a projector to be the auxiliary device and that cannot be used in outdoor environment and it is time-consuming when estimating the correspondence relationship between difference zoom in factor .

Another type visual monitoring system is a camera network comprising numerous cameras. Micheloni *et al*. [33] used several wide-angle cameras and PTZ cameras in a camera network designed for surveillance. The wide-angle camera supplied information regarding object position that was transferred to the PTZ camera for object tracking. Singh *et al*. [43] developed a coopetitive framework for optimal multicamera placement. Natarajan *et al*. [44] proposed a decision-theoretic approach to control and coordinate multiple active cameras for observing and tracking multiple moving targets in a surveillance system. Cai *et al*. [45]; Micheloni *et al*. [46] designed camera networks composed entirely of PTZ cameras. Such systems can be used to observe the entire field of view and concurrently collect detailed information, concomitantly tracking multiple targets. However, high-resolution background information is not captured, multiresolution information is not integrated, and all moving objects cannot be easily tracked with the same number of cameras.

All types of camera systems present advantages and disadvantages. The fixed-camera system featuring single or multiple wide-angle cameras can monitor a wide field of view, but cannot collect detail information regarding areas of interest. The PTZ-camera monitoring system is more flexible than is the wide-angle system, but it cannot concurrently monitor distinct areas and can either monitor a large-area view or collect detailed information. Although, Kang *et al*. [24]; Kang *et al*. [25] constructed a high-resolution panoramic view, the system did not adapt to scene changes, particularly when various foreground targets were present. The hybrid-camera system offers the advantages of the fixed-camera and PTZ-camera systems, but currently available systems [35] [42] [26] [30] cannot yet monitor an entire view at high resolution. The camera network can be used to observe an entire field of view by using a wide-angle camera to track multiple targets concurrently, but the current system [47] cannot monitor the entire view at high resolution and does not integrate the entire view with detailed information.

In this study, we used the advantages of the three aforementioned systems to construct a large-area high-resolution visual monitoring system, integrating the images of hybrid dual-cameras. In the proposed system, the details of the entire monitored field can be observed, enabling those monitoring the scene to accurately grasp the relevant information and allowing them react suitably to the events that transpire; all moving objects at low resolution to make the users aware of suspicious targets, and the target of interest is tracked at high resolution to examine details by using the automatic PTZ-camera control.

2.2 Background Modeling

The background subtraction technique is aim to detect the foreground region form comparing an observed image with an estimated image which is referred to as the background model. In recent year, numerous of background subtraction have been proposed.

Among these methods, the most intuitive one is the temporal differencing mechanism [6] [7] [48] that subtracts the current image with the previous one(s). The method is adaptive to different environments but easily extracts incomplete foreground objects. Pixel-based background modeling methods have been widely used in visual surveillance. [49] [50] [51] [52] used a statistical model of background that modeling of each pixel with a Gaussian which store as a color mean and covariance matrix. Those researchers assumed that each pixel in background is independent and that can be described as a uni-modal of Gaussian. However, a single Gaussian cannot deal with non-stationary background, because of multi-modal backgrounds which contains more than one peak is needed. To overcome the problem of non-stationary background, Stauffer and Grimson [1] [8] proposed Mixture-of-Gaussian (MoG) model which is probably the most popular parametric technique for background modeling. Several Gaussian distributions are used to generate the background candidates for each pixel. Harville *et al*. [53] adopted YUV color plus depth information instead of using RGB color. However, the drawbacks of MoG are that its parameters are sensitive to the background, needs the scene with empty of foreground objects at the beginning of the sequence for a learning phase and the capability of adapting sudden changes, such as the re-positioning of a static object or the turning on of a light is weak.

Instead of Mixture of Gaussian, Kim *et al*. [2] [9] modeled each pixel as a codebook with one or more codewords depending upon the background variation. They observed that the pixel values change over time under lighting variation and that are mostly distributed along the axis going toward the origin point (0, 0, 0) in the RGB space. Based on the observation, they developed a cylinder color model to describe the change of the pixel values over time. Guo *et al*. [54] [55] proposed a hierarchical scheme with block-based and pixel-based codebooks for foreground detection. The model performs well under illumination changes and also shows some advantages in dealing with dynamic backgrounds. However, they need a period of time to construct the background model and lacks of the mechanism of re-modeling.

In the recent year, Barnich and Van Droogenbroeck [10] [56] used a random technique to estimate the background. They modeled each background pixel with a set of samples which is randomly choose from the neighbors. To classify an incoming pixel, calculating the distances between and the set of samples and comparing with a thresholded Euclidean distance. When an incoming pixel has been classified as the background, the random propagation process determines whether this value is used to update the corresponding pixel model and to update the models of neighboring pixels.

The algorithms that are closer in spirit to ours are Kim *et al*. [2] [9] and Barnich and Van Droogenbroeck [10] [56]. We integrate the characteristics of modeling background with the random propagation process for updating background model from Barnich and Van Droogenbroeck [56] [10] and cylinder color model from Kim *et al*. [2] [9] to generate the spatial background model (SBM). Besides, we proposed the two-way propagation policy to cope with the problems of non-stationary

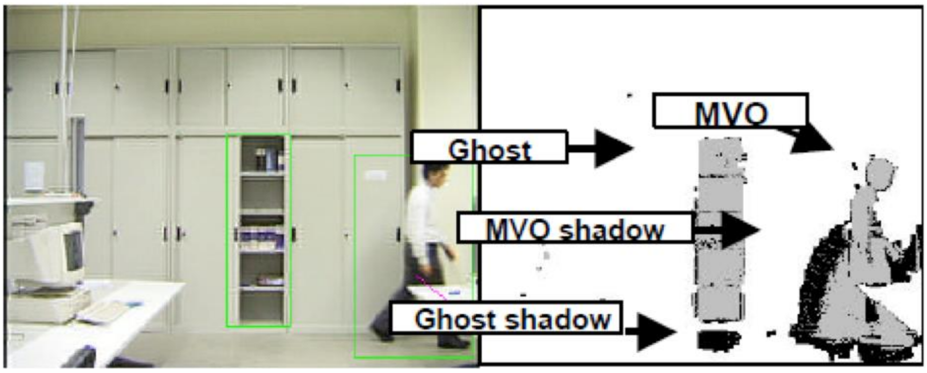


Figure 2.1: moving object, moving object shadow, ghost and ghost shadow [57].

background and ghost effect, and the forbidden propagation policy to keep the completeness of the detected foreground.

2.3 Cast Shadow Removal

Cast shadow is considered as part of moving objects and that affect the results of objects detection. Numerous frameworks have been proposed for cast shadow removal, Cucchiara *et al*. [57] [58] using background subtraction methodology for extracting moving objects which include four components: moving object, moving object shadow, ghost and ghost shadow, as shown in Figure 2.1. The ghost and ghost shadow are removed using optical flow methodology, because the momentum is lacking. For removing cast shadow, the illumination and chromaticity features are considered that is based on the assumption of when the chromaticity of background is similar to that of shadow, the illumination of shadow is darker than background. The main drawback of shadow removal used in this method is sensitive threshold.

Huang and Chen [59] assumed that the shadow distribution (SD) is between direct light (DL) and ambient light (AL) as shown in Figure 2.2. The direct light is white like sun light and ambient light is bluer than direct light. There are four steps for removing shadow. First, the moving objects are extracted by using GMM background model. Then, the weak shadow detector which is based on

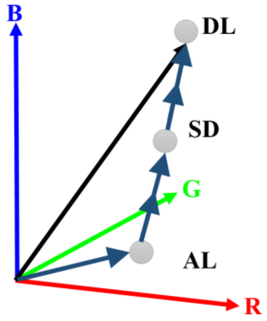


Figure 2.2: The contribution of all direct light sources and ambient illuminance [59].

the RGB color model is used to find the potential shadow as shown in Figure 2.3. The potential shadow falls into the gray area and that can be used to remove the impossible background and foreground. Next, the global shadow model and local shadow model are generated by using the physical-based and gradient features, respectively. Finally the posterior probabilities are evaluated by using the global shadow model and local shadow model which are used to identify foreground, background and shadow.

Sanin *et al*. [60] integrated two features, chromaticity and texture, for removing shadow effect. First, the candidate shadow regions are selected using chromaticity feature proposed by Cucchiara *et al*. [57] [58]. Then, the pixels which have the significant gradient magnitudes in the candidate shadow regions are selected. Finally, the shadow is removed according to the gradient direction which is the correlation between the shadow and the background.

Although, these methods can remove shadow using chromaticity or texture features, the capable

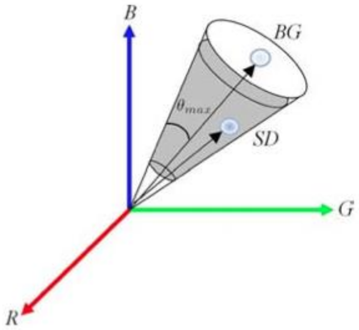


Figure 2.3: The weak shadow detector [59].

of adapting to different environments is lost, because they all need to tune the parameters and thresholds. For example, the parameters and thresholds are tuned for indoor environments that are not suitable for outdoors, because of the illumination of the ambient light and the sun are changed. However, it is difficult to find the best parameters for various environments. Therefore, the learning-based methods for shadow removal are proposed.

Different from the aforementioned methods, Wang *et al*. [61] firstly used Local-Patch Gaussian Mixture Model (LPGMM) which is extend from the Gaussian Mixture Model to detect the moving objects. Each pixel in LPGMM is modeled as a vector which is formed from its observed local neighborhood. Then, several features are used in the SVM shadow classifier to discriminate the foreground and shadow. The background model for detecting moving objects and the shadow classifier for removing shadow are integrated and that is suitable for various environment without tuning parameters or thresholds, because of the features for different scenes are used.

It is important to choose the useful features for shadow removal according to characteristics of features. In our study, the features used for shadow removal is classified into four categories, chromaticity, physical properties, geometric and texture, according to the survey paper [60] [62]. The chromaticity is one of the spectral features. [57] [58] [63] proposed to use HSV color space to be the chromaticity features. The HSV color space is converted from RBG color space that has three components: hue, saturation, and value. Hue is the property of color, such as yellow and blue. Saturation is the purity of color, the high value of saturation makes the color becomes bright. Value is the brightness. For instance, light red and dark red may have the same value of hue, but have different value of brightness. Chromaticity proposed that the hue and saturation are similar at the same region of shadow. Due to the illuminant is blocked, the brightness of shadow is darker than background. The region is classified as shadow, if its color is similar with background and the brightness is darker than background. Moreover, the region with different color or extreme light change is classified as the foreground. The advantages of chromaticity are simple to implement and low computational cost. However, the foreground is classified as background, because of the color space is used and the color of foreground is similar to background.

[59] [64] considered two major illumination sources, ambient light (blue light) and the sun (white light), in outdoor environment to be the physical properties. The shadow is generated by moving object. Although the moving objects block the illumination source of sun, the ambient light is always irradiating on the region. The illumination of sun is blocked and the ratio of ambient light is increasing that makes the color of shadow trend to be blue color. The physical property is the ratio of blue and brightness value, its accuracy is higher than other color features in outdoors and has common performance in indoors, because of the difference light source in indoors. However, when the foreground has the similar color with shadow, the physical property still misclassify, because of the color information is adopted.

Hsieh *et al*. [65] proposed the geometry features for specific objects, such as vehicles and standing people using its orientation, size and shape. First, the gravity of the foreground is used to calculate the angle between foreground and shadow. Then the cut point is found by using the gradient of object’s contour. According to the above results, the foreground and shadow can be separated. The geometry can overcomes the problem of when the colors between foreground and background are similar, because it is not color based. However, it has some limitations, for example, it can only use to detect some specific objects, the object and shadow must have different orientation and only accepts single light source.

Sanin *et al*. [66] proposed the texture feature for shadow removal. Although the region of shadow is darker than background, the texture of this region still exists. There are two steps: (1) rough foreground detection select the candidate region of shadow include foreground and shadow and (2) to classify the candidate region into foreground or shadow by using texture feature. Two texture features are calculated, the different of gradient and that of angle between pixel and its neighborhood. The texture features of shadow are smaller than foreground. The performance of texture feature is well, because it does not rely on color information and does not have the limitations as described in geometry feature. However, when texture feature are similar between foreground and background, it is misclassified, for example, the smooth object on the floor.

However, each feature is suitable for different scene. For example, physical property has the best performance in outdoors, because of it has significant difference between shadow and background compared to other features. When the scene has simple foreground (i.e. only working people in the hall), the feature of geometry substitute physical property and become the most suitable feature in this scene. Therefore, we combine more features without weighted and learn the properties of different scene to raise the performance of shadow removal.

CHAPTER 3

LARGE-AREA HIGH-RESOLUTION VISUAL MONITORING USING A DUAL-CAMERA SYSTEM

3.1 Introduction

Since the 911 terrorist attacks in New York and the subway bombings in London in 2005, security has been a critical research topic in both academia and industry. In recent years, various visual surveillance systems have been employed for diverse purposes, for example monitoring traffic and ensuring the security of communities, businesses, schools, hospitals, homes, and the public. A large-area high-resolution visual monitoring system is required in numerous practical surveillance situations in which a wide area must be monitored and details must be concurrently captured. For instance, when a traffic accident occurs, the scene of the accident must be observed and detailed information must be captured such as the license plate numbers of the vehicles involved and the specific injuries caused to the people involved in the accident.

To construct a large-area high-resolution visual monitoring system, both high-quality image capturing and display devices are required. Recently, the size and resolution of monitors have rapidly increased, and monitor prices have decreased. For example, the Hon Hai Precision Industry Company Ltd sold a 60-inch LED monitor featuring 2K resolution for US$999 in 2012 and Samsung Corporation announced a huge, 85-inch display featuring 4K resolution at the 2013 exhibition of the international Consumer Electronics Show (CES). Compared with the enhancement in the quality of displays, the image quality of surveillance cameras has improved slowly. Our market research

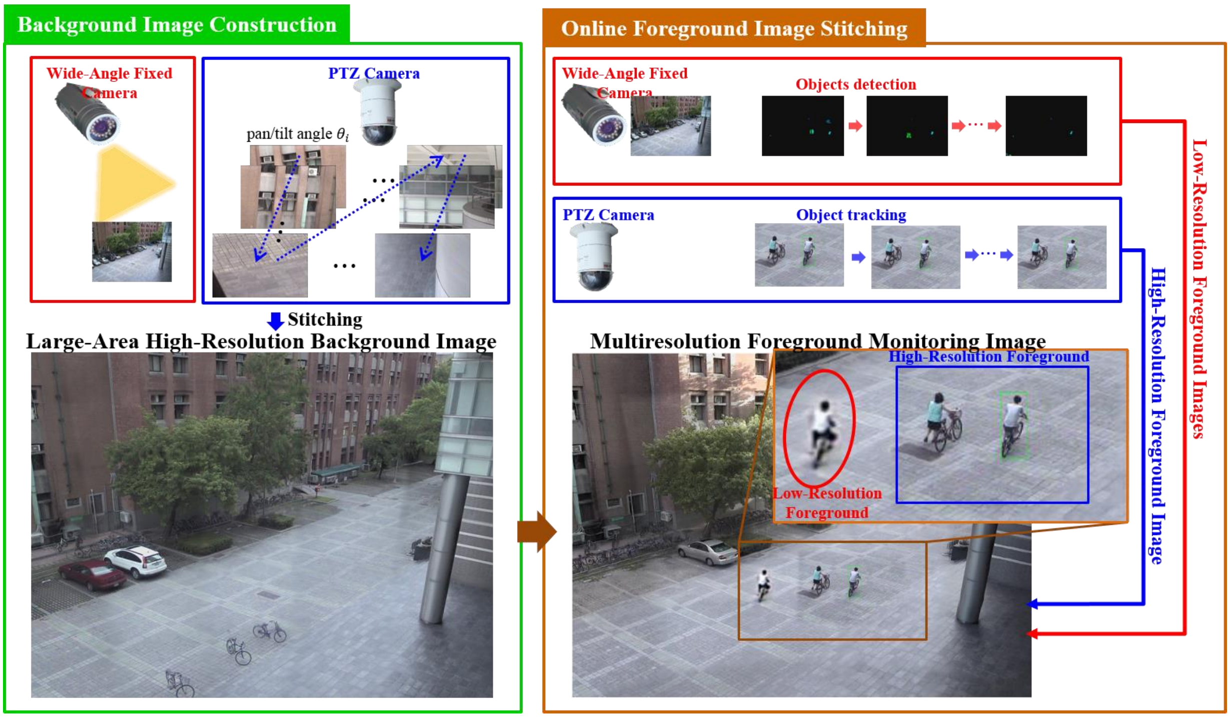

Background Image Construction
Wide-Angle Fixed Camera
PTZ Camera
pan/tilt angle θ_i
Stitching
Large-Area High-Resolution Background Image
Online Foreground Image Stitching
Wide-Angle Fixed Camera
Objects detection
PTZ Camera
Object tracking
Multiresolution Foreground Monitoring Image
High-Resolution Foreground
Low-Resolution Foreground
Low-Resolution Foreground Images
High-Resolution Foreground Image

Figure 3.1: An overview of our system.

indicates that, the highest resolution surveillance camera currently on the market is a 10-megapixel camera with a resolution of 3648×2752 pixels that captures images at only 7 fps (frames per second), sold by Arecont Vision for more than US$1100. At the 2013 CES, Bosch announced an ultrahigh-definition 4K surveillance camera featuring 3840×2160 pixels (8.3 megapixels) that captures images at 30 fps or captures images at a full 12-megapixel resolution at 20fps. Although this camera is not yet commercially available and its price has not been announced, we expect the camera to be substantially more expensive compared with current high-resolution surveillance cameras. To achieve the large-area and high-resolution qualities required, an alternative to high-end fixed camera is the pan-tilt zoom (PTZ) camera, which is an inexpensive and suitable compromise, although most of the monitored area is not covered when the PTZ camera is zoomed in. Another solution is constructing a camera network comprising multiple cameras to cover a large area; however, integrating multiple views is difficult.

In this study, we take advantage of the aforementioned cameras to develop a hybrid dual-camera system that includes a PTZ camera and a wide-angle fixed camera. Notice that, only two standard cameras are used in our system, costing US$2590. Furthermore, we proposed a framework to integrate both the zoomed-in and wide-angle views to construct a large-area high-resolution view that concurrently provides the large-area and high-resolution coverage required for visual monitoring. An overview of our system is presented in Figure 3.1. The wide-angle fixed camera can cover a fixed but wide area, and the PTZ camera can be controlled to focus on a specific target and then zoom in and out to capture videos of various scales. All images obtained from the PTZ camera were integrated to construct a large-area high-resolution background image with a resolution of 4320×2880 pixels (12.4 megapixels) as shown in the left side of Figure 3.1, achieving a maximal resolution of 11520×8160 pixels (94 megapixels), which is considerably higher than 12-megapixel resolution yielded by the Bosch. After moving objects are detected by a wide-angle fixed camera, the user can select a target of interest and use the PTZ camera to continuously and automatically track that target. Finally, the interesting target is integrated into the background and a large-area and high-resolution view is generated. Employing the proposed system, a user can observe both a target of interest at a high resolution and all activities in the synthesized monitoring area at a comparatively lower resolution (Figure 3.1, right) as part of a process called multiresolution foreground image monitoring.

This chapter is organized as follows. First, we focus on the calibration of the camera, namely, the calibration between the wide-angle fixed camera and the PTZ camera, the PTZ camera turning calibration, and the multilayer calibration of the PTZ camera. Next, we demonstrate the construction of a large-area high-resolution visual monitoring system by using the results of the camera calibration. Last, we present the experimental results, and conclusion.

To enhance readability, in this chapter, the image captured using the wide-angle fixed camera is referred to as the “*overview image*“, and the image captured using the PTZ camera is the “*detail image*“. The detail image captured using a pan/tilt angle $\theta_0 = [p_0, t_0]$ and the zooming factor 1 that

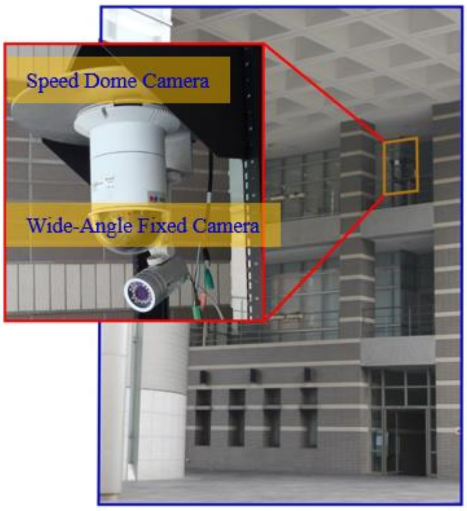


Figure 3.2: The hardware system.

make the detail image similar to the overview image is termed as the “*reference image*“. The PTZ camera captured at various zooming factors or distinct magnifications are called “*layers*“.

3.2 System Architecture

The proposed system was designed to cover a wide monitoring area at high-resolution and provide focused on targets. We installed a dual-camera system in a building (Figure 3.2) at a height of approximately 10 m. The wide-angle camera used was ACM-1311N (resolution 680 × 480 pixels), and the PTZ camera was CAM-6510N (resolution 720 × 480 pixels). In the installation, the cameras can be considered concentric because of the long distance between the cameras and the monitored area; in Section 3.3, we explain how this concentric property facilitates calibrating the cameras.

