NEW SIFT-BASED CALIBRATION METHODS

FOR HYBRID-CAMERA SYSTEM

LOW YI QIAN

MASTER OF ENGINEERING SCIENCE

FACULTY OF ENGINEERING SCIENCE
UNIVERSITI TUNKU ABDUL RAHMAN
DECEMBER 2013

NEW SIFT-BASED CALIBRATION METHODS

FOR HYBRID-CAMERA SYSTEM

By

LOW YI QIAN

A dissertation submitted to the Department of Electrical and Electronic Engineering,

Faculty of Engineering Science,

Universiti Tunku Abdul Rahman,

in partial fulfillment of the requirements for the degree of

Master of Engineering Science

December 2013

ABSTRACT

NEW SIFT-BASED CALIBRATION METHODS FOR HYBRID-CAMERA SYSTEM

LOW YI QIAN

Camera system is an important system in order to observe and capture daily activities, human and environmental behavior and etc. Traditionally, camera system mainly employs the usage of static cameras. Static cameras usually provide low resolution and poor image quality in modeling. With the enhancement of technology, the functions of camera system have been gradually increasing. For example, Pan-Tilt-Zoom (PTZ) camera is able to obtain multi-angles of views and multi-resolution information; it provides more functionality as compared to static camera. In a camera network, two or more various types of camera are employed in order to obtain different perspective and view of a scene (e.g. in a shop).

On the other hand, camera calibration is one of the challenging stages in camera network to enable different cameras to connect and communicate to each other together. This is because information of the positions of the cameras, focal length, different scaling factors for pixels and lens distortion are needed in order for the connection to be established. One of the main challenges in camera calibration is to obtain an accurate and fast estimation of disparities between two different views in a scene. Besides disparities, differences between the two views might be due to occlusion of the object, specular reflection, sensor noise and various other causes. The traditional approach of calibration takes longer time because it needs manual input or reference objects to find match between correspond images. Therefore, the recent works by some researchers have brought remarkable increase of automation to these problems.

In this project, a new hybrid camera system has been designed and constructed. Instead of using two static cameras, our hybrid camera system consists of a static wide angle camera and a PTZ camera. Both cameras obtain different optical elements and resolutions. We proposed a master-slave concept to represent both cameras. The static camera will be used as a master camera with wide angle view. It is used to monitor the environment from a distance. On the other hand, the PTZ camera will be used as a slave camera. It is used to perform panning, tilting and zooming. The PTZ camera will point towards the region of interest (ROI) in high resolution as well as providing different parameters of PTZ values. We also proposed a new calibration method for our hybrid camera system. In particular, we employed the Scale Invariant Features Transform (SIFT) algorithm in our work. SIFT is a popular feature extraction algorithm used to detect and describe keypoints. However, before performing the calibration between static camera and PTZ camera, PTZ camera needs to acquire the images from different values of parameters. Therefore, we adopted

image stitching approach to create a panorama image before proceed to image calibration between the static camera images with panorama images.

Another major contribution of this project is the increase of the determination and detection rate in image calibration. The affine transformation such as Hough transformation and RANSAC were adopted to identify the positive and negative keypoints. This helps to calibrate both static master camera and PTZ camera images without the use of any objects as reference. By eliminating the use of reference objects, it enables higher performance and efficiency in handling the estimation of disparities between different views in a scene from different angle and scale. Therefore, the obtained empirical results will be of higher accuracy.

ACKNOWLEDGEMENT

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Prof. Dr. Lee Sze Wei, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I would like to thank Prof. Dr. Goi Bok Min, who let me experience the research of freshwater mussels in the field and practical issues beyond the textbooks, patiently corrected my writing and financially supported my research. I would also like to thank Mr. Ng Mow Song for guiding my research for the past several years and helping me to develop my background in physiology, biochemistry, and ecology. Special thanks go to Dr. Tay Yong Haur was always willing to help and give his best suggestions. My research would not have been possible without their helps.

Lastly, I very appreciate to all my fellow friends who were willing to participate in my final defense committee at the last moment.

APPROVAL SHEET

This thesis/dissertation entitled "**NEW SIFT-BASED CALIBRATION METHODS FOR HYBRID-CAMERA SYSTEM**" was prepared by LOW YI QIAN and submitted as partial fulfillment of the requirements for the degree of Master of Engineering Sciences at Universiti Tunku Abdul Rahman.