To effectively integrate a wide-angle fixed camera and PTZ camera, multiple key tasks must be considered. As shown in Figure 3.3, the proposed system includes three primary technical components: offline camera calibration, online construction of large-area high-resolution background images, and online multiresolution foreground-image stitching. First, we determined the relationship between the two cameras. When a target of interest is selected from the overview image, its

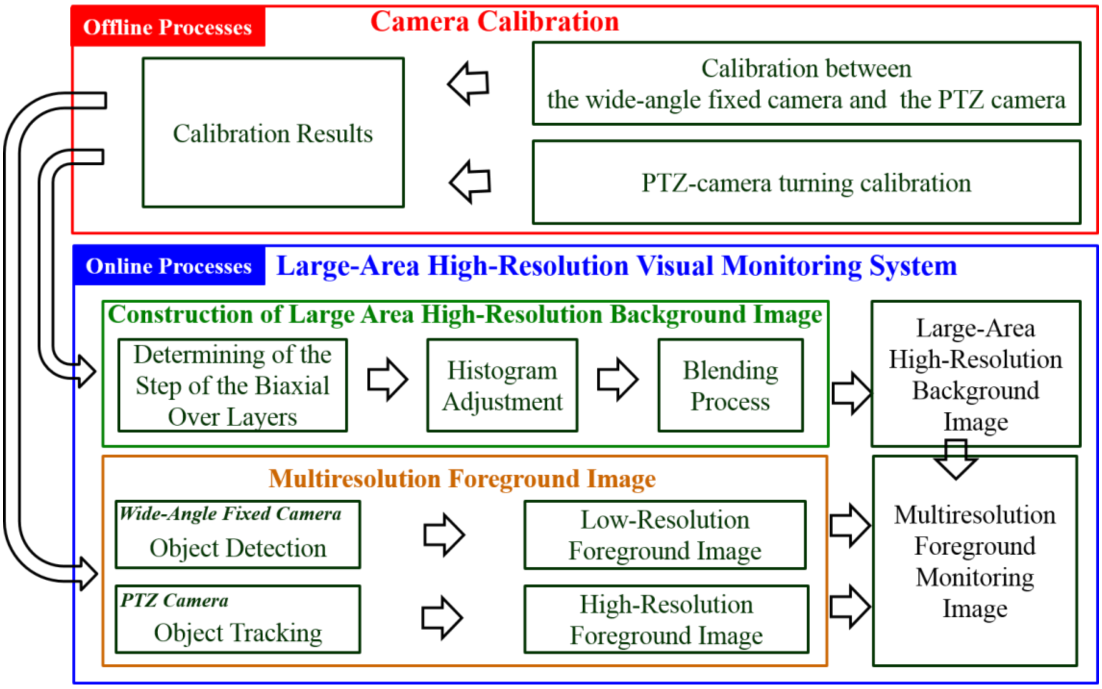

Offline Processes
Camera Calibration
Calibration Results
Calibration between the wide-angle fixed camera and the PTZ camera
PTZ-camera turning calibration
Online Processes
Large-Area High-Resolution Visual Monitoring System
Construction of Large Area High-Resolution Background Image
Determining of the Step of the Biaxial Over Layers
Histogram Adjustment
Blending Process
Large-Area High-Resolution Background Image
Multiresolution Foreground Image
Wide-Angle Fixed Camera
Object Detection
Low-Resolution Foreground Image
PTZ Camera
Object Tracking
High-Resolution Foreground Image
Multiresolution Foreground Monitoring Image

Figure 3.3: The work flow of our system.

corresponding coordinates must be obtained in the reference image and the PTZ camera is subsequently turned to the corresponding pan/tilt to acquire a detail image of the target. Thus, we designed calibration schemes to estimate the relationship that transforms the image coordinates from the overview image to the reference image, and to calculate the turning calibration of the PTZ camera that transforms the image coordinates of the reference image to the corresponding detail image at the pan/tilt. After calibrating the cameras, we used the calibration results to generate a large-area high-resolution background by stitching the detail image captured at each pan/tilt. This process includes the technical components of determining the step of the biaxis over layers, histogram adjustment, and blending. When the system initiates, moving objects are detected in the overview image and low resolution foreground images are generated. We can select one of the objects of interest as a target, and the proposed system controls the PTZ camera to continuously track the target, capturing high-resolution foreground images. Finally, we integrate the low- and high-resolution foreground images into the preconstructed large-area high-resolution background image, generating the multiresolution foreground monitoring image.

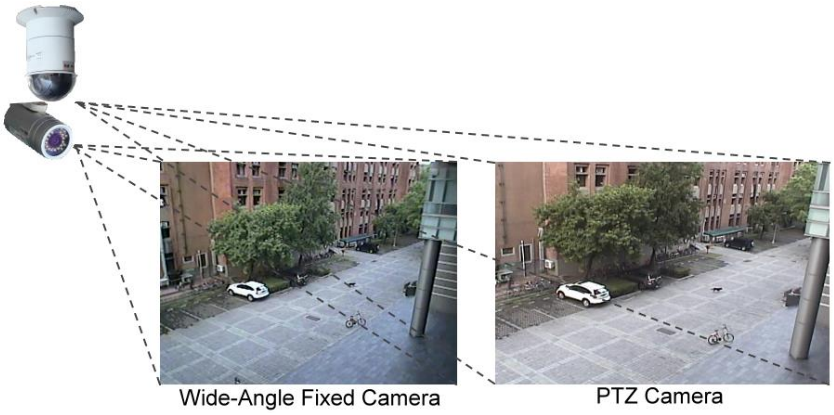


Figure 3.4: The concentric structure.

3.3 Camera Calibration

Using the proposed system requires turning the PTZ camera and zooming it toward any point selected on the overview image. Therefore, the image coordinates should be transformed from the overview image to the reference image, and then the corresponding pan/tilt should be calculated, enabling the capture of the detail image focused on the selected point in the overview image. Thus camera calibration involved two steps when implementing the proposed system. The first step was estimating the transformation between the coordinates of the overview and reference images. The second step was estimating the relationships between the coordinates of the reference image and the corresponding pan/tilt.

3.3.1 Calibration between Wide-Angle and PTZ Cameras

The proposed system requires estimating the transformation of coordinates between the overview and reference images (i.e., we must calculate the corresponding image coordinates from the overview image to the reference image). When using two concentric cameras, the image coordinates of two cameras can be transformed using homography or a 3 × 3 perspective-transform matrix [67] [68]. Numerous surveillance applications have adopted using concentric cameras; for example, Elder *et al*. [69] used such cameras to generate fused imagery, Prince *et al*. [70]; Wheeler *et al*. [71] used it to

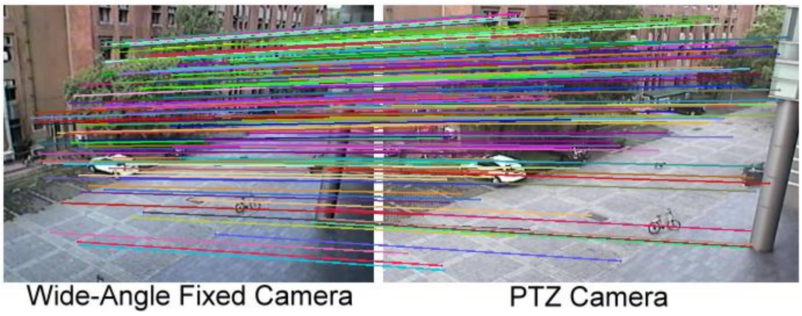


Figure 3.5: The corresponding SIFT features points.

detect and recognize faces, and Elder *et al*. [72] used it to detect people. In the proposed installation, the two cameras were nonconcentric, but could be considered to be concentric because the centers of their lenses were close to each other and the monitored area was adequately far from the camera (Figure 3.1). This installation is commonly accepted in implemented surveillance applications [15] [30] [47] [73]. A view captured using the wide-angle fixed camera and the PTZ camera in the proposed installation is shown in Figure 3.4.

To estimate homography, the Scale Invariant Feature Transform (SIFT) algorithm [67] [74] was used to estimate the corresponding feature points between the overview image and the reference image because of the robustness and distinctiveness of this algorithm (Figure 3.5). Next, we calculated the homography H_{RI}^{OI} between the overview and reference images by using at least four corresponding points from the two images, yielding the following:

$$s\begin{bmatrix} x^r & y^r & 1 \end{bmatrix}^T = H_{RI}^{OI}\begin{bmatrix} x^o & y^o & 1 \end{bmatrix}^T, \tag{3.1}$$

where s is a scale factor, and (x^o, y^o) and (x^r, y^r) are the coordinates of the corresponding feature points in the overview and reference images, respectively.

3.3.2 PTZ Camera Calibration

After translating the image coordinates from the overview image to the reference image, we determined the corresponding pan/tilt of the PTZ camera required to turn the PTZ camera toward the

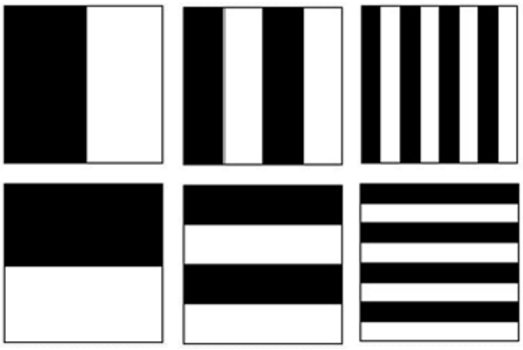


Figure 3.6: Gray-Code patterns.

position chosen from the overview image. In this subsection, we explain two calibration methods, in-factory and on-site calibrations, both of which can be performed without human control.

In-Factory Calibration

In-factory calibration is performed to generate the look-up table required to turn the PTZ camera; this method was used in our previous studies [35] [42]. The in-factory calibration method involves using an auxiliary fixed projector to project gray-code patterns [75] (Figure 3.6) in front of a wall and subsequently capture the projected pattern images by using the PTZ camera at each pan/tilt angle and at the pan/tilt angle $\theta_0 = [p_0, t_0]$, serving as the detail images and reference image, respectively. Next, the corresponding feature points between each detail image and the reference image are estimated and used to calculate the homographies for each detail image and the reference image, after which we can compute the corresponding coordinates of the image center of each pan/tilt in the reference image based on these homographies. By using interpolation, the corresponding pan/tilt can be computed when a user selects an arbitrary point on the reference image. Additional details of this process are available in [35] [42] .

On-Site Calibration

Although the in-factory calibration method is satisfactory, in various applications the dual cameras are installed at elevated sites, making dismantling the devices and using the in-factory calibration method challenging. Therefore, we propose using a novel camera calibration method, called on-site

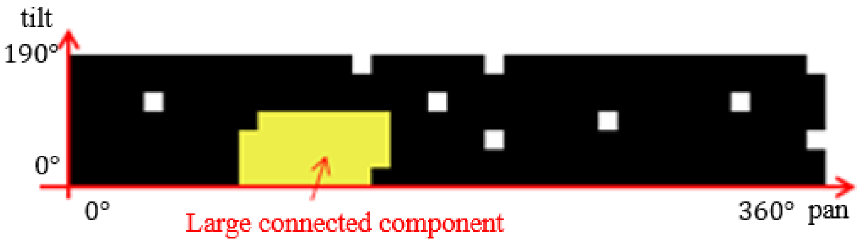


Figure 3.7: Accepted pan/tilt map.

calibration in the proposed system; this method is more convenient and robust compared with in-factory calibration. On-site calibration features two automatic procedures. The first procedure is used to automatically determine the calibration boundary, and the second is used to calibrate between Layer 1 (zooming factor equal to one) and the reference image, a step called PTZ-camera turning calibration.

In the proposed system, the maximal pan and tilt angles of the PTZ camera are 360° and 190 °, respectively. Before calibrating the turning of the PTZ camera, we determined the rotation range in which the PTZ camera was to be calibrated, enabling the detail and overview images to be superimposed. Thus, first procedure is determining the calibration boundary, which includes (1) determining the accepted pan/tilt map; and (2) deciding the calibration boundary.

In Step 1, we generated an accept pan/tilt map that indicated the rough pan/tilt of the detail images that can cover the overview image. The symbols are defined as follows: the detail image captured by the PTZ camera at pan/tilt angle $\theta_i = [p_i, t_i]$ with the use of Layer 1 is expressed as I_{DI}; the pan/tile angle $\theta_i = [p_i, t_i]$ with Layer 1 is termed as θ_{i,Z_1} for each I_{DI_i}, and I_{OV} is the overview image. In this step, the pan and tilt angles of the PTZ camera used for each interval rotation can be changed by the user; in the proposed implementation, we sampled these using angles of 11 ° and 9 °, respectively. For each I_{DI_i}, and I_{OV}, we use a feature-based method, SIFT [74], to determine the corresponding feature points between the two images. If the number of corresponding feature points is higher than a threshold T, we accept the pan/tile angle $\theta_i = [p_i, t_i]$ and set the value as 1 in the accepted pan/tilt map; if the number is lower than the threshold, we reject the pan/tile angle $\theta_i = [p_i, t_i]$ and set the value as 0 in the accepted pan/tilt map. If the pan/tile angle $\theta_i = [p_i, t_i]$ is accepted, then the homography H_{OI}^{DI} between I_{OV} and I_{DI} can be calculated. An example of accepted pan/tilt map is shown in Figure 3.7, wherein the white and yellow points indicate the accepted pan/tilt angles, and the black points indicate the rejected pan/tilt angles.

In Step 2, we determined the largest connected component in the accepted pan/tilt map, shown as the yellow area in Figure 3.7, this was considered the possible range of the field of view of the wide-angle fixed camera. Next, we used homography, H_{OI}^{DI} to translate the coordinate of the center of each detail image at θ_{i,Z_1} within the area of the largest connected component estimated into the overview image coordinate, which is called C_{DI_i}. We bilinearly interpolated C_{DI_i} to generate C_{DI} which is the center coordinate of detail image at all valid pan/tilt angle in the overview image. We determined the center coordinate of reference image C_{RI} in the overview image by using the following:

$$C_{RI} = \arg\min_{C_{DI}} \left\| C_{DI} - C_{OI} \right\|_{norm}, \tag{3.2}$$

where C_{OI} represents the center coordinate of the overview image. Furthermore, we chose the minimal and maximal valid C_{DI} values and set the corresponding pan/tilt angles as the minimal and maximal pan/tilt angles of the calibration boundary and denoted them as $p_{\min}$, $t_{\min}$, $p_{\max}$, and $t_{\max}$.

After determining the calibration boundary and reference image, the second procedure is calibrating the turning of the PTZ camera. We estimated the relationship between the coordinates of the reference image and turning angle of the PTZ camera, allowing the PTZ camera to be turned to a corresponding pan/tilt angle and then focused on the position selected in the reference image. To accelerate the calibration process, we calibrate only for every interval of the pan and tilt angles p_{step}

and t_{step}, respectively, using interpolation to calculate the results of the other angles. To determine the calibration interval of the pan and the tilt angles, we first calculate the number of intervals N by using the following equation:

$$N = \sqrt{\frac{Area_{OI}}{Area_{DI,z=MAX}}}, \tag{3.3}$$

where $Area_{OI}$ is the area of the overview image, and $Area_{DI,z=MAX}$ is the area of the corresponding field of view of the PTZ camera in the overview image obtained using the maximal zooming factor, which was 15 in the experiments. Ideally, distinct zooming factors should exhibit distinct numbers of the intervals N, but for simplicity, we choose the maximal number, ensuring that the overview image can be covered by merging all the detail images used for the calibration.

After calculating the number of calibration intervals N, we used the following equations to calculate the rotation interval p_{step} and t_{step}:

$$p_{step} = \frac{p_{\max} - p_{\min}}{N-1}, \tag{3.4}$$

$$t_{step} = \frac{t_{\max} - t_{\min}}{N-1} \tag{3.5}$$

Figure 3.8 shows the coordinates in the reference image that corresponded to all calibration positions with the calibration intervals p_{step} and t_{step} in the experiment. The red rectangle indicates the range of the wide-angle fixed-camera view, the green rectangle is the range of the PTZ-camera view, and the blue rectangle is the range of the PTZ-camera view with maximum zooming factor which is 15 in our experiment. The p_{step} and t_{step} angles were approximately to 1° in the proposed implementation.

After determining the sampled calibration positions, the PTZ-turning calibration involves

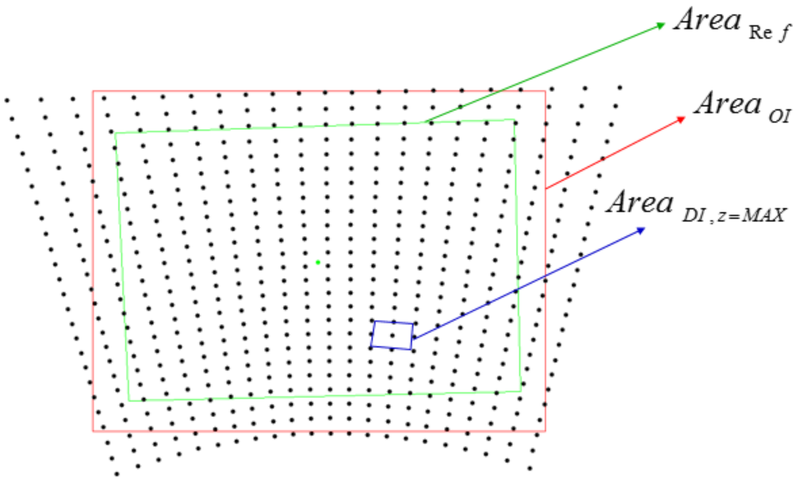
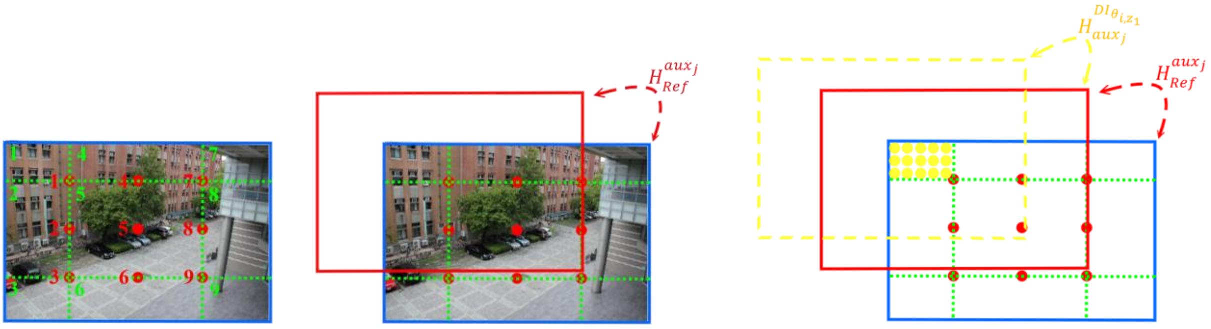


Figure 3.8: The coordinates map for background update.

obtaining the homography $H_{\mathrm{Re}\,f}^{DI_{z_1}}$, which is the relationship between the coordinates of the reference and detail images at each pan/tilt angle with the use of Layer 1. Here, we added the auxiliary views to avoid inadequate corresponding feature points between the reference and detail images at pan/tilt angle θ_{i,Z_1}. First, we obtained the reference image and collected 9 auxiliary images from the predefined positions, which are shown as red dots in Figure 3.9 (a). We use a method similar to that described in Section 3.3.1 and determined the homographies based on distinct auxiliary views j to the reference image $H_{\mathrm{Re}\,f}^{aux_j}$ where j =1 to 9. Second, we divided the image into 9 regions, and each detail image exhibiting the pan/tilt angle θ_{i,Z_1} was mapped to the predefined auxiliary view. The regions and their corresponding auxiliary view are shown in Figure 3.9. For example, the θ_{i,Z_1} angles that fall into the Green Region 1 map to the auxiliary view of Red Point 1. Next, we determined the homographies for all views θ_{i,Z_1} angles that were appropriate for the auxiliary view j, $H_{aux_j}^{DI_{\theta_{i,z_1}}}$. Finally, we determined the relationship between the reference and detail images at pan/tilt angles θ_{i,Z_1} according to the following equation (Figure 3.9 (b)(c)):

$$H_{aux_j}^{DI_{\theta_{i,z_1}}} = H_{\mathrm{Re}\,f}^{aux_j} \times H_{aux_j}^{DI_{\theta_{i,z_1}}} . \qquad (3.6)$$

Multilayer PTZ-Camera Calibration



(a) The position of the auxiliary views (b) The relationship between the auxiliary views and the reference image (c) The relationship between the auxiliary views and detail images

Figure 3.9: The transformation between Layer 1 and the reference image.

Although the aforementioned calibration procedure is satisfactory, PTZ-turning calibration is ineffective given distinct zooming factors because the focal length changes as Layer k changes; thus, the principal point of the PTZ camera in the image is offset from the exact center of the image even at the same pan/tilt angle. Therefore, we calibrated PTZ camera by using distinct zooming factors, using multilayer PTZ-camera calibration.

The proposed multilayer PTZ-camera calibration is based on the consistent property of distinct pan/tilt angles at various zooming factors (i.e., the relationship between the detail images obtained using distinct zooming factors is independent of the pan/tilt angle). The consistent property does not change when the pan/tilt angles are changed; thus, this method does not rely on the accuracy of pan/tilt operation. By contrast, the PTZ camera calibration described in Section 3.3.2 relies on the repeatability of the pan/tilt operation performed using modern PTZ cameras. The relationships between the detail images that exhibit various zooming factors can be transformed using the homography of $H_{\theta_{z_k}}^{\theta_{z_{k+1}}}$, which represents the transformation between the Layers k and $k+1$ at the pan/tilt angle θ. The homographies between distinct layers are independent of the pan/tilt angle θ (i.e., $H_{\theta_{i,z}}^{\theta_{i,z'}} = H_{\theta_{k,z}}^{\theta_{k,z'}}$, for all pan/tilt angle θ_i and θ_k). We apply this feature of multilayer PTZ-

ALGORITHM 1: The algorithm developed for multilayer PTZ camera calibration

```
Input:  The SIFT feature points determined from the reference image
Output: The homographies between distinct layers
for number of layers do
    for number of feature points do
        To calculate the number of feature points, $N$, which can be viewed using Layer $i$
        based on this feature point.
        if $N$ is the most then
            To set the position of this feature point to be the calibration position ($p$, $t$) for
            this Layer $i$
        end
    end
    To capture images for the Layers $i$ and $i+1$ based on the position ($p$, $t$) and determine
    the feature points on these images. To calculate the homography between these two
    layers according to these feature points
end
```

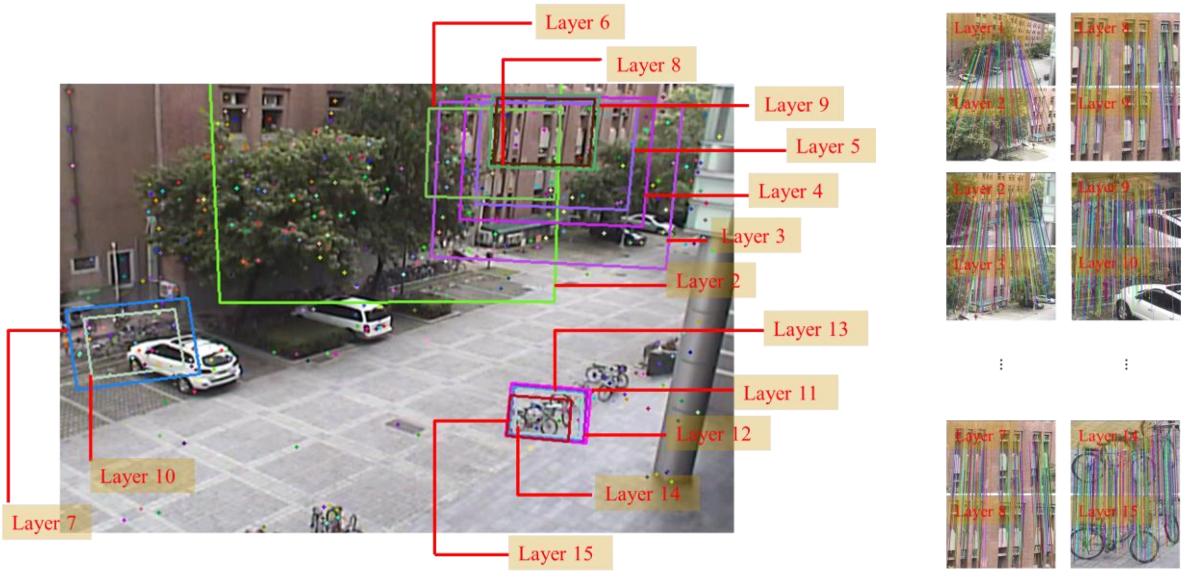
camera calibration to both in-factory and on-site calibration.

In on-site calibration, we detected the feature points in the reference image by using the SIFT [74] method. To overcome the noise problem caused by moving objects, we used a series of images captured by the PTZ camera as the reference view at the pan/tilt angle $\theta_0 = [p_0, t_0]$, and then determined the corresponding points based on this image series to identify the feature points that are fixed in all images. After identifying the stable feature points, we determined the optimal regions in the reference image in which the maximal numbers of feature points were covered, calibrating the coordinate transformation between Layers i and $i+1$ by using Algorithm 1.

The calibration regions between various layers can be distinct. Because of the consistency of distinct pan/tilt angles at various zooming factors, we can determine the corresponding feature points between various layers by using distinct views. Figure 3.10 shows the feature points in the reference image and the colored rectangles indicate the optimal regions that were used for multilayer calibration in each layer.

3.4 Large-Area High-Resolution Visual Monitoring

After calibrating the camera, we obtain the transformation relationship between the wide-angle fixed



(a)

(b)

Figure 3.10: The feature points in the reference image and the correspondence between different layers.

camera and the PTZ camera, the relationship of the PTZ camera turning, and the relationship between the multiple layers of the PTZ camera. On the basis of these calibration results, we can construct a large-area high-resolution background image combined with the multiresolution foreground image.

3.4.1 Construction of High-Resolution Background Images

Constructing the large-area high-resolution background image involves determining the step of the biaxis over layers, histogram adjustment, and blending.

Determining the Step of the Biaxis Over Layer

A large-area high-resolution background image was constructed by integrating the detail images captured using the PTZ camera at each pan/tilt angle $\theta_i = [p_i, t_i]$. The quality of the integrating image depends on the percentage of overlap region between the images. When the overlap percentage is small, the blended region is small; thus the quality of the integrated image is high. Therefore, the first step is determining the step of the biaxial over layers to minimize the overlap region.

The expected result of this step is similar to the map of coordinates (Figure 3.8) produced in the

ALGORITHM 2: The step-determination algorithm of the biaxis over each layer

```
Input: $z_i, \mathcal{T}_{overlap}$
Output: $p_{step_i}$ and $t_{step_i}$ for distinct layers, i
Initial: $p_{max\_step} = p_{max} - p_{min}$ and $p_{min\_step} = 1$
while ($p_{max\_step} \neq p_{min\_step}$)
    $p_{avg\_step} = \left\lceil \frac{p_{max\_step} + p_{min\_step}}{2} \right\rceil$
    Initial: $t_{\mathrm{max}\_step} = t_{max} - t_{min}$ and $t_{min\_step} = 1$
    while ($t_{max\_step} \neq t_{min\_step}$)
        $t_{avg\_step} = \left\lceil \frac{t_{max\_step} + t_{min\_step}}{2} \right\rceil$
        $Tilt_{non-overlap} = \frac{Area_{HRB}}{\sum_k Region_k}$     $k$: the region number
        if $Tilt_{non-overlap} \leq \mathcal{T}_{non-overlap}$ && all pixel are fixed in then
            $t_{min\_step} = t_{avg\_step}$
        else
            $t_{max\_step} = t_{avg\_step} - 1$
        end if
    end
    $t_{max\_step} = t_{min\_step}$
    $t_{step_i} = t_{max\_step}$
    $Pan_{non-overlap} = \frac{Area_{HRB}}{\sum_k Region_k}$     $k$: the region number
    If $Pan_{non-overlap} \leq \mathcal{T}_{non-overlap}$ && all pixel are fixed in then
        $p_{min\_step} = p_{avg\_step}$
    else
        $p_{max\_step} = p_{avg\_step} - 1$
    end if
end
$p_{step_i} = p_{max\_step} = p_{min\_step}$
```

calibration-boundary determining procedure (Section 3.3.2). However, the map of coordinates cannot be applied because distinct layers must exhibit distinct steps for each axis, called p_{step_i} and t_{step_i}, to minimize the overlap regions. Therefore, we developed Algorithm 2 to determine the step of the biaxis over each layer by using p_{min} 、 t_{min} 、p_{max} and t_{max}, which were estimated as described in Section 3.3.2.

The input parameters in the algorithm are z_i and $\mathcal{T}_{non-overlap}$, where z_i is the zooming factor and $\mathcal{T}_{\mathrm{non-overlap}}$ is the percentage of the nonoverlapping region area determined by the user. In the proposed implementation, we set the value of $\mathcal{T}_{\mathrm{non-overlap}}$ as 0.9. The parameters p_{max_step}, p_{min_step}, and p_{avg_step} were the temporary variables used to determine the exact value of p_{step_i}

for Layer i and represent the maximal, minimal, and average pan steps in Algorithm 2. Similarly, the parameters, t_{max_step}, t_{min_step} and t_{avg_step}, were used to determine t_{step_i}. $Tilt_{non-overlap}$ is the ratio of $Area_{HRB}$ to $\sum_k Region_k$ and it is used to determine the step of the tilt. $Area_{HRB}$ is the area of the high-resolution background image and $\sum_k Region_k$ is the sum of the stitched areas of the detail image at each pan/tilt angle. $Pan_{non-overlap}$ has a similar meaning as $Tilt_{non-overlap}$ and it determines the step of the pan. We used Algorithm 2 to determine the step of p_{step_i} and t_{step_i} for each Layer i in the proposed system.

As the Algorithm 2 shows, the results of t_{step_i} and p_{step_i} are influence each other. For example, the optimal t_{step_i} can be identified that generates the smallest overlapping region in they-axis, but this may worsen $p_{step_i}$and cause a large overlapping region in the x-axis. To avoid this bias, we use the following equation:

$$\frac{Tilt_{non-overlap}}{Pan_{non-overlap}} \geq 0.8, \tag{3.7}$$

which balances the estimates p_{step_i} and t_{step_i}. When the ratio of $Tilt_{non-overlap}$ to $Pan_{non-overlap}$, is closer to 1, the overlap region of the biaxis is similar..

Histogram Adjustment

After stitching the images of the PTZ camera by using distinct pan/tilt angles, the image can appear a mosaic (Figure 3.11). The mosaic phenomenon is caused by the variation in the illumination of distinct regions captured by the PTZ camera at various pan/tilt angles or by the function of the camera's automatic-gain and white-balancing circuits; for example, the regions in the red and blue boxes in Figure 3.11 are illuminated dissimilarly. To solve this problem, two image-processing methods are used: "histogram adjustment" and "blending."

Histogram adjustment applies a histogram-matching step, converting the target source image to the specified-histogram image. The proposed system features two choices of the specified-histogram image: the overview image or the reference image. In the proposed implementation, we selected the

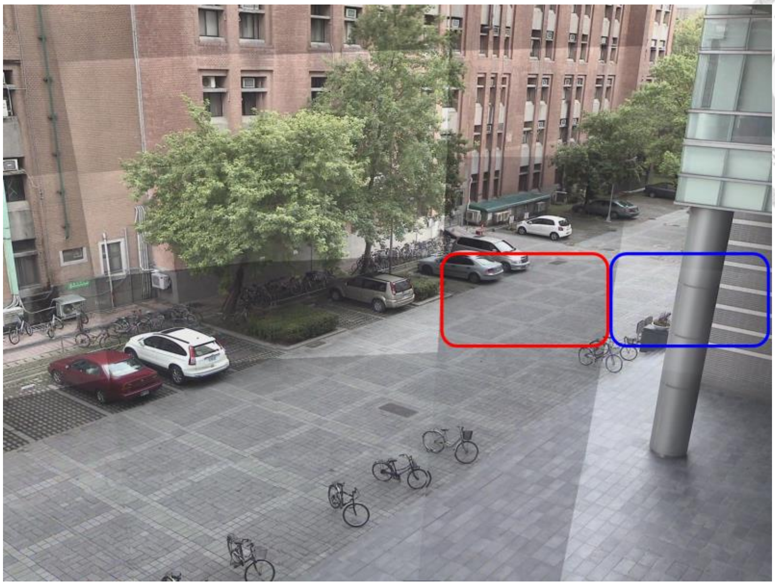
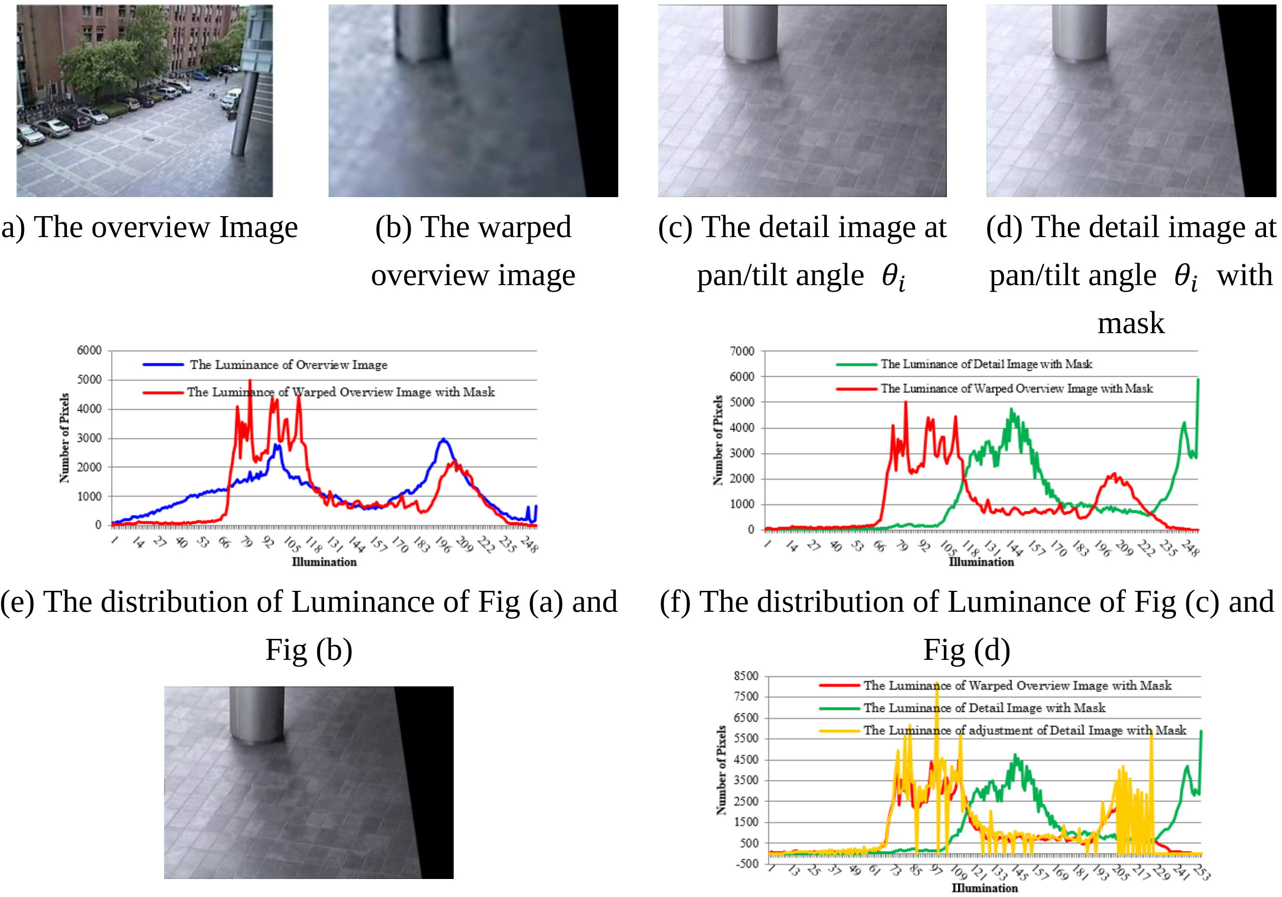


Figure 3.11: The mosaic phenomenon.

overview image as the specified histogram image during the histogram-adjustment process because the image captured using the wide-angle fixed camera was constantly updated and effectively represented the changes in the monitored environment.

The histogram adjustment is region based in the proposed system. We warped the overview image into the detail images and matched the histograms of the detail images with the coordinating regions of the warped overview image. This is similar to splitting the overview image into small regions and matching them to the corresponding detail images. Figure 3.12 shows an example of histogram adjustment, where Figure 3.12 (a)(c) present the overview and detail images at pan/tilt angle θ_i. The warped overview image may include some black areas, as shown in Figure 3.12 (b). We use the black area as a mask to make the detail image keep consistent with the warped overview image, as shown in Figure 3.12 (d). Region-based adjustment is more effective compared with using entire- image adjustment for two reasons. First, the illumination of the detail image is similar to that of the corresponding region in the overview image. An example is presented in Figure 3.12 (f), in which the red and green curves indicate the distribution of luminance of the warped overview image and the detail image, respectively. The distribution shapes of the images are similar. The second



(a) The overview Image

(b) The warped overview image

(c) The detail image at pan/tilt angle θ_i

(d) The detail image at pan/tilt angle θ_i with mask

(e) The distribution of Luminance of Fig (a) and Fig (b)

(f) The distribution of Luminance of Fig (c) and Fig (d)

(g) The Image of Adjustment Result

(h) The Histogram of Adjustment Result

Figure 3.12: An Example of Histogram Adjustment.

reason is that illumination is highly dissimilar in the entire overview image and the corresponding overview image region as shown in Figure 3.12 (e), in which the blue curve represents the entire overview image. This example shows that the luminance distribution of the entire overview image cannot accurately represent the illumination condition of each region of the overview image.

Figure 3.12 (g) shows the image obtained after histogram adjustment and Figure 3.12 (h) presents the distributions of the luminance of the images in Figure 3.12 (b), (d), and (g). The yellow curve represents the result of histogram adjustment of the image in Figure 3.12 (g).

Blending in Large-Area High-Resolution Background Images

After histogram adjustment, we compensated for the disparities in the illuminations of distinct regions; however, the mosaic phenomenon can still appear at image boundaries, as shown in Figure

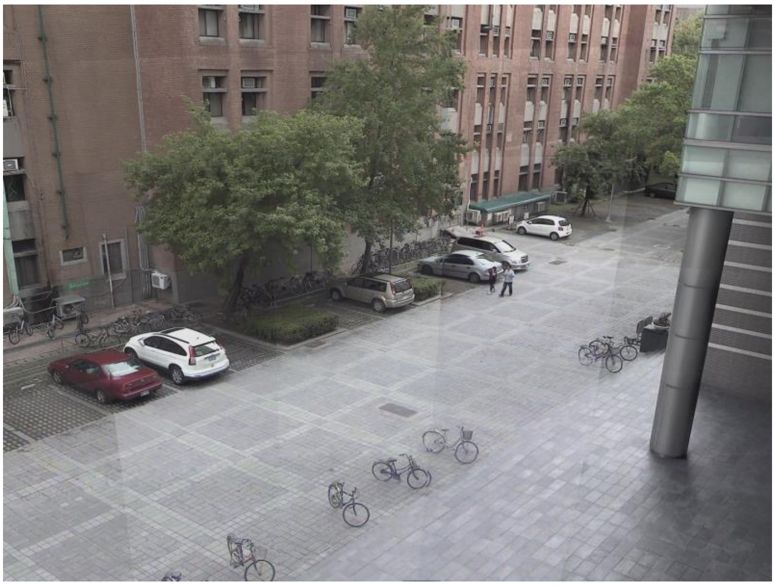


Figure 3.13: The result after histogram adjustment.

3.13. Therefore, we developed a blending algorithm, Algorithm 3. Over the past two decades, numerous algorithms have been developed to blend images, such as the algorithm of [76] in which the Poisson equation was used, and that of [77] in which watersheds and graph cuts were used; by using both of these algorithms, images have been effectively blended. However, such methods require substantial amounts of processing time. Thus, in this study, we adopted the alpha blending (feathering) method, which is faster compared with the aforementioned methods and more suitable for use in visual- surveillance systems.

In the blending process, we design a weighted map, $\mathcal{W}$, as shown in Figure 3.14. The weighted

ALGORITHM 3: The blending algorithm

```
Input: Each region that is viewed at the pan/tilt angle $\theta_i = [p_i, t_i]$
Output: Large-area and high-resolution image blending of $I_{HRBG_{z_k}}$
To calculate the weighted image $\mathcal{W}$
for each region that is viewed at $\theta_i$ do
    To calculate the warped weighed image $\mathcal{W}_{w_{i,z_k}}$
end
for each region that is viewed at $\theta_i$ do
    To calculate the alpha map $M_{\alpha_{i,z_k}}$
end
for each region that is viewed at $\theta_i$ do
    To calculate the final image $I_{HRBG_{z_k}}$
end
```

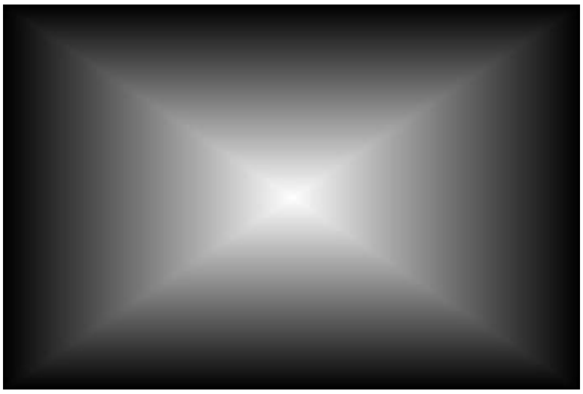


Figure 3.14: The Weighted Map.

map $\mathcal{W}$ is calculated according to the following equation, which is used to assign larger blending weights to the points close to the center compared with the points far from the center:

$$\mathcal{W}(x,y) = min\left(1-\frac{|2x-width|}{width}, 1-\frac{|2y-height|}{height}\right) \tag{3.8}$$

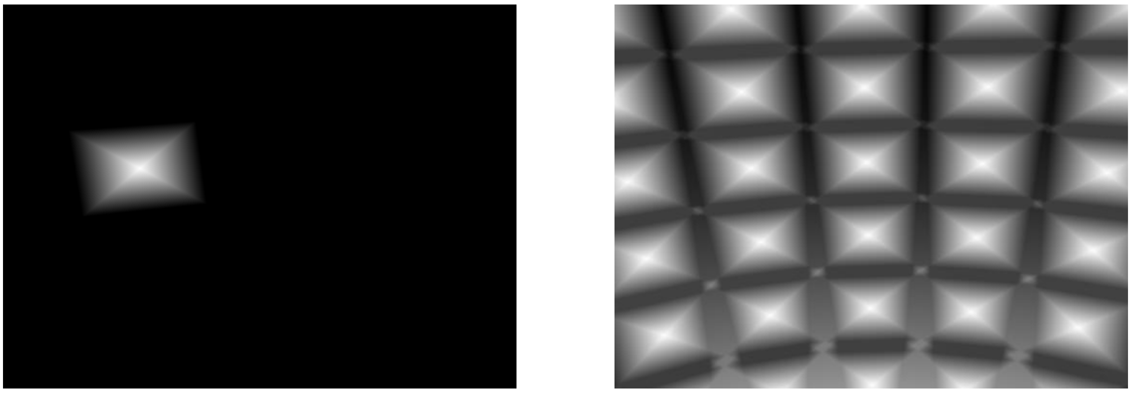
where width and height are the width and the height of the weighted map, respectively. Thus, the warped weighted map $\mathcal{W}_{w_{i,z_k}}$ is calculated for the wide-angle view at each pan/tilt angle $\theta_i = [p_i, t_i]$ by using the following equation,

$$\mathcal{W}_{w_{i,z_k}}(x,y) = \mathcal{W}\left(\left(H_{OI}^{Ref}\cdot H_{Ref}^{DI_{\theta_{i,z_1}}}\cdot H_{DI_{\theta_{i,z_1}}}^{DI_{\theta_{i,z_k}}}\right)^{-1}\begin{bmatrix}x\\y\\1\end{bmatrix}\right) \tag{3.9}$$

where the parameters H_{OI}^{Ref}, $H_{Ref}^{DI_{\theta_{i,z_1}}}$, and $H_{DI_{\theta_{i,z_1}}}^{DI_{\theta_{i,z_k}}}$ are the homographies between the overview image and the reference image, the reference and detail images at pan/tilt angle $\theta_i = [p_i, t_i]$ captured at Layer 1, and between distinct layers. Figure 3.15 (a) presents an example of the translation of one of the weighted map, $\mathcal{W}$, at pan/tilt angle $\theta_i = [p_i, t_i]$ to the overview image. The sum of the warped weighted map $\sum_i \mathcal{W}_{w_{i,z_k}}$ is shown in Figure 3.15 (b).