Approved by:

____
(Prof. Ir. Dr. Lee Sze Wei)

Date:…………………..
Professor/Supervisor
Department of Electrical and Electronic Engineering
Faculty of Engineering Science
Universiti Tunku Abdul Rahman

____
(Prof. Dr. Goi Bok Min)

Date:…………………..
Professor/Co-supervisor
Department of Electrical and Electronic Engineering
Faculty of Engineering Science
Universiti Tunku Abdul Rahman

FACULTY OF ENGINEERING SCIENCE
UNIVERSITI TUNKU ABDUL RAHMAN

Date: ____________________

SUBMISSION OF THESIS / DISSERTATION FOR EXAMINATION

It is hereby certified that **LOW YI QIAN** (ID No: **10UEM2136**) has completed this thesis/dissertation* entitled "**NEW SIFT-BASED CALIBRATION METHODS FOR HYBRID-CAMERA SYSTEM**" under the supervision of Prof. Ir. Dr. Lee Sze Wei from the Department of Electrical and Electronic Engineering, Faculty of Engineering Science, and Prof. Dr. Goi Bok Min from the Department of Electrical and Electronic Engineering, Faculty of Engineering Science.

I understand that the University will upload softcopy of my thesis/dissertation* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

____

(LOW YI QIAN)

DECLARATION

I, LOW YI QIAN hereby declare that the thesis/dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

____

(LOW YI QIAN)

Date ______________________

TABLE OF CONTENT

Titles **Page**

LISTS OF TABLE(S)

LISTS OF FIGURE(S)

CHAPTER 1

INTRODUCTION

1.1 Background and Introduction

Camera system plays an important role in video surveillance systems. Video surveillance is an application that requires embedded image capture capabilities that allows video images or extracted information to be compressed, stored or transmitted over communication networks. In surveillance scenarios, it is hard to monitor environment entirely using single camera type i.e. static camera or pan-tilt-zoom (PTZ) camera.

Generally, various types of cameras are used to form a camera network. Camera networks can be designed for indoor and outdoor purposes. In fact, different angles of viewpoint captured by a camera can result in large differences in term of the appearance of the moving objects due to illumination, casted shadows, occlusion, etc. These significantly affect the performance of object detection and tracking as well as of object recognition. To solve these problems, camera systems are employed in order to acquire multiple views of an environment from different angles and zooming levels. More information could be collected as compared to using a single camera capturing single view direction. For example, in a shop, we need to install a number of good quality cameras in order to capture the situation inside the shop, staffs and customers as a whole from all views. Thus, using PTZ cameras in the design of camera system will be very advantageous as they allow pan-tilt and high resolution zooming. Pan-tilt function allows the camera to capture a wider range of views in the shop. This is beneficial as it helps to reduce the number of camera usage inside a shop. Furthermore, it reduces the cost of camera installation. On the other hand, high resolution zooming allows better detection of thefts. It also enables better face recognition of intruders should there be any incidents happened in the shop.

Camera network obtains two or more perspectives of slight relative displacement of objects in the multiple monocular views of scene. By comparing the information of the image from vantage points, we can implement 3D reconstruction, scene analysis and other depth related applications (Dingrui Wan & Jie Zhou, 2008; D. Forsyth & J. Ponce, 2003). Conventionally, camera network research uses static cameras to achieve lower cost and relative simplicity in modeling (Dingrui Wan & Jie Zhou, 2008). Although a fixed static camera network has been applied successfully in real application scenarios, it suffers from the inherent problem of sensor quantization. The fixed optics and fixed sensor resolution share most of the information. Nonetheless, there are differences that are caused by occlusions of objects, specular reflections, which move independently of the surfaces of objects, sensor noise, etc. (R.D. Henkel, 1997).

For instance, the employment of more cameras can help to increase the monitoring coverage area. However, these may also increase the camera cost, processing power and architectural complexity of the application. Recently, high resolution camera such as PTZ camera offered an alternative way to obtain the understanding of the dynamic scenes to fixed static camera. It enables the attainment of different angles of view and resolution information at any time. Theoretically, PTZ camera is capable of monitoring a wide area with the acquisition of images of fine quality. However, we can only be focus on one area at a time just like any other static camera when it is without human supervision. Incidents might have happened outside the instantaneous views of a PTZ camera even though it happened within its coverage area as shown in Figure 1.1.