Next, the alpha map $M_{\alpha_{i,z_k}}$,, which is at the pan/tilt angle $\theta_i = [p_i, t_i]$ with the use of Layer i can be calculated using the following:

$$M_{\alpha_{i,z_k}}(x,y) = \frac{\mathcal{W}_{w_{i,z_k}}}{\sum_i \mathcal{W}_{w_{i,z_k}}}(x,y) \tag{3.10}$$



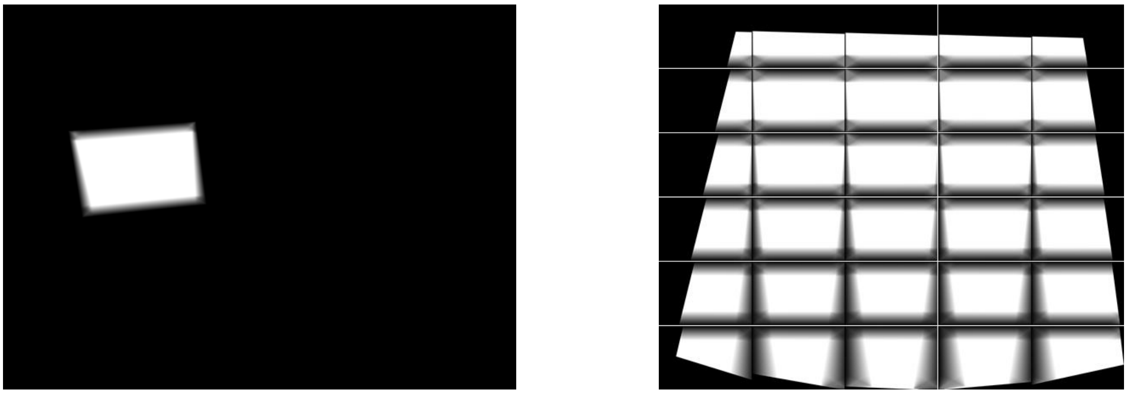
(a) The warped weighted map (b) The sum of the warped weighted map

Figure 3.15: An example of the warped weighted image.

An example of the estimated alpha map is shown in Figure 3.16 (a). Figure 3.16 (b) presents an example of the sum of the alpha map, comprising 30 pan-tilt positions determined using the following equation (the widths and heights of the subimages in Figure 3.16 (b) are the same as those of the detail image):

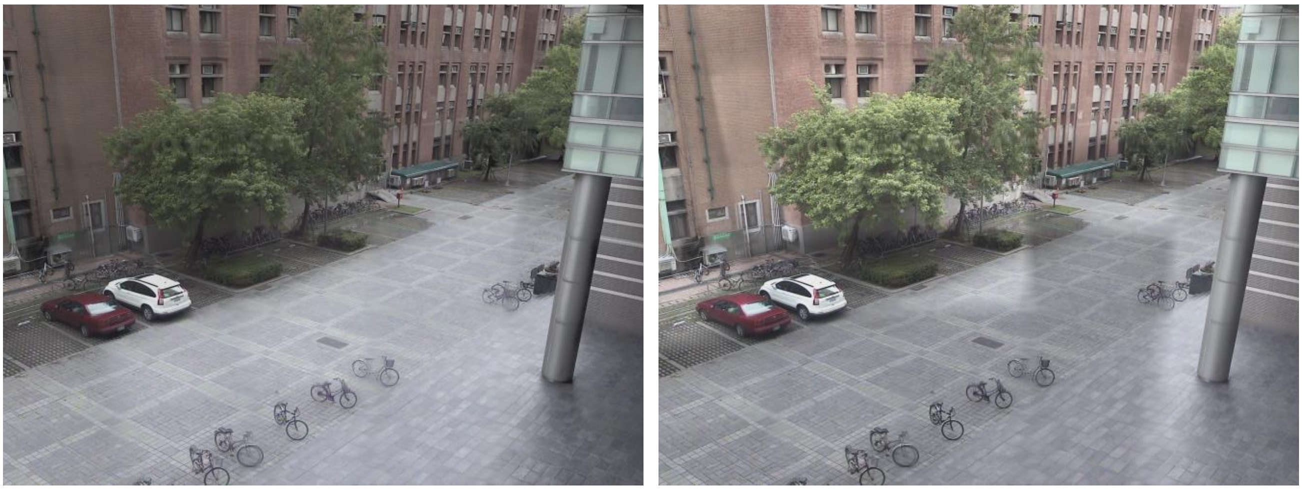
$$M_{\alpha_{i,z_k}}(x,y) = \frac{w_{w_{i,z_k}}(x,y)}{\sum_i w_{w_{i,z_k}}} \left(\left(H_{OI}^{Ref} \cdot H_{Ref}^{DI_{\theta_{i,z_1}}} \cdot H_{DI_{\theta_{i,z_1}}}^{DI_{\theta_{i,z_k}}} \right) \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \right) \tag{3.11}$$

Finally, the detail image at each pan/tilt angle $\theta_i = [p_i, t_i]$ was transformed to the overview image and multiplied by the alpha map, yielding the high-resolution background image I_{HRBG,z_k} by using the following equation:



(a) The alpha map (b) The sum of the alpha map

Figure 3.16: An example of the alpha map.



(a) The result of the blending after the histogram adjustment

(b) The result of the blending without first applying the histogram adjustment

Figure 3.17: The result of the large-area high-resolution background image.

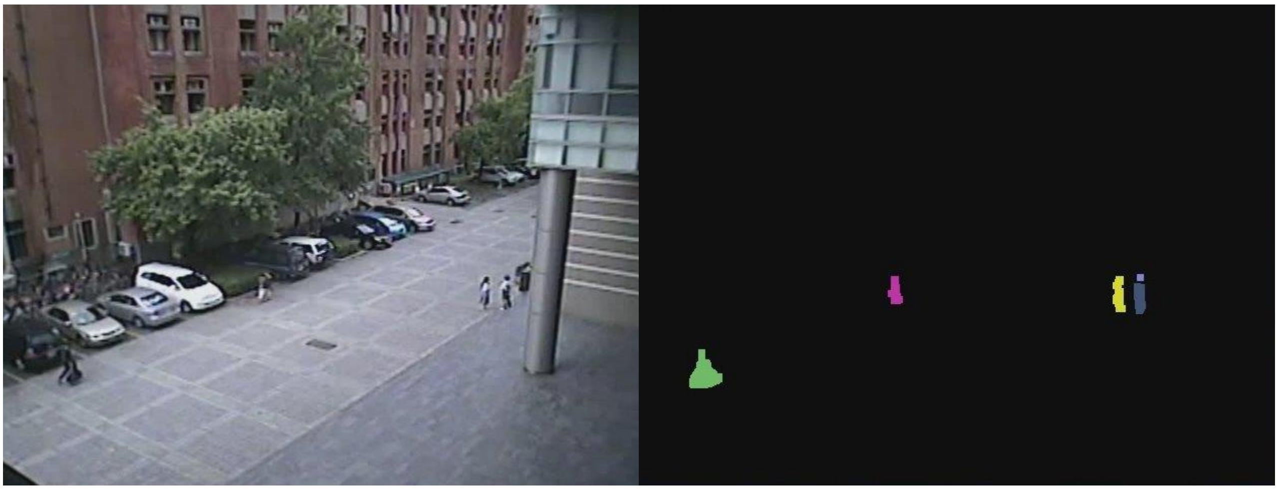
$$I_{HRBG,z_k} = \sum_i \left(M_{\alpha,i,z_k} \cdot DI_{\theta_{i,z_k}}\right)\left(\left[H_{OI}^{Ref} \cdot H_{Ref}^{DI_{\theta_{i,z_1}}} \cdot H_{DI_{\theta_{i,z_1}}}^{DI_{\theta_{i,z_k}}}\right]^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}\right) \tag{3.12}$$

Figure 3.17 shows the obtained large-area high-resolution background image; Figure 3.17 (a) presents the results of blending after applying histogram-adjustment step, showing that the mosaic phenomenon evident in Figure 3.11 has been eliminated.

Figure 3.17 (b) shows the results of using the blending algorithm without first applying the histogram adjustment. The various regions captured using the PTZ camera are almost all brighter compared with the regions in the overview image, suggesting that the image in Figure 3.17 (b) is of higher quality compared with that in Figure 3.17 (a). Comparing with Figure 3.17 (a) shows that although the boundaries of the adjacent images are smooth, the same mosaic phenomena exist because the illuminations of distinct regions are clearly dissimilar. Figure 3.17 shows that the histogram must be adjusted before the blending step to synthesize a high-quality high-resolution background image.

3.4.2 Multiresolution Foreground Images

A key feature of the proposed system is the multiresolution display of the foreground. To produce multiresolution foreground images, the proposed system uses two technical components: low-



(a) The overview image (b) The result of foreground image

Figure 3.18: The result of object detection.

resolution foreground detection and high-resolution foreground tracking.

Low-Resolution Foreground Detection

During the low-resolution foreground-detection step, we detected foreground objects in the overview images. The moving objects were detected using the mixture of Gaussian (MoG) approach [8], which is a classic background-subtraction method, in which each pixel is labeled as a background or foreground pixel comprising *k* Gaussian distributions. Using the MoG approach, moving objects are detected in the overview image. An example of foreground detection is shown in Figure 3.18, in which Figure 3.18 (a) presents the overview image, and Figure 3.18 (b) shows the foreground-detection result obtained using that image.

Subsequently, we attached all the foreground objects at a low resolution; these were scaled according to the zooming factor of the PTZ camera, to the large-area high-resolution background image according to the positions using the object-detection procedure (Figure 3.19).

High-Resolution Foreground Tracking

The relationship between the overview and detail images was calculated as described in Section 3.3. After selecting a target of interest in the large-area high-resolution background image containing low-resolution foreground objects, we controlled the PTZ camera to continuously track that target

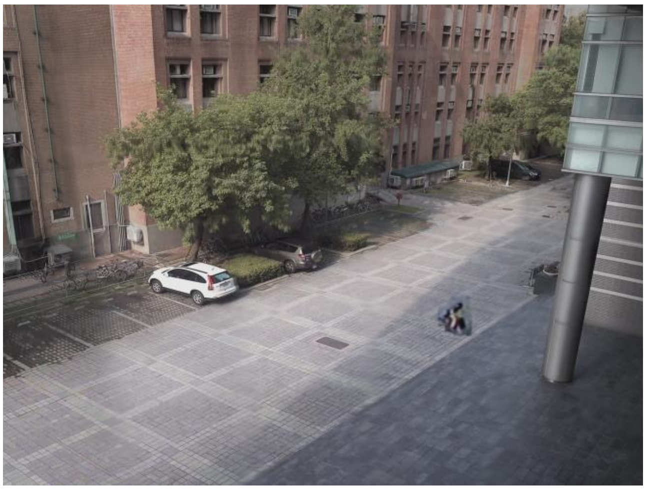


Figure 3.19: The large-area high-resolution background image with low-resolution foreground objects.

according to the transformation of the coordinates.

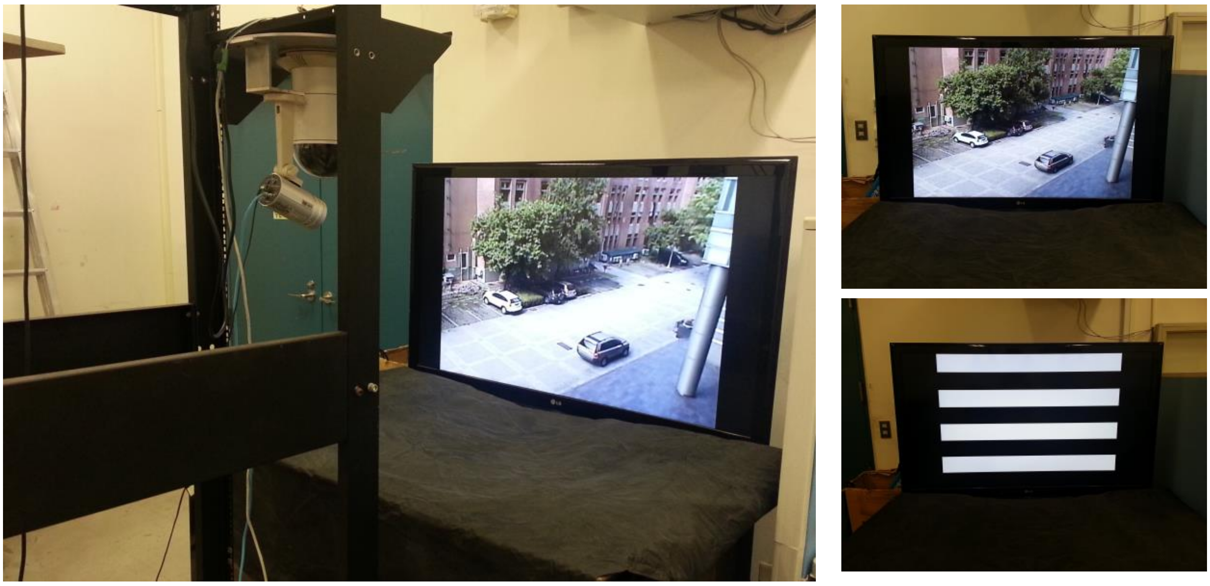
Given arbitrary pan/tilt coordinates, the corresponding homographies can be estimated by using the interpolation method. The region in which the target of interest is continuously tracked by capturing detail images is called the “fovea region“ in the proposed system. The automatic tracking is based on the results of foreground detection results and the velocity of the moving target [8]. Last, we obtained the high-resolution texture of the foreground object of interest from the detail image and pasted it onto the large-area high-resolution background image.

3.5 Experiments

In this section, we first describe the experiments that were used to analyze the accuracy of PTZ-camera calibration by comparing the in-factory and on-site calibration methods (described in Section 3.3.2) in both indoor and outdoor environments. After the analysis, we present the results obtained using the proposed large-area high-resolution visual-monitoring system.

3.5.1 Experimental Analysis of PTZ-Camera Calibration

First, we compared the two PTZ-camera calibration methods described in Section 3.3.2: in-factory and on-site calibration. In the experiments, both indoor and outdoor environments were used. In the



(a) The setting of indoor environment (b) The patterns

Figure 3.20: The indoor experimental environment.

indoor environment, the patterns used for camera calibration were displayed on an LED screen (Figure 3.20). Figure 3.20 (a) shows the setting of the indoor experimental environment; the distance between the LED screen and the PTZ camera was approximately 1 m. The upper image in Figure 3.20 (b) shows the pattern used for on-site calibration, and the lower image shows the gray-code patterns used for in-factory calibration. In the outdoor experimental environment, we installed the PTZ camera on the third floor of a building as shown in Figure 3.1, and the image in Figure 3.4(right) shows the view captured using on-site calibration.

The in-factory calibration method can only be used in indoor environments because the calibration patterns are manually produced and a screen or a projector is required. The on-site calibration process can be used both indoors and outdoors. Therefore, we evaluated three calibration cases: in-factory calibration applied in the indoor environment (indoor in-factory calibration), on-site calibration applied in the indoor environment (indoor on-site calibration), and on-site calibration applied in the outdoor environment (outdoor on-site calibration).

We evaluated each calibration case in the indoor and outdoor environments by generating measuring points to calculate the calibration error. To uniformly generate the measuring points, we

Table 3.1: The error analysis of the in-factory calibration.

The indoor in-factory calibration						The outdoor in-factory calibration	
Outdoor error analysis			Indoor error analysis			Outdoor error analysis	Indoor error analysis
Zoom	Error (pixel)	Std. (pixel)	Zoom	Error (pixel)	Std. (pixel)	✕	✕
1x	6.1	2.2	1x	1.9	1.0		
5x	32.2	12.4	5x	8.4	4.7		
10x	55.2	24.6	10x	14.1	6.9		
(a)			(b)			(c)	(d)

Table 3.2: The error analysis of the on-site calibration.

The indoor on-site calibration						The outdoor on-site calibration					
Outdoor error analysis			Indoor error analysis			Outdoor error analysis			Indoor error analysis		
Zoom	Error (pixel)	Std. (pixel)	Zoom	Error (pixel)	Std. (pixel)	Zoom	Error (pixel)	Std. (pixel)	Zoom	Error (pixel)	Std. (pixel)
1x	7.4	2.5	1x	2.0	0.9	1x	1.8	0.8	1x	5.3	2.1
5x	29.3	12.3	5x	9.5	4.4	5x	8.0	4.1	5x	22.6	8.6
10x	54.2	25.0	10x	15.5	9.1	10x	14.0	8.3	10x	41.3	16.6
(a)			(b)			(c)			(d)		

split the reference image into a 4×8 grid (Figure 3.21) and manually specified the measuring point in each grid, shown as the colored points, allowing users can readily recognize the measuring points in the reference image. Using the calibration results, the PTZ camera was controlled to focus on the selected point, as shown in Figure 3.22; the hollow circle is the ground truth that is defined by the user and the solid circle is the center of the detail image. The distance between these points represents calibration error.

Table 3.1 presents the error-analysis results of in-factory calibration. Table 3.1 (a) lists the results of the outdoor test environment and Table 3.1 (b) lists the results of the indoor environment. The errors in Table 3.1 (b) are smaller than those in Table 3.1 (a) are because the calibration environment and the test environment were not the same and the turning of the PTZ camera was not exactly concentric. Table 3.1 (c) and Table 3.1 (d) contain no results because in-factory calibration cannot be

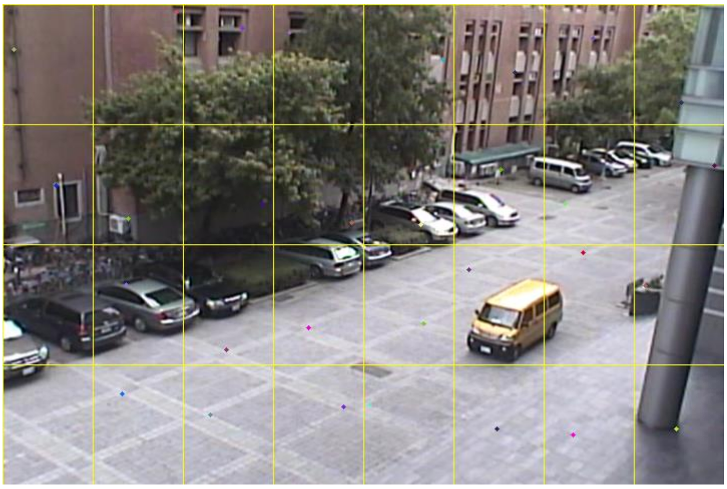


Figure 3.21: The grid of the reference image.

used in the outdoor environment. The results in Table 3.1 show that as the magnification factor increased, the error increased. However, the accuracy was adequate for surveillance applications because the target of interest could be monitored even at a 10x zooming factor. Table 3.2 presents the error-analysis results of on-site calibration. The calibration and test environments used to obtain the results shown in Table 3.2 (c) were outdoors and those shown in Table 3.2 (b) were indoors. The results in Table 3.2 (b) and (c) show that if the calibration and test environments are the same, the error of the testing in both indoor and outdoor environments is smaller and more similar than it is when the calibration and test environments are distinct. Based on these experiments, we can conclude that because the PTZ camera is not exactly concentric when the calibration and the test environments are the same, the high levels of accuracy of calibration results are achieved. The proposed on-site

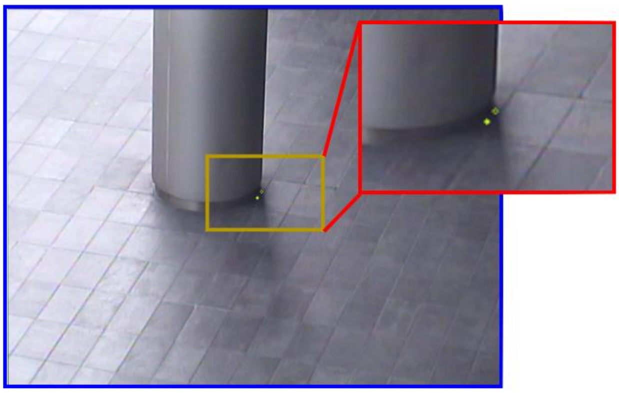


Figure 3.22: An Example of the accuracy analysis.

calibration method exhibits high levels of accuracy in both indoor and outdoor environments; thus, it is considered more effective for calibrating the proposed camera system compared with the in-factory calibration method [35] [42].

3.5.2 System Demonstration

We installed a large-area high-resolution visual-monitoring system on the third floor of a building to monitor a courtyard (Figure 3.1). The targets of interest can include pedestrians or vehicles crossing the courtyard or other user-chosen targets.

Large-Area High-Resolution Background Images

In the proposed monitoring system, a large-area high-resolution background containing distinct layers can be generated using the methods described in Section 3.4.1. Images containing moving objects can be generated by collecting the areas of detail images that display moving objects and stitching these to the low-resolution images captured using the wide-angle camera. After initialization, the system updates the background even when users do not control the PTZ camera or when no moving objects are present in the monitored area.

The synthesized large-area high-resolution background images were created using the 4x and 15x zooming factors of the PTZ camera, and the resolutions were approximately 4320 × 2880 and 11520 × 8160, respectively. In the images captured using the proposed system, users can see details of the monitored scene such as the license plates of parked cars.

Large-Area High-Resolution Background and Multiresolution Foreground Images

After a user selects a target of interest, the PTZ camera ceases updating the background and begins tracking the target. Figure 3.23 shows an example image captured using the proposed system after a user selected an object to be tracked. We did not blend the online high-resolution image with the preconstructed background image because a user can easily distinguish the offline high-resolution background from the online high-resolution region. After generating the large-area high-resolution

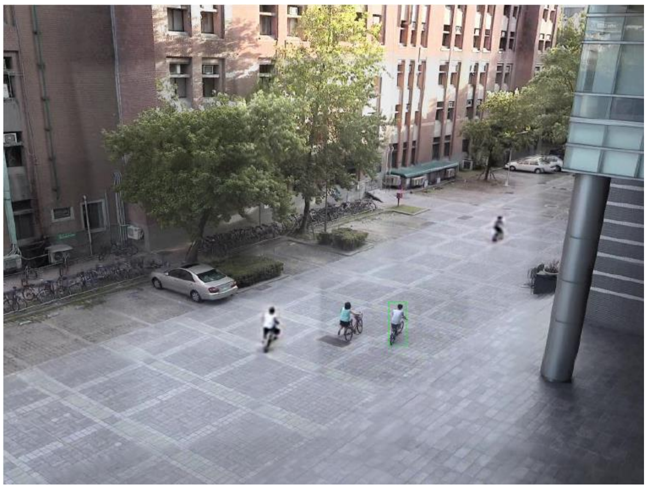


Figure 3.23: The result of the object tracking with the fovea region.

background, the system begins to update the background or track the target by capturing high-resolution images; the proposed system is effective because only one patch of the monitoring view is required to update.

As Figure 3.23 shows, in addition to the high-resolution region, moving objects are pasted at low resolution in the image that allowing users to follow. Additional details of this imaging method can be seen in this demonstration video: http://youtu.be/HbQZ6f1qxfk. In the video, the stitching results obtained without using histogram adjustment and blending are first shown, and both histogram adjustment and blending are subsequently applied. Finally, the entire system is displayed. After generating the large-area high-resolution background, we tracked the target of interest within the fovea region through distinct layers.

3.6 Summary

In this chapter, we designed a framework by using a dual-camera to construct a large-area high-resolution visual-monitoring system. We proposed two PTZ camera calibration methods: in-factory calibration and on-site calibration. The in-factory calibration method is useful when stable features are extremely difficult to obtain. The on-site calibration method is more flexible than in-factory calibration method is because it can be applied both indoors and outdoors. To perform multilayered

PTZ-camera calibration, the consistent property of distinct pan-tilt angles at various zooming factors is used, allowing the calibration between each layer to be reduced to only calibrating a pair of images in distinct layers at an arbitrary pan/tilt angle. This consistency substantially improves the proposed calibration method, and a literature review indicates that no scholars have previously used this property to calibrate PTZ cameras.

Comparing the PTZ camera calibration with our previous work [35] [42], the pros of the proposed PTZ camera turning calibration are automatic processing, without time-consuming, could be calibrated on site, be re-calibrated on site without manual and without decay the accuracy of calibration results. To compare the multilayered PTZ-camera calibration method, the proposed method is faster than our previous work [35] [42].

We also propose a framework that using the calibration results of two cameras to construct the proposed visual-monitoring system, combining the preconstructed large-area high-resolution background image with online multiresolution foreground images. We demonstrated the robustness of this camera-calibration method by using quantitative experiments and the practical use of this system by installing it in an outdoor environment.

CHAPTER 4

SPATIAL BACKGROUND MODEL USING A SINGLE-LAYER CODEBOOK MODEL

4.1 Introduction

The number of cameras in cities is getting increased in recent years, and thus there is a growing need of constructing intelligent visual surveillance system based on cameras. There are numerous research topics in association with visual surveillance, for example vehicle classification, objects stolen, objects left, and objects classification. In which, segmenting moving objects from the scene is a relevant issue in those research areas. Moving objects can be effectively extracted by comparing with the background model.

Background subtraction can be defined as a separation of significant differences (foreground) inside of the video frame from the non-significant components (background). The problem is relatively simple when there is a significant difference between the colors (or gray levels) of foreground and background, and is also easier to handle when the background is kept static. According to the environment situation, the background can be classified as the stationary background also called the static background and the non-stationary background also called the dynamic background.

For static background, it is appropriate for the constrained indoor environment, because of the camera jittering and signal noise caused from the outdoor environment may still cause non-stationary problems in most practical situations. The non-stationary backgrounds, such as waving leaves, ripple water, fluorescent light, monitor flicker, are one of the difficulty problems in background modeling in practice.

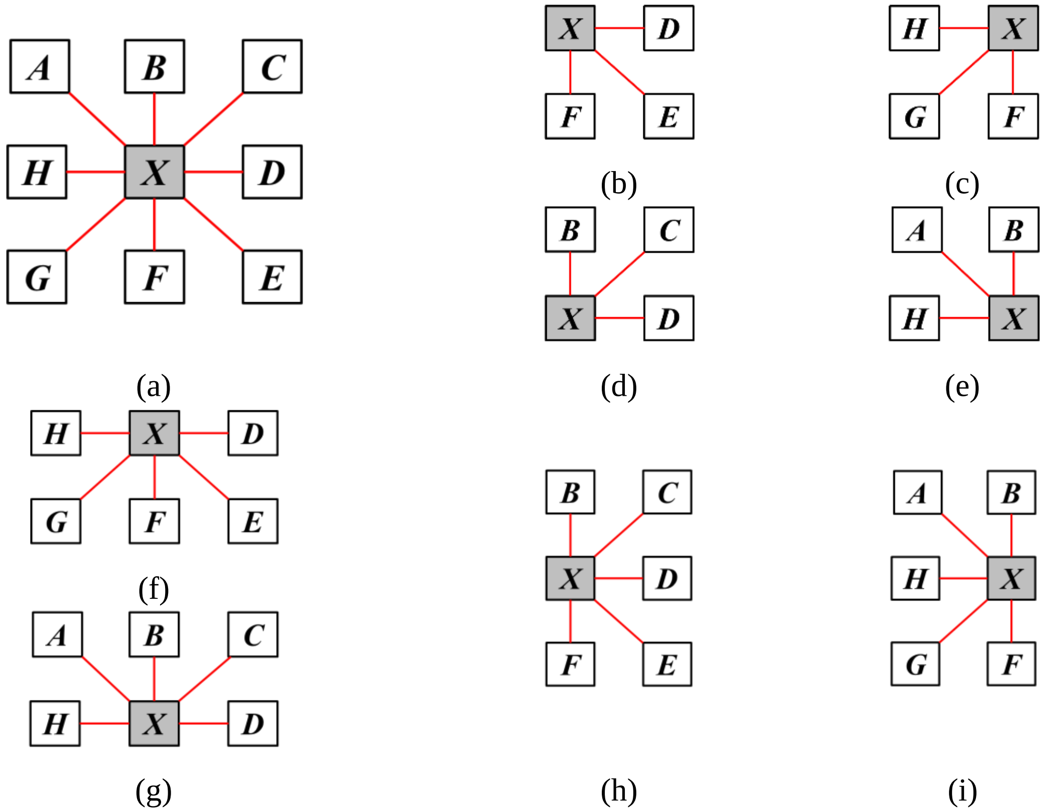
Moreover, the initialization is the other important issue in background modeling. Most of the algorithms need a period of the clean frames at the beginning of the sequence to build the background model; however, this requirement is difficult to be satisfied in practice, because of moving objects. When a static object start to move in the scene at the beginning of video, the ghost effect likes a hole is the background may appears and the detected object is broken as hole inside of the detected region.

In this chapter, we present a background subtraction algorithm, spatial background model (SBM), for detecting moving objects from a non-stationary background and that is able to cope with the effect caused by initiating of background modeling. The rest of this chapter is organized as follows. First, we focus on the spatial background model that includes the background model and the background gradient extractor. Then, we show the experiment results and conclusions.

4.2 Spatial Background Model (SBM)

The spatial background model (SBM) includes two fundamental components: the background model (BM) and the background gradient extractor (BGE) for foreground detection strategy is proposed, in which the structure of the codebook is involved in the component of BM. The proposed codebook in BM is similar to the former codebook [2] [9], but a single layer of codebook is adopted. To our observation, the updating strategy of background model in the former codebook is high efficiency; however, it still involves many redundancies. To ease this problem, the random technique of ViBe [10] [56] is adopted in SBM for preserving the advantage of codebook, and then further promotes the accuracy of the classification of foreground and background.

Although the performance of estimating the noise in the background in BM is well, the foreground become incomplete. The gradient information of the foreground is adopted in BGE to



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 4.1: The chart of connected neighborhoods around the current pixel ***X***. (a) the current pixel is inside the image. (b)-(e) the current pixel ***X*** is at the corner of the image.(f)-(i) the current pixel ***X*** is at the border of the image.

promote the completeness of the foreground.

4.2.1 Background Model (BM)

The background model (BM) is constructed as a single-layer codebook model refer to [2] [9]. Differing from [2] [9] that constructs a background model from a period of video sequence, we only need a single frame to construct BM and take the information from the neighbors as the features for each pixel in BM.

Features Used in Background Model

Traditionally, the background model constructed with codebook extracts the features for each pixel in background model utilizes the continuous property in time axis which is time consuming [2] [9] [54] [55] and that cannot effectively estimate the problem of the non-stationary background. In order to reduce the time of training for constructing background and solve the problem of non-stationary background model. In this study, we adopt a single frame to construct the background model which efficiently solve the problem of time-consuming.

Although, using a single frame to construct the background model can save the time of training, the information for each pixel in background model is not sufficient. Thus, we adopt the RGB color information extract from the eight-connected neighbors to be the features for each pixel in background model as shown in Figure 4.1. In Figure 4.1, the current pixel in background model is referred to as ***X***. The number of neighbors around ***X*** is dependent on its location. In Figure 4.1 (a), as the current pixel ***X*** is inside the image and the neighborhoods ***A***-***H*** are adopted. Figure 4.1 (b)-(e) and (f)-(i) illustrate the other cases of positions of the current pixel ***X*** when it is located at the corners or the border of an image. The blue, green, red, and illumination are took as the features used in background model. In addition, the features obtained from the neighbors can effectively solve the problem of non-stationary background.

Construction of Background Model

We use the structure of codebook to store the information of background model, slightly modified from the version presented by Kim *et al.* [2] [9] to perform background subtraction in the color domain. The BM algorithm is designed for color imagery and each pixel x_p in background model is modeled as a single-layer codebook $\mathcal{C}_{x_p} = \{c_i | 1 \leq i \leq n\}$ consisting of n codewords, in which p is the position of that pixel and t is the number of frame. Each codeword c_i is composed of two elements, a RGB vector $v_i = (\bar{R}_i, \bar{G}_i, \bar{B}_\imath)$ and a six-tuple $aux_i = \langle \check{I}_i, \hat{I}_i, f_i, \lambda_i, p_i, q_i \rangle$. Where $\check{I}_i$ and $\hat{I}_i$ are the minimum and maximum brightness of each codeword c_i, respectively, f_i is the occurred frequency, λ_i is the longest time interval that the codeword c_i is **NOT** recurred, and p_i and q_i are the first and last access times, that the codeword has occurred, respectively. The algorithm of constructing background model is shown in Algorithm 4.

Algorithm 4. The Contraction of Background Model

```
Input: The first frame of image sequence
Output: The background model
$\mathcal{C} \leftarrow \emptyset$ (empty set)
for each pixel in background model    do
        To construct the codebook $\mathcal{C}_{x_p} = \{c_i | 1 \le i \le \mathrm{n}\}$
        according to the adjacent neighborhoods and each
        codeword $c_i$ is composed of $v_i$ and $aux_i$
        for neighbor $i = 1$ to $k$ do
                $I_{x_i^{t=1}} = \sqrt{R^2_{x_i^{t=1}} + G^2_{x_i^{t=1}} + B^2_{x_i^{t=1}}}$
                $v_i = (R_{x_i^{t=1}}, G_{x_i^{t=1}}, B_{x_i^{t=1}})$
                $aux_i = (\check{I}_i = I_{x_i^{t=1}}, \hat{I}_i = I_{x_i^{t=1}}, f_i = 1, \lambda_i = 1,\ p_i = 1, q_i = 1)$
        end
end
```

In BM, the first frame of the image sequence is used to construct the background model and each codebook $\mathcal{C}_{x_p}$ is constructed base on a set of neighbors. As shown in Figure 4.1, the number of codewords in each codebook $\mathcal{C}_{x_p}$ in the initialization is depending on its location. For example, In Figure 4.1 (a), current pixel ***X*** is inside the image and the neighborhoods ***A-H*** are adopted. Figure 4.1 (b)-(e) and (f)-(i) illustrate the other cases of positions of the current pixel ***X*** when it is located at the corners or the border of an image. We extract the features from the neighbors for each codebook according to the following equations,

$$I_{x_i^{t=1}} = \sqrt{R^2_{x_i^{t=1}} + G^2_{x_i^{t=1}} + B^2_{x_i^{t=1}}}$$

$$v_i = (R_{x_i^{t=1}}, G_{x_i^{t=1}}, B_{x_i^{t=1}}) \tag{4.1}$$

$$aux_i = (\check{I}_i = I_{x_i^{t=1}}, \hat{I}_i = I_{x_i^{t=1}}, f_i = 1, \lambda_i = 1, p_i = 1, q_i = 1)$$

where $x_i^{t=1}$ is the pixel at i position on first frame of the image sequence, R, G, B are the red, green, and blue color information, and I is the illumination of that pixel.

According to the algorithm 4, the background model is built, and the number of each codebook is between 4 and 8. The BM algorithm is different from [2] [9]. The construction is faster because only a single frame is used. The neighborhoods contribute spatial information, which is reflected in

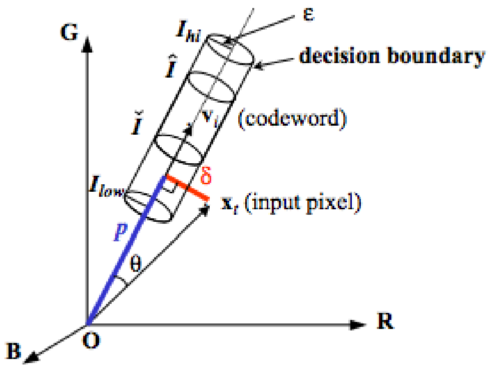


Figure 4.2: The cylinder color model proposed by [9]

the simplified single-layer codebook model and useful to overcome the non-stationary background problem. Notice that, because a single frame is used in this procedure, the minimum and maximum brightness of each codeword c_i, $\check{I}_i$ and $\hat{I}_i$, are set to be the same.

Pixel classification Policy

When background model is built, there are two procedures for incoming pixel x^t : pixel classification and update policies.

In pixel classification stage, the cylinder color model is adopted, as shown in Figure 4.2 and two measurements: the color distortion and the range of brightness, are used provided by [2] [9].

In color distortion, the matched codeword c_m with RGB vector v_m is found from its codebook $\mathcal{C}$ according to the following color distortion measure:

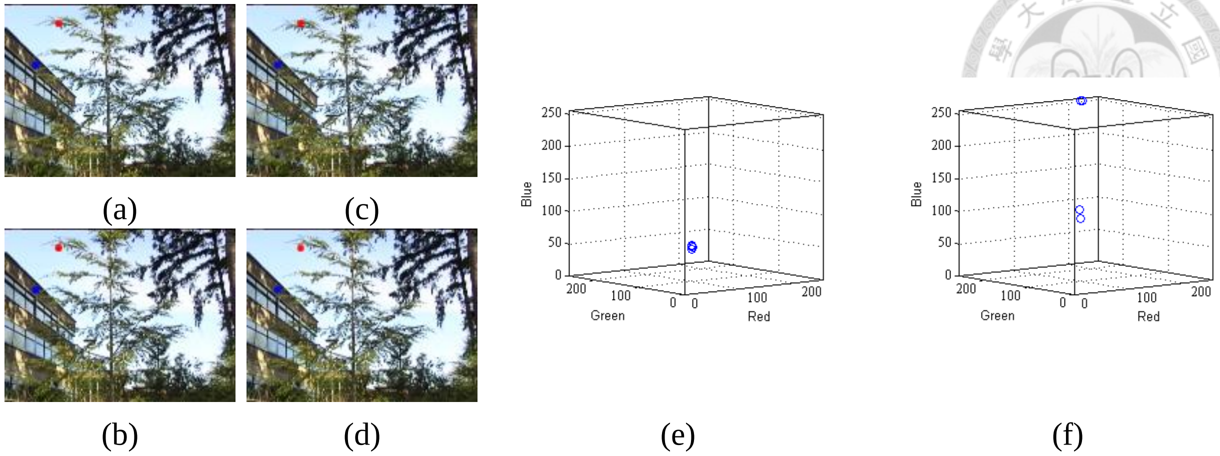
$$p^2 = \|x^t\|^2 \cos^2\theta = \frac{\langle x^t, v_i\rangle^2}{\|v_i\|} = \frac{(\bar{R}_i R + \bar{G}_i G + \bar{B}_i B)^2}{\|\bar{R}_i + \bar{G}_i + \bar{B}_i\|}, \text{and}$$

$$colordis(x^t, v_i) = \delta = \sqrt{\|x_t\|^2 - p^2} \leq \varepsilon, \tag{4.2}$$

where $x_t = (R, G, B)$ is the incoming pixel at time t with a RGB vector and ε is the threshold for color distortion.

To handle the changes of illumination, the brightness of the incoming pixel is considered:

$$brightness(I_{x^t}, \langle\check{I}, \hat{I}\rangle) = \begin{matrix} true & if\ I_{low} \leq I_{x^t} \leq I_{hi} \\ false & otherwise \end{matrix} \tag{4.3}$$



(a) (c)

(b) (d) (e) (f)

Figure 4.3: The temporal and spatial information. (a) WaveTree #36 (b) WaveTree #38 (c) WaveTree #40 (d) WaveTree #42 (e) the temporal distribution of the red pots in RGB color space (f) the spatial distribution of the blue pots in RGB color space.

where $I_{low} = \alpha\hat{I}$and $I_{hi} = min\left\{\beta\hat{I}, \frac{\hat{I}}{\alpha}\right\}$ are the lower and upper bounds of illumination, and I_{x^t} is the illumination of incoming pixel x^t. The parameters, α and β, are set as that in [2] [9] which are used to allow large brightness bound I_{hi} and limiting I_{low}, respectively. If the incoming pixel simultaneously satisfies the above two criteria about color distortion and brightness, it is classified as background.

The Update Policies of Background Model

When the incoming pixel is classified as background, it is updated into background model according to the update policies of BM. To our observation, two characteristics have been considered during update processing: the temporal and spatial characteristics. The temporal characteristic is to consider the relationship of a pixel at various timing and the spatial characteristic strengthen the relationship between a pixel and its neighbors. For example, Figure 4.3 (a)-(d) are the image sequences of waving trees. The blue and red points show in Figure 4.3 (a)-(d) are the pixels of the stationary and non-stationary background pixels. The temporal characteristic of the stationary background pixels at various timing in RGB color space is shown in Figure 4.3 (e). From Figure 4.3 (e), the stationary background is slowly changed with time, and their distribution is concentrated. The spatial characteristic of the non-stationary background pixels at various timing in RGB color space is

Algorithm 5. The update policies of BM

```
Input: The incoming pixel $x^t$, the position of current
        pixel X, and the positions of randomly chosen
        neighbors C and H.
Output: The updated background model
Regular update policy:
        if $x^t$ is classified into background then
            The matched codeword $c_m$ is updated.
        end if
Two-way propagation policy:
    First-propagation direction:
            To propagate the color information from the
            current pixel X to a randomly chosen neighbor C.
    Second-propagation direction:
            To propagate the color information from a
            randomly chosen neighbor H to current pixel X.
end
```

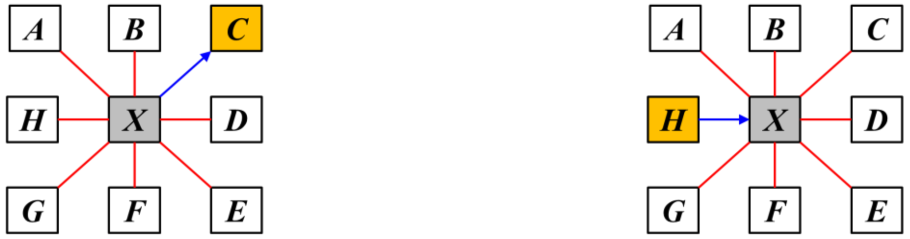
shown in Figure 4.3 (f). From Figure 4.3 (f), the distribution of non-stationary background is broader than stationary background. Therefore, we propose two updating polices, regular update and two-way propagation, for aiming to incorporate the consistency of temporal and spatial characteristics. The algorithm of the update policies of BM is shown in Algorithm 5.

In regular update policy, the temporal characteristic is considered which is inspire of [2] [9], the similar pixel is found in time axis and that is added into background model, the detail steps are described in the below. When the incoming pixel is classified as the background, the matched codeword c_m is updated according to the following equations:

$$v_m = \left(\frac{f_m\bar{R}_m+R}{f_m+1}, \frac{f_m\bar{G}_m+G}{f_m+1}, \frac{f_m\bar{B}_m+B}{f_m+1}\right), \text{ and}$$
$$aux_m = \begin{Bmatrix} min\{I_{x_t}, \check{I}_m\}, max\{I_{x_t}, \hat{I}_m\}, f_m = f_m + 1, \\ max\{\lambda_m = t - q_m\}, p_m, q_m = t \end{Bmatrix}. \tag{4.4}$$

The RGB vector v_m is calculated by averaging the value of red, green, and blue colors with the incoming pixel x_t and the parameters in aux_m are all updated.

In addition, for non-stationary background pixel, only consider temporal characteristic is not enough as shown in Figure 4.3 (f). Here, the two-way propagation policy is developed to integrate the spatial characteristic to solve the problem of non-stationary background. The main idea of two-way propagation is that when at time t, we want to obtain the information at time $t+1$ or further



(a) (b)

Figure 4.4: The two-way propagation policy. (a) the first-propagation direction. (b) the second-propagation direction.

to a pixel. As shown in Figure 4.3 (f), the information can be contributed from the neighbors at time t, we use the two-way propagation technique to achieve this function, the current pixel propagates the color information to a neighbor and a neighbor propagates the color information to the current pixel, details are described as below.

In the first direction of the two-way propagation, we randomly choose a neighbor from 8 connected neighborhoods around the current pixel and propagate the color information to the chosen neighbor. To find the matched codeword from that neighbor according to two criteria: the color distortion and the range of brightness, as described in Section 4.2.1. If the matched codeword is found, we update the color information into the matched codeword according to the equation 4.4. Otherwise, we create a new codeword c_L for the chosen neighbor and propagate the color information of the incoming pixel x^t which is classified as background to the codeword c_L by the following setting:

$$v_L = (R_{x^t}, G_{x^t}, B_{x^t})$$

$$aux_L = \{\check{I}_L = I_{x^t}, \hat{I}_L = I_{x^t}, f_L = 1, \lambda_L = 1, p_L = 1, q_L = 1\} \tag{4.5}$$

On the other direction, we randomly choose a codeword from the 8 connected neighborhoods around the current pixel and propagate the color information to the current pixel. If there is a matched codeword in the current pixel, we updated the matched codeword according to equation 4.4. Otherwise, we create a new codeword c_L for the current pixel and assign the value according to the equation 4.5. An example of two-way propagation is shown in Figure 4.4. Figure 4.4 (a) shows the first direction which propagates the color information from the current pixel ***X*** to a randomly chosen neighbor ***C***, as indicated by the blue line. . Figure 4.4 (b) shows the second direction of the two-way propagation.

The proposed update polies combine the temporal and spatial characteristics for efficiently tackling the problems of stationary and non-stationary backgrounds, respectively and capture the foreground accurately.

The filter and reconstruction policies of BM

Endless increase the codewords, makes the size of codebook became huge and that increased the time of matching. Therefore, we rejected the unsatisfied codewords which have the longest interval λ according to the following equation:

$$\lambda > T_{\lambda} \tag{4.6}$$

where T_{λ} is the number of frames which is set to be 500 in our implementation.