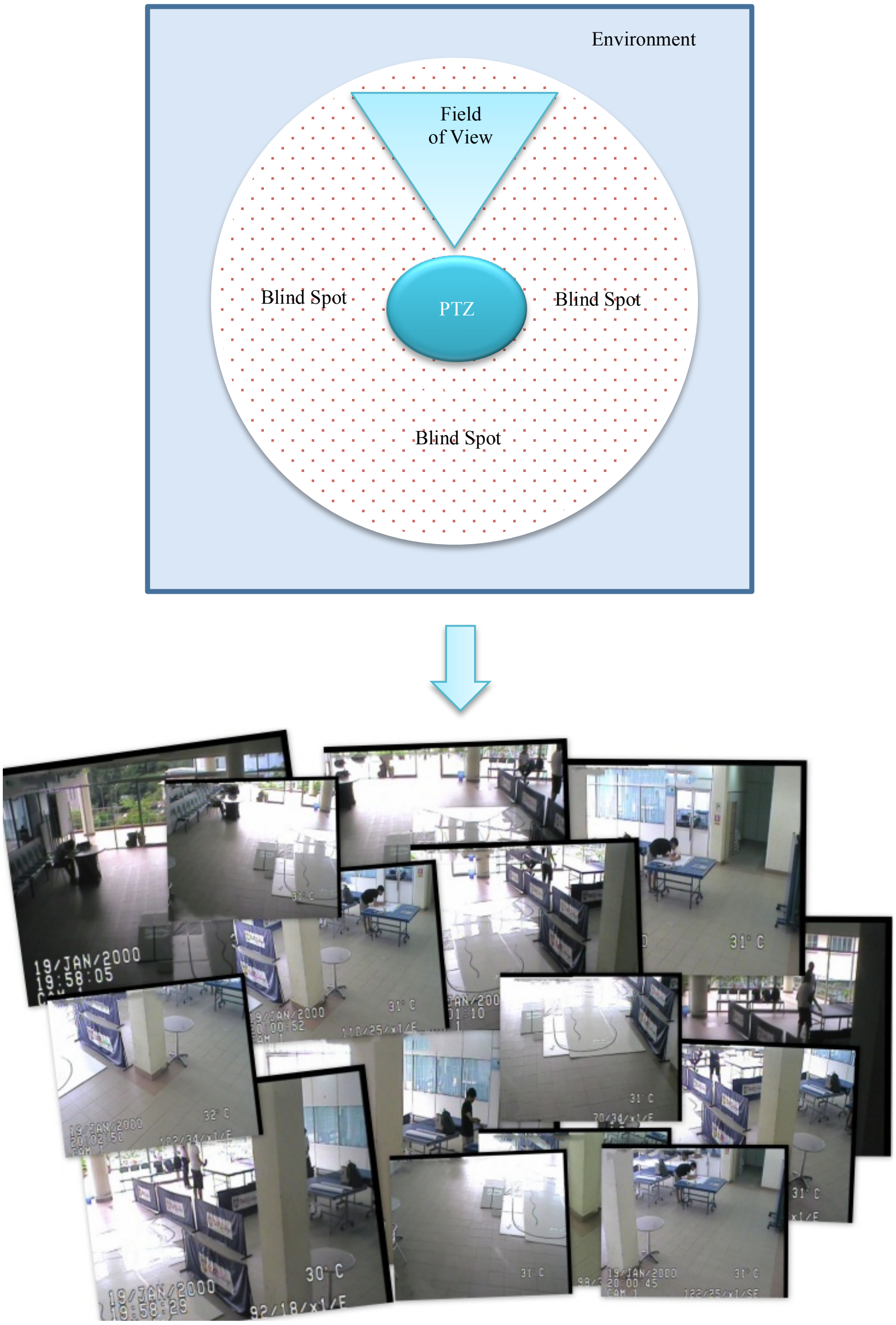


Figure 1.1: Sample images acquired by PTZ camera in limited angle of view.