The filter policy would make the codebook became empty if the threshold T_{λ} is too small or the codeword is not recurred for a longest interval. Therefore, we proposed the reconstruction policy, if the codebook became empty, the reconstruction policy is triggered according to the construction of BE, as descripted in Section 4.2.1.

4.2.2 Background Gradient Extractor (BGE)

BM is useful to cope with the stationary, non-stationary background and sudden changes of illumination. However, the propagation policy passes the background information to a neighbor and that would blur and decrease the foreground intensities. Hence, we introduce the forbidden propagation policy that maintains the completeness of foreground. The forbidden propagation policy is based on the BGE introduced below.

Construction and update of BGE

In BM, we use RGB color to describe the background and foreground information. Here, the additional information, the gradient information, is used to assist to the segmentation result of foreground. The gradient information can clearly describe the boundary of objects. Therefore, we use gradient to construct background gradient model by using accumulation, and then obtain the foreground gradient information.

In BGE, we calculate the gradient of each frame, and accumulate the least N frames to construct the background gradient image for each frame. The current foreground gradient is extracted by comparing the current gradient frame with the gradient background model.

In the beginning of BGE, we used **AND** operator to background gradient model because of the least N frames is required, the equation is shown as below:

$$\nabla I_{BG_i} = \nabla I_{BG_{i-1}} \cap \nabla I_i , \tag{4.7}$$

where ∇I_{BG_i} is the background gradient image of the i^{th} frame, ∇I_i are the gradient results the i^{th} frame and $\nabla I_{BG_1} = \nabla I_1$.

After the first N frames, the background gradient image of the i^{th} frame is calculated by accumulating the least N frames according to the following equation:

$$\nabla I_{BG_i}(x,y) = \begin{cases} 255 & if \ \dfrac{\sum_{k=i-N}^{i} \frac{\nabla I_{BG_i}(x,y)}{255}}{N} \le T_{BG} \\ 0 & \qquad otherwise \end{cases} , \tag{4.8}$$

where $\nabla I_{BG_i}(x,y)$ is the background gradient value of the i^{th} frame at position (x,y), and T_{BG} is the threshold. In our experiment, the value of T_{BG} is equal or larger than 0.6 which has sufficient amount of reliable samples and that shows the higher performance of forbidden propagation.

Forbidden propagation policy

The forbidden propagation policy is used to ensure the completeness of the foreground that is based on BGE.

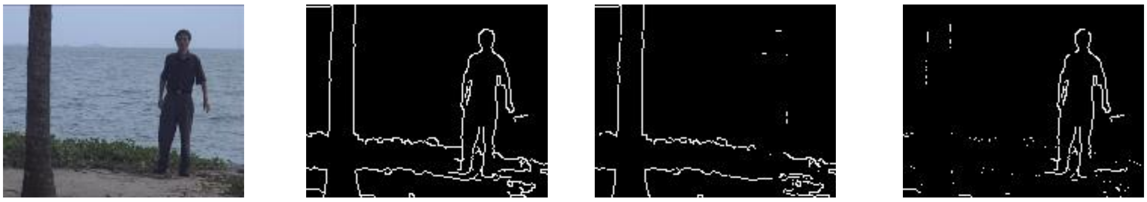


Figure 4.5: The results of background gradient extractor. (a) the original image of water surface; (b) the gradient of the current frame; (c) the gradient of background model; (d) the gradient of foreground.

There are two phases in the forbidden propagation policy. In phase 1, the background-gradient Image ∇I_{BG_i} is used to find the foreground gradient of the current frame by using the following equation:

$$\nabla I_{FG_i} = \nabla I_i - \nabla I_{BG_i} \tag{4.9}$$

where ∇I_{FG_i} is the foreground-gradient image. In phase 2, the color information of the incoming pixel is forbidden to propagate if the foreground-gradient value of itself or a randomly choice neighbor is 255.

Figure 4.5 shows the gradient results of current, background and foreground images. By BGE, the foreground gradient can be extracted currently. The performance of the forbidden propagation is demonstrated by using the data of water surface on the Perception dataset [12]. Figure 4.6 (a) is the original scenery of water surface. Figure 4.6 (b) and (c) are the results of Codebook and ViBe that have a lot of false positive because of tree and wave of the sea and has a huge broken on the human

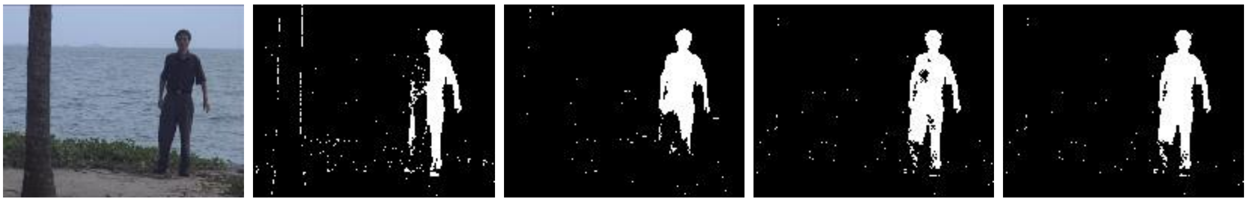


Figure 4.6: The comparison of forbidden propagation. (a) the original image of water surface; (b) the result of codebook; (c) the result of ViBe; (d) Proposed method without forbidden propagation; (e) Proposed method with forbidden propagation.

Compared methods	Mixture of Gaussian (MoG)
	Codebook
	Vibe
Test Videos	Dataset
1$^{\#}$: LightSwitch	Wallflower [11]
2$^{\#}$: TimeOfDay	
3$^{\#}$: WavingTrees	
4$^{\#}$: Campus	Perception [12]
5$^{\#}$: Curtain	
6$^{\#}$: Escalator	
7$^{\#}$: WaterSurface	

Table 4.1: List of experiment items.

body. Figure 4.6 (d) is the result of our proposed method without forbidden propagation which is considerably improves the performance of Figure 4.6 (b) and (c).

The result of forbidden propagation policy is shown in Figure 4.6 (e). The body is become more completeness and the false positive is not rising compared to the Figure 4.6 (c) and (d).

The foreground gradient is a criterion that effectively keeps the completeness of foreground. The completeness performance of the forbidden propagation policy is shown in the experiment.

4.3 Detection results and comparison

In this section, we analyze the accuracy of SBM by comparing with MoG, Codebook, and ViBe. It should be noted that, the parameters used in these methods are all optimized.

Seven videos are selected from two popular surveillance video datasets, Wallflower [11] and Perception [12], which are used to evaluate the accuracy of SBM with MoG, Codebook and ViBe. The surveillance scenes of the selected test videos include indoor and outdoor. Some scenes of these videos are crowded with people, some with dynamic background caused by trees and water, others are the sudden changes of illumination caused by human and nature. The detailed list of the

Total Error(TE)				
Video	MoG	Codebook	ViBe	SBM
$1^\#$	15828	11887	15053	4221
$2^\#$	1044	1081	1143	854
$3^\#$	1807	1011	1172	378
$4^\#$	1168	1535	605	317
$5^\#$	511	1471	1768	993
$6^\#$	362	1205	746	557
$7^\#$	376	1091	1172	478
Average	3013	2754	3047	1114

Table 4.2: Experiment results.

experiment is shown in Table 4.1.

Three indicative items: false positive (FP), false negative (FN) and total error (TE), are used to evaluate the performance of these methods and the hand-segments are used to be the ground truth. The false positive refers to the number of pixels which is the background pixel but marked as foreground. The false negative is contrary to the false positive which is the foreground pixel and be marked as background. The sum of the false positive and false negative is expressed as the total error.

The TEs of each video for each algorithm are shown in Table 4.2. Some results of MoG have the least TEs; however, the proposed method (SBM) has the least average error which is more suitable for different environments. Table 4.3 shows the results with images. In Table 4.3, the performance of MoG, Codebook and ViBe in the sudden changes of illumination are not well, as shown in video $1^\#$. Because MoG and Codebook are all need a period of time to update the background model, ViBe did not has a policy of detecting the sudden change of illumination and reconstructing the background model. SBM performs well in the situation of the dynamic background, such as video $3^\#$, $4^\#$ and $7^\#$, and sudden changes of illumination, such as video $1^\#$ and $2^\#$. In all of the tested video, SBM has the least false negative and has the most completeness foreground.

Figure 4.7 shows the performance in term of FP, FN for each algorithm. The blue bar and the

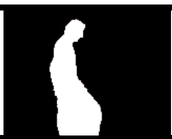
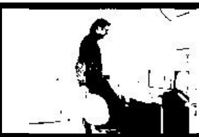
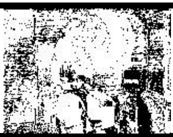
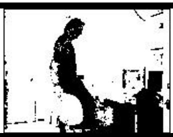
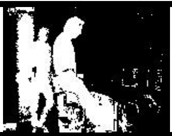
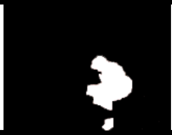
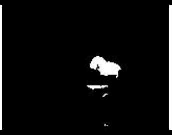
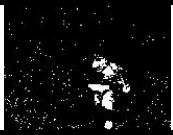
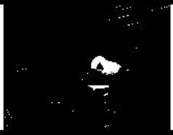
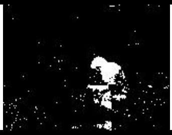
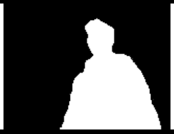
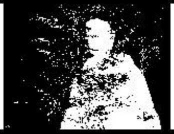
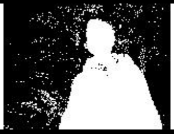
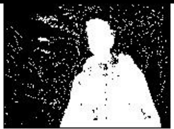
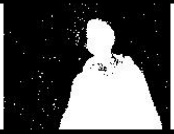
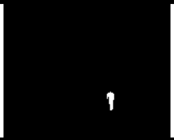
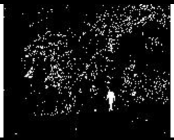
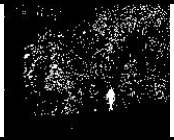
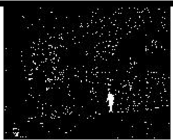
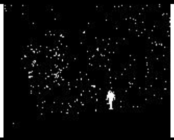
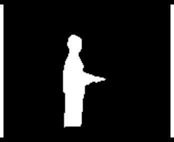
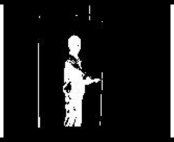
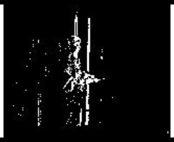
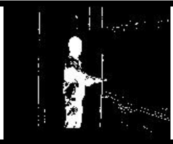
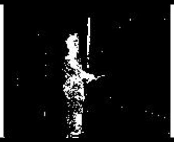
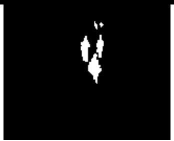
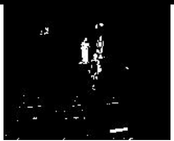
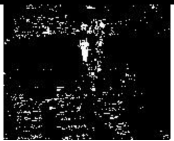
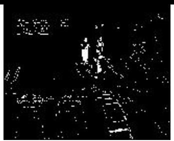
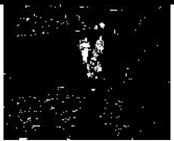
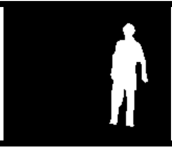
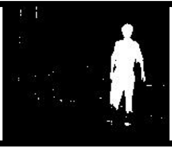
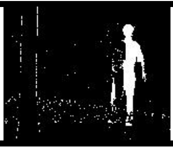
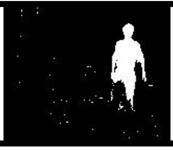
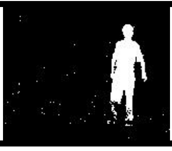
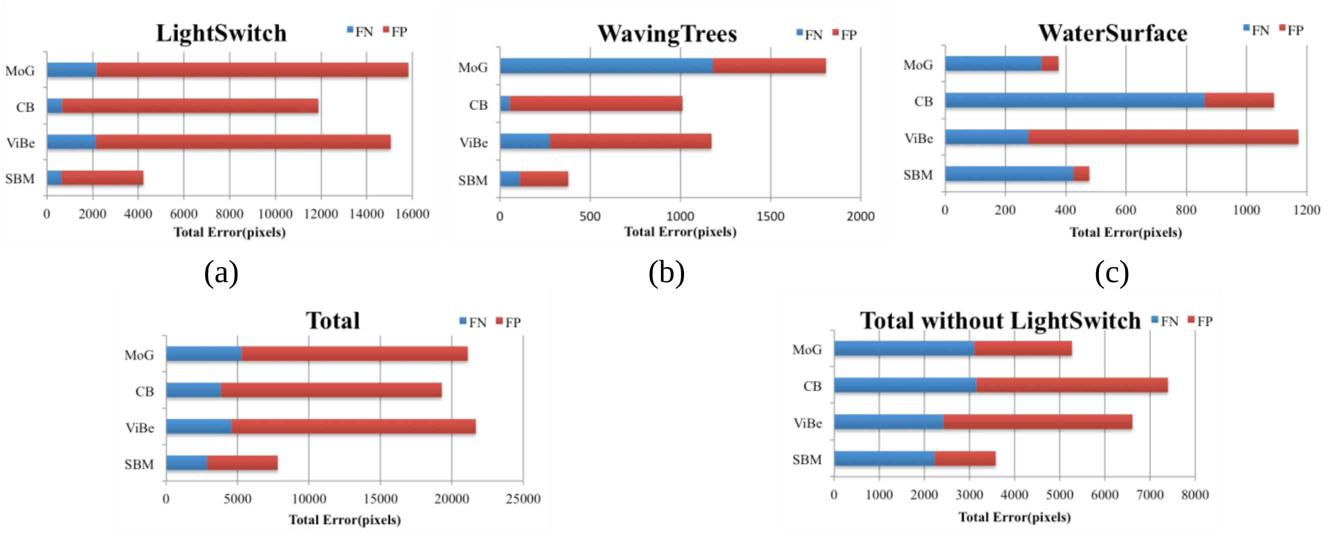
Video	Test image	Ground Truth	MoG Stauffer *et al.* [1]	Codebook Kim *et al.* [2]	ViBe Barnich *et al.* [10]	Proposed method (SBM)
1#						
2#						
3#						
4#						
5#						
6#						
7#						

Table 4.3: Results on the dataset.

red bar in Figure 4.7 are the value of FP and FN, respectively. We especially illustrate the results of dynamic background and sudden changes of illumination in Figure 4.7 (a) to (c). The total error with seven videos is shown in Figure 4.7 (d) and the total error without the sudden change of illumination is shown in Figure 4.7 (e). Although the TE of MoG is the least in Figure 4.7 (c), the performance of all videos of MoG is worst, as shown in Figure 4.7 (d) and (e). To compare with Codebook, because of SBM consists the spatial information, the difference of FP between SBM and Codebook is obvious and the total error of proposed method is the least. Comparing with ViBe, the total error of SBM is the least, even in the case of sudden change of illumination.



(a) (b) (c)

(d) (e)

Figure 4.7: The chart of the performance of each algorithm. (MoG: Mixture of Gaussian, CB: Codebook, ViBe, SBM: Proposed method). (a) LightSwitch. (b) WavingTrees. (c) WaterSurface. (d) Total Error. (e) Total Error without LightSwitch.

4.5 Summary

In this chapter, we design a new framework, the spatial background model (SBM), for addressing the dynamic background and the sudden changes of illumination in background subtraction. Two main components are proposed: the background model (BM) and the background gradient extractor (BGE). The BM is proposed to capture foreground and eliminate the noise in background. To model the BE, only a single frame is used which is based on a single-layer codebook model and the spatial information is propagated from the adjacent neighbors. The BM can efficiently eliminate the dynamic background and the sudden changes of illumination. However the propagation makes the foreground incompleteness. Therefore, the propagation is forbidden according to the BGE that keeps the completeness of foreground. The BGE is synchronously constructed with BM. To construct the BGE, the stable background gradient is used which is used to find foreground gradient of the current frame.

Experimental results on a set of background extractor databases, Perception and Wallflower. We analyze the accuracy of the proposed method (SBM) to compare with MoG, Codebook, and ViBe. The total error of the proposed method is the least, and the capabilities of handling the dynamic background and the sudden changes of illumination are greater than others.

CHAPTER 5

COMBINING SPATIAL BACKGROUND MODELING AND RANDOM FOREST CLASSIFIER FOR FOREGROUND SEGMENTATION AND SHADOW REMOVAL

5.1 Introduction

Recently, due to the terrorist attacks, such as 911 in New York, London bombing in 2005 and some local violent events occur in different countries, the security issue has been regarded. In the past, we consume lots of people and money to solve the problems of security. However, the resource of surveillance can be saved because of the developing of the technology. The visual surveillance spans several applications, such as human detection, people counting, object left, and object stolen, in which the object detection is the most important part. The useful information, such as the location, the shape, and the size of the object can be obtained from object detection and be applied to different applications.

A challenging class of object detection is the cast shadow. Cast shadows generated by moving objects and that are detected as the foregrounds. However, incorrect object detections caused by shadows make the applications of visual surveillance become unreliable. Therefore, shadow removal is an important issue in visual surveillance.

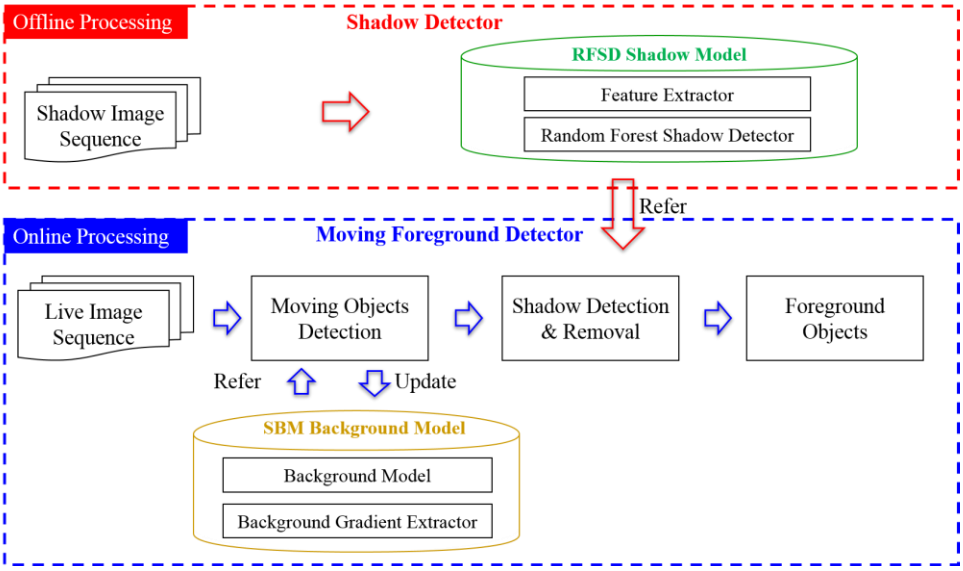


Figure 5.1: The work flow of Shadow Removal.

There have been numerous works dedicated to solve the problem of shadow removal. In the method [78], the shadow removal is classified into two categories: parameter-based and model-based. The parameter-based method used the feature of shadow, such as RGB, HSV, and texture, to classify the foreground and shadow. The drawbacks of these methods are sensitive to the parameter and the threshold for each feature. The model-based method used the classifier which is modeled using features to discriminate the shadow from moving objects. Although, the model-based method needed the prior knowledge, such as the ground truth and object's class, it is suitable for different scene.

In this chapter, we propose two primary technical components: shadow detector and foreground detector. First, we use a learning technique that employs random forest algorithm to learn the chromaticity, physical property, and texture characteristics to construct the shadow detector which is model-based. After that, we propose a process that combining a spatial background model which constructs the background model with a single frame and is useful to overcome the dynamic background and the suddenly changes of illumination with the shadow detector to discriminate shadow from moving objects. Figure 5.1 gives the flowchart of the proposed method.

5.2 System Architecture

The proposed system includes two primary technical components: shadow detector and moving foreground detector as shown in Figure 5.1. In the offline processing, we extract the features from training data and label each pixel as foreground or shadow according to the ground truth. Then, we associate the training data with the classification methodology, Random Forest, to generate the shadow detector. In the online processing, moving foreground detector, the candidate region of moving objects are detected using the spatial background model (SBM) which has two components: the background model (BM) and the background gradient exactor (BGE). The BM is used to detect the moving objects with the less noise, and the BGE is used to keep the completeness of detected moving objects. After SBM, the candidate region of moving objects is obtained and that is recalculated with shadow detector to obtain the moving foreground without shadow.

The spatial background model has been presented in chapter 4. In the following, we first introduce the properties of each feature and the reasons that we chose these features. After extracting the features, the algorithm of Random Forest is used to train the classifiers for shadow removal.

5.3 Random Forest Shadow Detector

In this section, we proposed the Random Forest shadow detector to discriminate shadow from the moving objects. We first explain the algorithm for extracting features. Then, we describe the classifier that is used in our algorithm.

5.3.1 Feature Extraction

In this subsection, we adopted three features, chromaticity, physical property, and texture characteristics, which are analyzed in [60] for the Random Forest classifier. The details of these features are described in the following.

A. Chromaticity

To choose the color space for separating the intensity and chromaticity is an important issue. Several color spaces, such as HSV [58], c1c2c3 [79] and YUcUv [80] have great performance for shadow removal. We chose the method proposed by Cucchiara *et al*. [58] to obtain the chromatic features, because of the value of chromaticity and that of intensity are divided in this color space and it is widely used in shadow removal. Although, the value of hue and saturation on the region of shadow are similar to that of background, the intensity in the shadow is lower than in the background. The features of chromaticity are calculated according to the following equations:

$$C_{H_p} = \left| F_p^H - B_p^H \right| , \tag{5.1}$$

$$C_{S_p} = \left| F_p^S - B_p^S \right| , \tag{5.2}$$

$$C_{V_p} = {F_p^V} \big/ {B_p^V} \quad , \tag{5.3}$$

where F_p^H, F_p^S and F_p^V represent the value of hue, saturation and intensity at pixel p of current frame. The B_p^H, B_p^S and B_p^V represent the value of hue , saturation and intensity at pixel p of background. We use equation 5.1 and equation 5.2 to calculate the difference between the current frame and background and use equation 5.3 to calculate the proportion of intensity.

B. Physical Properties

The characteristics of the ambient light which is blue and the sun which is white are considered in the physical properties. Because of the region of shadow remain the source of sun light, it is bluer than the background which is not be blocked. We used the physical properties which are discussed in [60] based on RGB color space. The physical properties are calculated according to the following equations:

$$P_{\phi_p} = \cos^{-1}\left({F_p^B} \big/ {\left\| F_p \right\|} \right), \tag{5.4}$$

$$P_{\theta_p} = \tan^{-1}\left(F_p^G \big/ F_p^R \right), \tag{5.5}$$

$$P_{\alpha_p} = \left\| F_p \right\| \big/ \left\| B_p \right\|, \tag{5.6}$$

where F_p^R, F_p^G and F_p^B represent the components of red, green and blue of the current frame at the pixel p. The F_p and B_p represent the value of combining the RGB component of current frame and that of background at the pixel p, respectively.

P_{ϕ_p} is the most important feature in the physical properties that presents the angle between blue and color at pixel p. P_{θ_p} is the angle between green and red of current frame at pixel p to confirm the color is similar to the background. P_{α_p} is the illumination attenuation of the current frame and background.

C. Texture

Various formulations are proposed to extract the texture features from image. However, none is perfect for various scenery. Therefore, we chose the widely used method proposed by [60] [66] which is classic and easily implement without time consuming. The texture features are calculated as shown in the following equations:

$$T_{\lambda_P} = \left(F_p / B_p \right) \big/ \left(F_p - B_p \right), \tag{5.7}$$

$$T_{\nabla_p} = \sqrt{\nabla F_x^2 + \nabla F_y^2} - \sqrt{\nabla B_x^2 + \nabla B_y^2}\,, \tag{5.8}$$

$$T_{\theta_p} = \tan^{-1}\left(\nabla F_y \big/ \nabla F_x \right) - \tan^{-1}\left(\nabla B_y \big/ \nabla B_x \right), \tag{5.9}$$

where ∇_x and ∇_y are the horizontal and vertical gradient of image, respectively. T_{λ_P} is the difference between current frame and background image. T_{∇_p} and T_{θ_p} are the gradient and

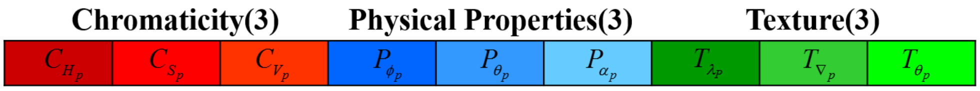


Figure 5.2: A 9-dimensional vector of three features.

orientation of each pixel p, respectively.

Notice that, we do not choose the geometry feature in this work, because of the proposed method is pixel-based, and the geometry feature is region-based.

5.3.2 Random Forest Classifier

In this study, the Random Forest learning mechanism is adopted, because of the properties of feature that have different performance in each scene is similar to the concept of Random Forest.

Each feature described in Section 5.3.1 has three dimensions. However, it is difficult to choose a feature for a scene. Therefore, in this work, we merge three features to be one instance with nine dimensions, as shown in Figure 5.2. An instance can be expressed as a vector $f_i = \left\langle C_{H_p}, C_{S_p}, C_{V_p}, P_{\phi_p}, P_{\theta_p}, P_{\alpha_p}, T_{\lambda_p}, T_{\nabla_p}, T_{\theta_p} \right\rangle$, in which f_i is a combined instance for training data at pixel p. The classifier is obtained according to the Random Forest algorithm as shown in Figure 5.3. For each tree of Random Forest, it randomly choose a subsample from f to be the training data. For each node of tree, Random Forest chose the most suitable part of features according to the distribution of the selected training data, and train the best threshold for that node. The number of node of each tree is growing, until the training data is small enough.

5.4 Experiments

In this section, we first compare the result of Random Forest shadow classifier with other methods and then demonstrate the performance of the proposed method, SBE + Random Forest classifier.

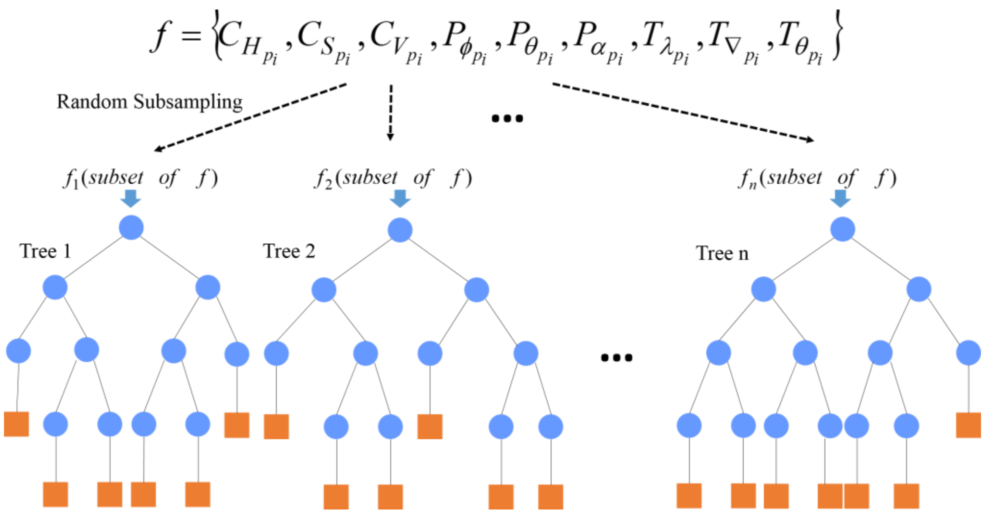

$f = \{C_{H_{Pi}}, C_{S_{Pi}}, C_{V_{Pi}}, P_{\phi_{Pi}}, P_{\theta_{Pi}}, P_{\alpha_{Pi}}, T_{\lambda_{Pi}}, T_{\nabla_{Pi}}, T_{\theta_{Pi}}\}$
Random Subsampling
$f_1(subset\ \ of\ \ f)$
$f_2(subset\ \ of\ \ f)$
$f_n(subset\ \ of\ \ f)$
Tree 1
Tree 2
Tree n

Figure 5.3: The Random Forest Shadow Detector.

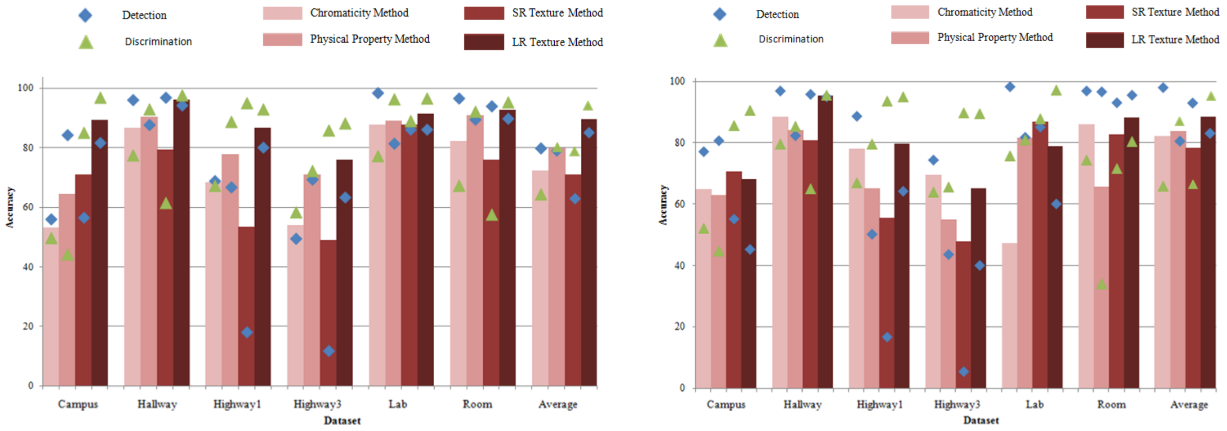
5.4.1 Experimental Results of Shadow Removal

In this subsection, we analyze the performance of shadow removal. First, we shown the results of each and combined features which is classified by SVM or Random Forest. Then, we demonstrate the performance of the proposed method with other methods which is described in [60]. Two metrics are used, the shadow detection rate (η) and the shadow discrimination (ξ), to evaluate the performance of shadow removal which are proposed by Prati *et al*. [81], as shown below:

$$\eta = \frac{TP_s}{TP_s + FN_s}, \tag{5.10}$$

$$\xi = \frac{TP_f}{TP_f + FN_f}. \tag{5.11}$$

Four indicative items are used: TP_s, TP_f, FN_s and FN_f. TP_s refers to the number of pixels which are the shadow pixel then marked as shadow. TP_f refers to the number of pixels which are the foreground pixel then marked as foreground. FN_s refers to the number of pixel which are the foreground pixel then marked as shadow. FN_f represents to the number of pixel which are the shadow pixel then marked as foreground.



(a) (b)

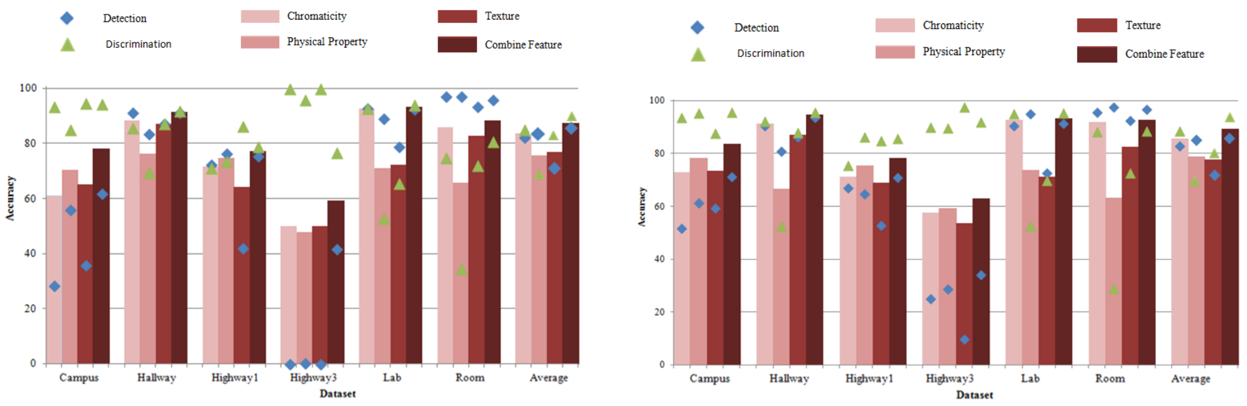
Figure 5.4: Comparison of shadow detection results on various sequences. (a) the results from [60]/ (b) the results of the program released from [60].

The value of η and ξ show the performance of each method. However, if a method has poor performance that marked all testing data as foreground, for example, the method may has one hundred percent accuracy of ξ and zero accuracy of η. Therefore, we use the average of η and ξ to represent the accuracy of performance.

Sanin *et al*. [60] releases the programs for geometry method, chromacity method, physical method, small-region texture method, and large-region texture method. Figure 5.4 shows the average shadow detection and discrimination rates on various sequences. Figure 5.4 (a) and (b) are the results from the paper of [60] and that of the released programs which has optimal parameters, respectively. From Figure 5.4, there are some of difference between Figure 5.4 (a) and (b), because the performance of those programs are all depend on the parameters tuning. In the following, the programs are used in the different experiments and the parameters are all optimal.

5.4.2 Comparison of SVM and Random Forest Classifiers

Various features have different performance in different environments. Here, we first demonstrate the performance of SVM classifier with individual features and the combined feature. Then, we compared the performance of SVM with Random Forest.



(a) (b)

Figure 5.5: Experimental results of different classifiers. (a) SVM. (b) Random Forest.

Figure 5.5 (a) is the results using SVM classifier with individual features and the combined feature in different environments. From Figure 5.5 (a), the combined feature has the best performance in SVM classifier. As the same as Figure 5.5 (a), the combined feature has the best performance in Random Forest classifier as shown in Figure 5.5 (b).

From Figure 5.5, the combined feature has the best performance in different environments. The performance of Random Forest classifier with combined feature is better than SVM classifier.

5.4.3 Experimental Results of Different methods

In this subsection, the results of SVM, Random Forest and each method which is described in [56] are shown in Figure 5.6.

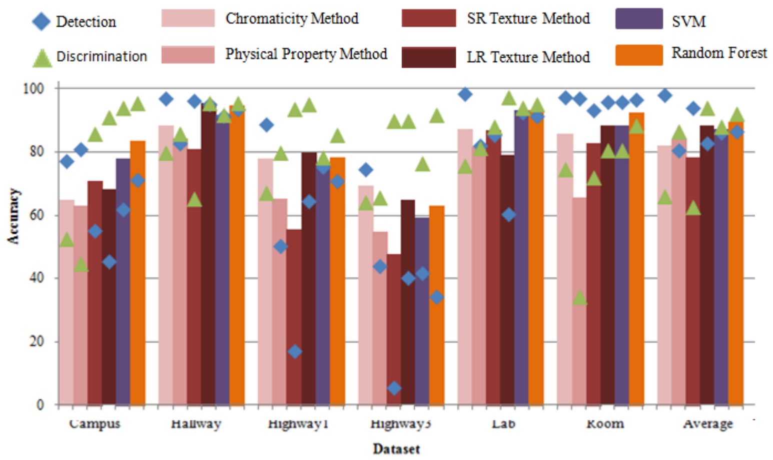


Figure 5.6: Experimental results of Different methods.

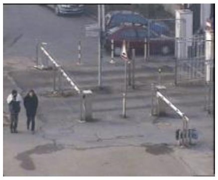
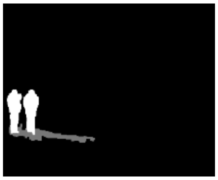
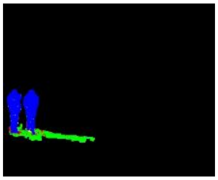
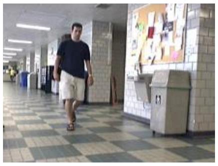
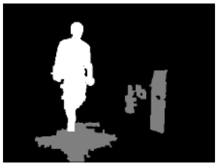
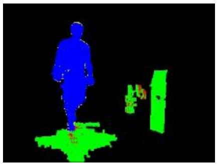
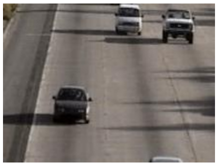
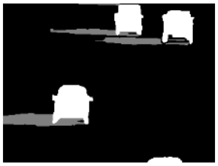
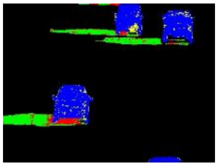
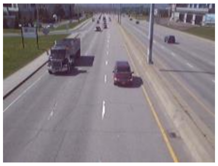
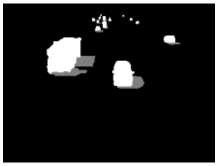
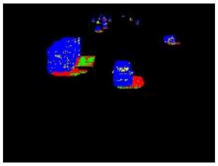
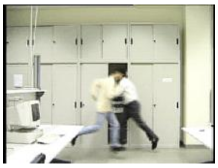
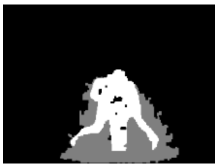
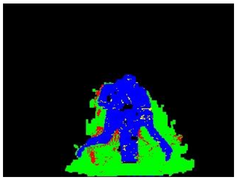
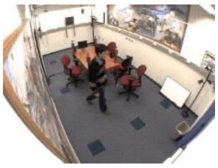
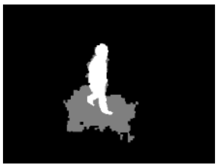
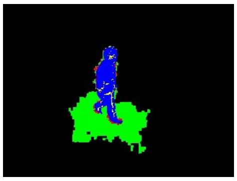
Dataset	Current Frame	Ground Truth	RFSD
Campus			
Hallway			
Highway1			
Highway3			
Lab			
Room			

Table 5.1: Experimental results of Random Forest shadow detector.

In Figure 5.6, each method has the worst performance in the dataset of Highway3, in which the chromacity method has the best performance, and SVM and Random Forest also performed well. Although, the large-region texture method has the best performance in some datasets, SVM and Random Forest have the best performance in average. To compare SVM with Random Forest, Random Forest has the best performance in the most datasets and has the best performance in average. The performance of each method described in [60] is depend on the parameter tuning, the proposed method, Random Forest shadow classifier, is suitable for various scenes.

Table 5.1 shows the proposed method, Random Forest shadow classifier. The first column is the

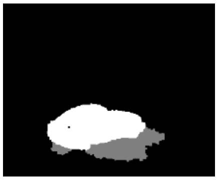
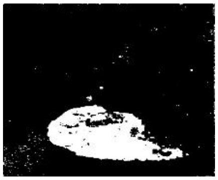
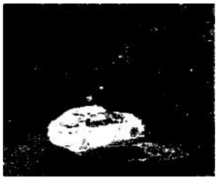
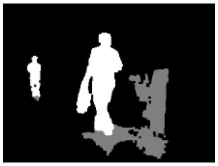
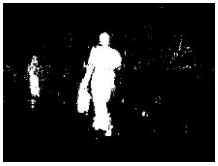
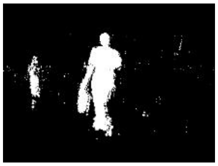
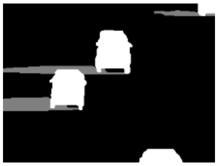
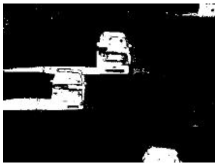
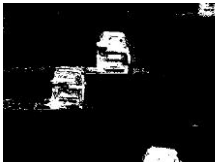
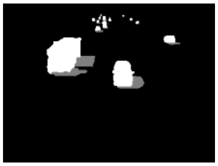
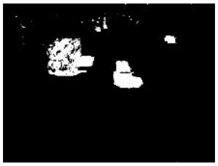
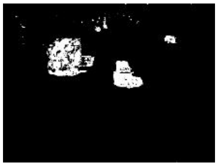
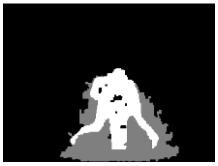
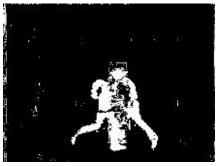
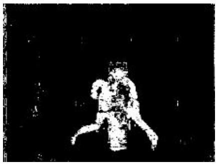
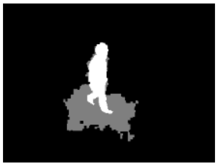
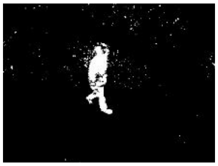
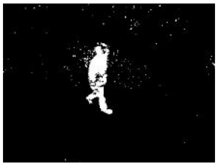
Current Frame	Ground Truth	SBM	SBM + RFSD
			
			
			
			
			
			

Table 5.2: Experimental results of the proposed method.

current frame, the second column is the ground truth in which the foreground pixels are indicated as white and the shadow pixels are marked in gray, and the third column is results of Random Forest classifier in which the shadow pixels which is classified as shadow are marked in green, the shadow pixels which are classified as foreground are marked in red, the foreground pixels which are classified as foreground marked as blue, and the foreground pixel which are classified as shadow are marked in yellow.

5.4.4 Experimental Results of the Proposed Method

We combine the SBM and Random Forest shadow detector to detect moving objects. The classifier

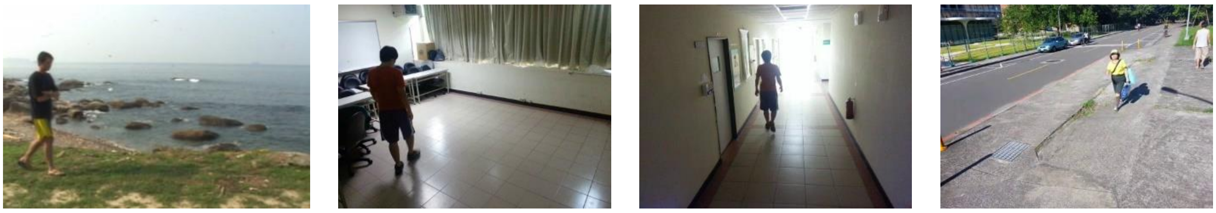
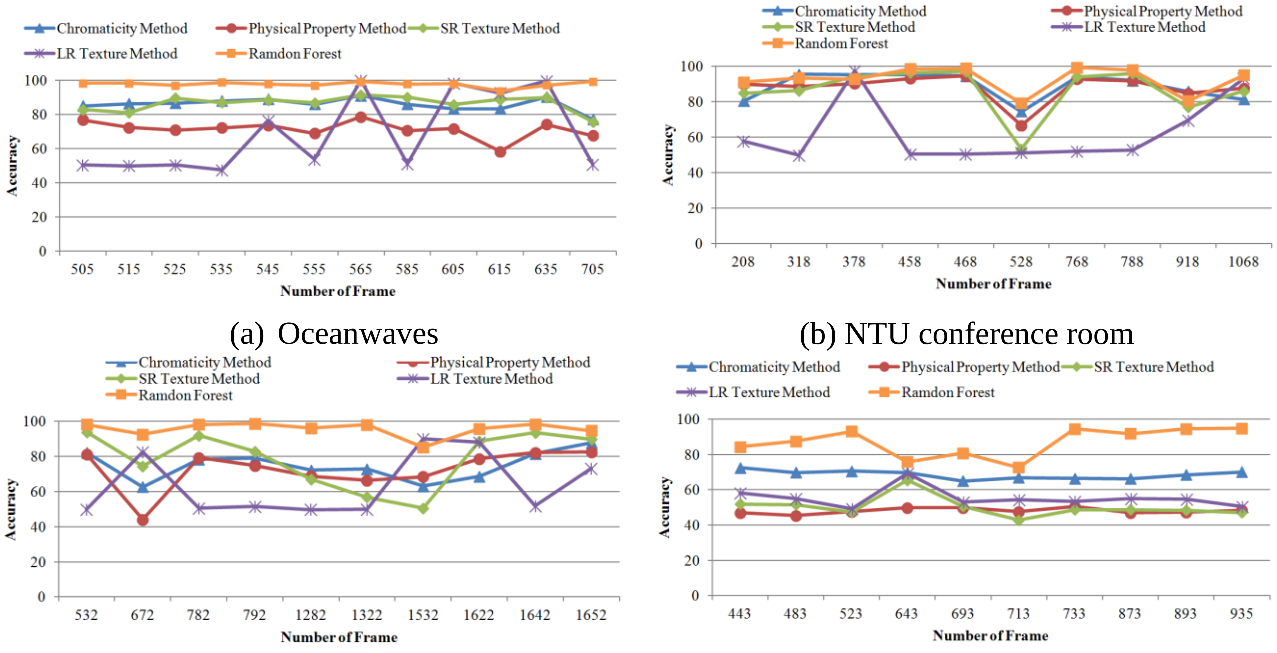

(a) (b) (c) (d)

Figure 5.7: The experimental environments of the proposed dataset. (a) Oceanwaves. (b) NTU conference room. (c) NTU hallway. (d) NTU outdoor.

is trained according to the ground truth of training data, and the shadow is removed from the results of SBM, the results are shown in Table 5.2. In Table 5.2, some results of SBM have great performance of removing shadows. We can see the results of Hallway, Lab and Room have the clear shadow without using shadow removal. Thus, the performance of shadow removal is not obvious. In the cases of Campus and Highway1, SBE did not perform well in cast shadow. After using the Random Forest classifier, the shadow is removed and the results of moving objects detection become accurately.

We also proposed a new dataset which includes four sequences: oceanwaves, NTU conference room, NTU hallway and NTU outdoor, for testing, as shown in Figure 5.7. The oceanwaves has the dynamic background, the NTU conference room has the reflective floor, the NTU hallway has the



(a) Oceanwaves (b) NTU conference room

(c) NTU hallway (d) NTU outdoor

Figure 5.8: Experimental results of the proposed method with the proposed dataset.

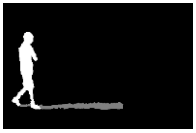
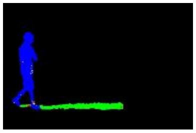
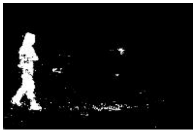
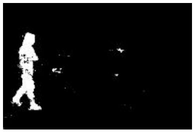
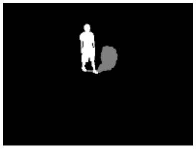
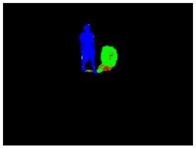
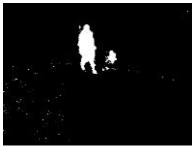
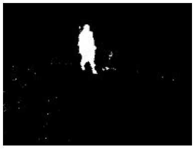
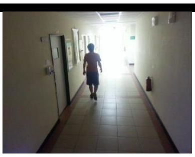
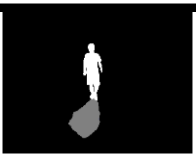
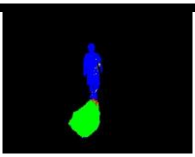
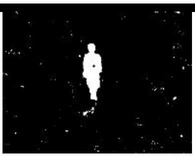
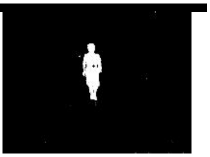
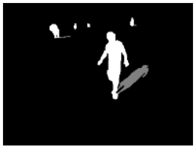
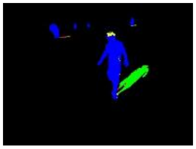
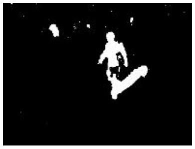
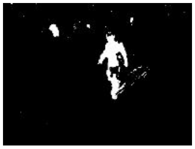
#frame	Current Frame	Ground Truth	RFSD	SBM	SBM+RFSD
00505					
00378					
00532					
00893					

Table 5.3: Experimental results of the proposed method with the proposed dataset.

strong light comes from the window, and the NTU outdoor has the strong light on the sidewalk and on the traffic lane. Figure 5.8 (a)-(d) show the results of shadow removal using the proposed dataset. From Figure 5.8, the proposed classifier, Random Forest classifier, has the best performance in all of the proposed dataset. Table 5.3 shows the results of proposed method, SBM + Random Forest shadow classifier, SBM is useful for dynamic background and Random Forest shadow classifier can effectively remove the shadow effect, as shown in Table 5.3.

5.6 Summary

We propose a comprehensive method that combines the object detection and shadow removal. First, we design a novel framework, the spatial background model (SBM) for modeling background that integrate the temporal and spatial information. Two main components are proposed: the background model (BM) and the background gradient extractor (BGE).

Then, to overcome the problem of shadow, we used the algorithm of Random Forest to be the classifier which is suitable for the properties of shadow features and combined different features, chromaticity, physical properties and texture, for shadow removal.

In the experimental results, we first show the results of shadow removal, the performance of combined feature are greater than individual features in both SVM and Random Forest classifiers. To compare the Random Forest classifier with other methods, the Random Forest classifier performs well and that is suitable for various environments.

Finally, we demonstrate the proposed method, SBE + Random Forest classifier, in both classic and new datasets.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Thesis

We have studied three common topics for visual surveillance: constructing the large- area high-resolution for visual monitoring system, design the spatial background model (SBM) for objects detection, and cast shadow removal using Random Forest classifier.

For constructing large-area high-resolution visual monitoring system, we proposed two PTZ camera calibration methods: in-factory calibration and on-site calibration and a framework that using the calibration results of two cameras to construct the proposed visual-monitoring system. The in-factory calibration method is useful when stable features are extremely difficult to obtain. The on-site calibration method is more flexible than in-factory calibration method is because it can be applied both indoors and outdoors. To achieve the multilayered PTZ-camera calibration, the consistent property of distinct pan-tilt angles at various zooming factors is used. To construct the large-are high-resolution visual monitoring system, the calibration results are used. The large-area high-resolution background image is constructed and that is combined with online multiresolution foreground image.

For detecting moving objects, an integrated methodology, the spatial background model, for addressing the dynamic background and the sudden changes of illumination in background subtraction is proposed. Two components are proposed: background model (BM) and background gradient extractor (BGE). In BM, the background model is constructed using a single frame that is based on a single-layer codebook model and that integrates the spatial information is propagated from the adjacent neighbors that can effectively eliminate the dynamic background and the sudden changes of illumination in the background. However the propagation makes the foreground incompleteness. Therefore, the propagation is forbidden according to the BGE that keeps the completeness of foreground. The BGE is synchronously constructed with BM. To construct the BGE, the stable background gradient is used which is used to find foreground gradient of the current frame.

For cast shadow removal, we constructed the shadow classifier utilizing the characteristic of features of shadow, such as chromaticity, Physical Property, and texture based on Random Forest algorithm. To compare the results of Random Forest shadow classifier with other state-of-the-art techniques in both classic and new datasets, the Random Forest shadow classifier performs well in most parts of testing data and that has the best average result. Random Forest shadow classifier is suitable for various environments without tuning parameters. We also demonstrated the combination of the spatial background model and Random Forest shadow classifier, the moving objects are completely detected and the shadow effects are removal.

6.2 Future Directions

We have investigated the visual surveillance techniques for monitoring and detecting moving objects. In this section, we introduce some research directions which include some extension of current works and the approaches for integrating the components.

For constructing large-area high-resolution visual monitoring system, we plan to integrate learning based super-resolution approaches to increase the overall resolution and enable users to monitor the details of multiple targets in a scene. Moreover, the proposed system was designed for use by only a single user, and only one PTZ camera was integrated because a single user can focus on only one region at a time. To customize the system for multiple users, several PTZ cameras can be added; the proposed system framework can be readily expanded to attain this functionality. Finally, in the current system, a standard background-subtraction mechanism was used [8], which cannot effectively address occlusion, but processes images in real time. In the future, we plan to adopt alternative background subtraction solutions to eliminate partial occlusion, such as the approaches developed by [10] and [82], to enhance the quality of the foreground image.

For moving objects detection, although the forbidden propagation can keep the completeness of the objects, some post-processing still needed which used to improve the accuracy of the spatial background model, such as the color segmentation.

For shadow removal, the features of shadow traditionally extract from a single image, we will attempt to consider the relationship between frame and frame in the future, and some features can be considered to apply into the Random Forest classifier, such as the wavelet feature and scattering transform.

LIST OF REFERENCES

[1] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *Pattern Analysis and Machine Intelligence,IEEE Transactions on* , vol. 22, no. 8, pp. 747-757, 2000.

[2] K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time imaging,* vol. 11, no. 3, pp. 172-185, 2005.

[3] M. Andriluka, S. Roth and B. Schiele, "People-tracking-by-detection and people detection-by-tracking," in *Proceeding of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[4] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on,* vol. 50, no. 2, pp. 174-188, 2002.

[5] P. N. Belhumeur, J. P. Hespanha and D. Kriegman, "Eigenfaces vs. fisherfaces:Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 19, no. 7, pp. 711-720, 1997.

[6] D. J. Dailey, F. W. Cathey and S. Pumrin, "An algorithm to estimate mean traffic speed using uncalibrated cameras," *Intelligent Transportation Systems, IEEE Transactions on,,* vol. 1, no. 2, pp. 98-107, 2000.

[7] A. J. Lipton, H. Fujiyoshi and R. S. Patil., "Moving target classification and tracking from real-time video," in *Proceeding of the 4th IEEE Workshop on Applications of Computer Vision*, 1998.

[8] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking.," in *the IEEE Computer Society Conference on the Computer Vision and Pattern Recognition*, 1999.

[9] K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, "Background modeling and subtraction by codebook construction," in *the IEEE International Conference on Image Processing*, 2004.

[10] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *Image Processing, IEEE Transactions on,* vol. 20, no. 6, pp. 1709-1724, 2011.

[11] K. Toyama, J. Krumm, B. Brumitt and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proceedings of the 7th IEEE International Conference on Computer Vision*, 1999.

[12] L. Li, W. Huang, I. H. Gu and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *Image Processing, IEEE Transactions on,,* vol. 13, no. 11, pp. 1459-1472, 2004.

[13] T. Zhao and R. Nevatia, "Stochastic human segmentation from a static camera," in *Proceedings of the IEEE Workshop on Motion and Video Computing*, 2002.

[14] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[15] M. Lalonde, S. Foucher, L. Gagnon, E. Pronovost, M. Derenne and A. Janelle, "A system to automatically track humans and vehicles with a ptz camera," in *Proceeding of the International Society for Optics and Photonics on Defense and Security Symposium,* 2007.

[16] A. Mian, "Realtime face detection and tracking using a single pan, tilt, zoom camera," in *Proceeding of the 23th International Conference on Image and Vision Computing New Zealand*, 2008.

[17] Y. Ye, J. K. Tsotsos, E. Harley and K. Bennet, "Tracking a person with pre-recorded image database and a pan, tilt, and zoom camera," *Machine Vision and Applications,* vol. 12, no. 1, pp. 32-43, 2000.

[18] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision,* vol. 57, no. 2, pp. 137-154, 2004.

[19] A. Ess, B. Leibe, K. Schindler and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[20] Q. Wang, F. Chen, W. Xu and M. H. Yang., "Object tracking via partial least squares analysis," *Image Processing, IEEE Transactions on,* vol. 21, no. 10, pp. 4454-4465, 2012.

[21] R. Wang, S. Shan, X. Chen, Q. Dai and W. Gao, "Manifold–manifold distance and its application to face recognition with image sets," *Image Processing, IEEE Transactions,* vol. 21, no. 10, p. 4466–4479, 2012.

[22] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, "Face recognition: A literature survey.," *Acm Computing Surveys (CSUR),* vol. 35, no. 4, p. 399–458, 2003.

[23] S. N. Sinha and M. Pollefeys, "Pan–tilt–zoom camera calibration and high-resolution mosaic generation," *Computer Vision and Image Understanding,* vol. 103, no. 3, p. 170–183, 2006.

[24] K. Xue, G. Ogunmakin, Y. Liu, P. A. Vela and Y. Wang, "Ptz camera-based adaptive panoramic and multi-layered background model," in *Proceeding of the 18th IEEE International*

Conference on Image Processing, 2011.

[25] K. Xue, Y. Liu, G. Ogunmakin and J. Zhang, "Panoramic gaussian mixture model and large-scale range background substraction method for ptz camera-based surveillance systems," *Machine vision and applications,* vol. 24, no. 3, p. 477–492, 2013.

[26] L. Marchesotti, L. Marcenaro and C. Regazzoni, "Dual camera system for face detection in unconstrained environments," in *Proceeding of the IEEE International Conference on Image Processing*, 2003.

[27] X. Zhou, R. T. Collins, T. Kanade and P. Metes, "A master-slave system to acquire biometric imagery of humans at distance," in *In First ACM SIGMM international workshop on Video surveillance*, 2003.

[28] C. C. Chen, Y. Yao, A. Drira, A. Koschan and M. Abidi, "Cooperative mapping of multiple ptz cameras in automated surveillance systems," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[29] V. K. Singh and P. K. Atrey, "Coopetitive Visual Surveillance Using Model Predictive Control," in *Proceedings of the Third ACM International Workshop on Video Surveillance and Sensor Networks*, 2005.

[30] Z. Zhang, Z. Liu and Q. Zhao, "Semantic saliency driven camera control for personal remote collaboration," in *Proceedings of the 10th IEEE workshop on multimedia signal processing*, 2008.

[31] C. H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan and M. Abidi, "Heterogeneous fusion of omnidirectional and ptz cameras for multiple object tracking," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 18, no. 8, p. 1052–1063, 2008.

[32] A. Alahi, D. Marimon, M. Bierlaire and M. Kunt, "A master-slave approach for object detection and matching with fixed and mobile cameras," in *Proceedings of the 15th IEEE International Conference on Image Processing*, 2008.

[33] C. Micheloni, E. Salvador, F. Bigaran and G. L. Foresti, "An integrated surveillance system for outdoor security," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance,* 2005.

[34] M. Reale, T. Hung and L. Yin, "Pointing with the eyes: Gaze estimation using a static/active camera system and 3d iris disk model," in *Proceedings of the IEEE International Conference on Multimedia and Expo,* 2010.

[35] K. W. Chen, C. W. Lin, M. Y. Chen and Y. P. Hung, "e-fovea: a multi-resolution approach with steerable focus to large-scale and high-resolution monitoring," in *Proceedings of the International Conference on Multimedia,* 2010.

[36] F. Dornaika and J. H. Elder, "Image registration for foveated panoramic sensing," *ACM*

Transactions on Multimedia Computing, Communications, and Applications, vol. 8, no. 2, p. 17, 2012.

[37] P. Baudisch, N. Good, V. Bellotti and P. Schraedley, "Keeping things in context: a comparative evaluation of focus plus context screens, overviews, and zooming," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2002.

[38] C. Plaisant, D. Carr and B. Shneiderman, "Image-browser taxonomy and guidelines for designers," *Software, IEEE,,* vol. 12, no. 2, p. 21–32, 1995.

[39] P. Baudisch, N. Good and P. Stewart, "Focus plus context screens: combining display technology with visualization techniques," in *Proceedings of the 14th annual ACM symposium on User interface software and technology*, 2001.

[40] T. T. Hu, Y. W. Chia, L. W. Chan, Y. P. Hung and J. Hsu, "im-top: An interactive multi-resolution tabletop system accommodating to multi-resolution human vision," in *Proceedings of the 3th IEEE International Workshop on Horizontal Interactive Human Computer Systems*, 2008.

[41] O. G. Staadt, B. A. Ahlborn, O. Kreylos and B. Hamann, "A foveal inset for large display environments," in *Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications*, 2006.

[42] K. W. Chen, C. W. Lin, T. H. Chiu, M. Y. Chen and Y. P. Hung, "Multi-resolution design for large-scale and high-resolution monitoring," *Multimedia, IEEE Transactions on,* vol. 13, no. 6, p. 1256–1268, 2011.

[43] V. K. Singh, P. K. Atrey and M. S. Kankanhalli, "Coopetitive multi-camera surveillance using model predictive control," *Machine Vision and applications,* vol. 19, no. 5-6, pp. 375-393, 2008.

[44] P. Natarajan, T. N. Hoang, K. H. Low and M. Kankanhalli, "Decision-theoretic approach to maximizing observation of multiple targets in multi-camera surveillance," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent*, 2012.

[45] Y. Cai, G. Medioni and T. B. Dinh, "Towards a practical ptz face detection and tracking system," in *Proceeding of IEEE Workshop on Applications of Computer Vision*, 2013.

[46] C. Micheloni, B. Rinner and G. L. Foresti, "Video analysis in pan-tilt-zoom camera networks," *Signal Processing Magazine, IEEE,* vol. 27, no. 5, pp. 78-90, 2010.

[47] C. Micheloni, G. L. Foresti and L. Snidaro, "A network of co-operative cameras for visual surveillance," *Vision, Image and Signal Processing,* vol. 15, no. 2, pp. 205-212, 2005.

[48] R. Jain, D. Militzer and H. H. Nagel, Separating non-stationary from stationary scene components in a sequence of real world TV-images, Hamburg: Institut für Informatik, 1977.

[49] C. R. Wren, A. Azarbayejani, T. Darrell and A. P. Pentland, "Pentland. Pfinder: Real-time

tracking of the human body," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 19, no. 7, p. 780–785, 1997.

[50] T. Horprasert, D. Harwood and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 1999.

[51] J. C. S. Jacques, C. R. Jung and S. Raupp Musse, "A background subtraction model adapted to illumination changes," in *Proceedings of the IEEE International Conference on Image Processing*, 2006.

[52] J. Davis and V. Sharma, "Robust background-subtraction for person detection in thermal imagery," in *Proceedings of the IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2004.

[53] M. Harville, G. Gordon and J. Woodfill, "Foreground segmentation using adaptive mixture models in color and depth," in *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[54] J. M. Guo and C. S. Hsu, "Cascaded background subtraction using block-based and pixel-based codebooks," in *Proceedings of the IEEE 20th International Conference on Pattern Recognition*, 2010.

[55] J. M. Guo, Y. F. Liu, C. H. Hsia and C. S. Hsu, "Hierarchical method for foreground detection using codebook model," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 21, no. 6, p. 804–815, 2011.

[56] O. Barnich and M. Van Droogenbroeck., "Vibe: a powerful random technique to estimate the background in video sequences," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[57] R. Cucchiara, C. Grana, M. Piccardi and A. Prati, "Detecting objects, shadows and ghosts in video streams by exploiting color and motion information," in *Proceedings of the IEEE 11th International Conference on Image Analysis and Processing*, 2001.

[58] R. Cucchiara, C. Grana, M. Piccardi and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, no. 10, p. 1337–1342, 2003.

[59] J. B. Huang and C. S. Chen, "Moving cast shadow detection using physics-based features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[60] A. Sanin, C. Sanderson and B. C. Lovell, "Shadow detection: A survey and comparative evaluation of recent methods," *Pattern recognition,* vol. 45, no. 4, p. 1684–1695, 2012.

[61] S. C. Wang, T. F. Su and S. H. Lai, "Detecting moving objects from dynamic background with shadow removal," in *Proceedings of the IEEE International Conference on Acoustics, Speech*

and Signal Processing, 2011.

[62] A. Amato, I. Huerta, M. G. Mozerov, F. X. Roca and J. Gonzàlez, "Moving cast shadows detection methods for video surveillance applications," in *Proceedings of the Conference on Wide Area Surveillance*, 2014.

[63] Y. S. Qin, S. F. Sun, X. B. Ma, S. Hu and B. J. Lei, "A shadow removal algorithm for vibe in hsv color space," in *Proceedings of 3rd International Conference on Multimedia Technology*, 2013.

[64] S. Nadimi and B. Bhanu, "Physical models for moving shadow and object detection in video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 8, p. 1079–1087, 2004.

[65] J. W. Hsieh, W. F. Hu, C. J. Chang and Y. S. Chen, "Shadow elimination for effective moving object detection by gaussian shadow modeling," *Image and Vision Computing*, vol. 21, no. 6, p. 505–516, 2003.

[66] A. Sanin, C. Sanderson and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *Proceedings of the 20th International Conference on Pattern Recognition*, 2010.

[67] M. Brown and D. G. Lowe, "Recognising panoramas," in *Proceedings of the International Conference on Computer Vision*, 2003.

[68] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends@ in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1-104, 2006.

[69] J. Elder, F. Dornaika, Y. Hou and R. Goldstein, Attentive wide-field sensing for visual telepresence and surveillance, Neurobiology of Attention, 2005.

[70] S. Prince, J. H. Elder, Y. Hou and M. Sizintsev, "Pre-attentive face detection for foveated wide-field surveillance," in *Proceedings of the IEEE Workshop on Applications of Computer Vision/IEEE Workshop on Motion and Video Computing*, 2005.

[71] F. W. Wheeler, R. L. Weiss and P. H. Tu, "Face recognition at a distance system for surveillance applications," in *Proceedings of the 4th IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010.

[72] J. H. Elder, S. J. Prince, Y. Hou, M. Sizintsev and E. Olevskiy, "Pre-attentive and attentive detection of humans in wide-field scenes," *International Journal of Computer Vision,*, vol. 72, no. 1, p. 47–66, 2007.

[73] PENPOWER, "Trackin idvr, auto ptz tracking system," [Online]. Available: http://www.penpower.com.. [Accessed 2012].

[74] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE international conference on Computer Vision*, 1999.

[75] G. Sansoni, M. Carocci and R. Rodella, "Three-dimensional vision based on a combination of gray-code and phase-shift light projection: analysis and compensation of the systematic errors," *Applied Optics,* vol. 38, no. 31, p. 6565–6573, 1999.

[76] P. Pérez, M. Gangnet and A. Blake, "Poisson image editing," *ACM Transactions on Graphics,* vol. 22, no. 3, p. 313–318, 2003.

[77] N. Gracias, M. Mahoor, S. Negahdaripour and A. Gleason, "Fast image blending using watersheds and graph cuts," *Image and Vision Computing,* vol. 27, no. 5, p. 597–607, 2009.

[78] W. Wu, J. Shao and W. Guo, "Moving-object Detection Based on Shadow Removal and Prospect Reconstruction," in *Lecture Note in information Technology*, 2012.

[79] E. Salvador, A. Cavallaro and T. Ebrahimi, "Cast shadow segmentation using invariant color features," *Computer vision and image understanding,* vol. 95, no. 2, p. 238–259, 2004.

[80] S. Zhenhong, S. Jinxia, C. Xiaofang, L. Hui, S. Guofeng, L. Hao and Q. Zhenping, "Shadow removing based on color and texture features," in *Proceedings of International Conference on Consumer Electronics Communications and Networks*, 2011.

[81] A. Prati, I. Mikic, M. M. Trivedi and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, no. 7, pp. 918-923, 2003.

[82] X. Wang, T. X. Han and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, 2009.

[83] A. Alahi, D. Marimon, M. Bierlaire and M. Kunt, "A master-slave approach for object detection and matching with fixed and mobile cameras," in *In Proceeding of the 15th IEEE International Conference on Image Processing*, 2008.