1.2 Motivation

In this project, a hybrid camera system which consists of a static camera and pan-tilt-zoom (PTZ) camera was designed as shown in Figure 1.2. Both cameras obtain different types of optics and sensor resolutions. The static camera plays the important role of monitoring a wide area at a distance. On the other hand, the PTZ camera captures multi-scale view images, moving from a region of interest (ROI) to another with different intrinsic and external parameters. Hybrid camera system is more challenging as the intrinsic and external parameters of each camera can be changed during operation compared to the dual-static-camera system. A hybrid-camera system is particularly useful in object detection, recognition and tracking. The static camera is set to have an overall view of the scene such that several entities can be tracked simultaneously. The PTZ camera is used to follow the target trajectory and generate close-up imagery of the entities.

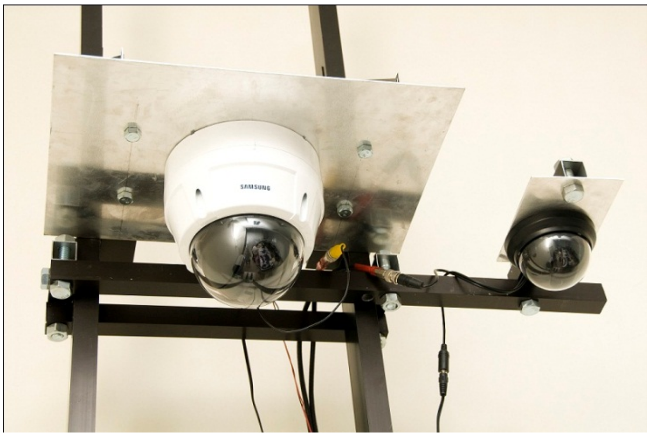


Figure 1.2: Hybrid camera system which is combined by PTZ camera and static camera.

The approach taken on hybrid camera calibration also differs from conventional methods where matching is perceived as an operation on multiple images which are acquired from PTZ camera to create a panoramic image and not just an image pair. The associated problems and conditions are stated as follows:

Problems and Conditions:

i. Field of view (FOV) is a measure of how large an area a camera is capable of viewing. The FOV is based on the camera optical lens. For example, in a 15’ x 15’ room as shown in the Figure 1.3 below is a static wide angle camera using 4mm lens (green arrows) allows better wide-angle viewing coverage than a PTZ camera. In a hybrid camera system where a close-up view is needed, a PTZ camera is able to perform dynamic-view-angle and dynamic-scales from 4mm to 12mm (from blue arrows to red arrows) FOV. The images acquired with different focal-length lenses may cause different distortion and FOV compared to the wide angle camera.

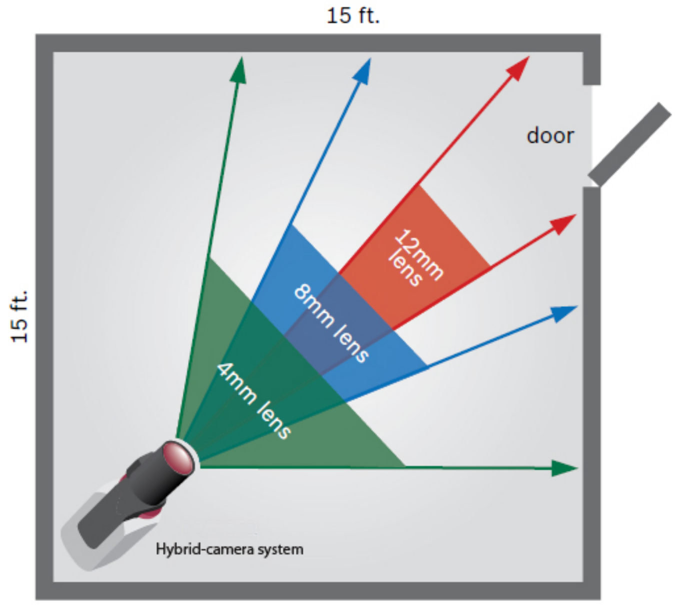


Figure 1.3: Field of view with different focal-length lenses.

ii. As shown in Figure 1.4, hybrid camera geometry imposes constraint on the finding out of the positions of the PTZ camera's views. PTZ camera enables flexibility for pan, tilt and zooms while wide angle camera allows static image capturing. The challenge is to identify the relationship of both cameras in the camera network.

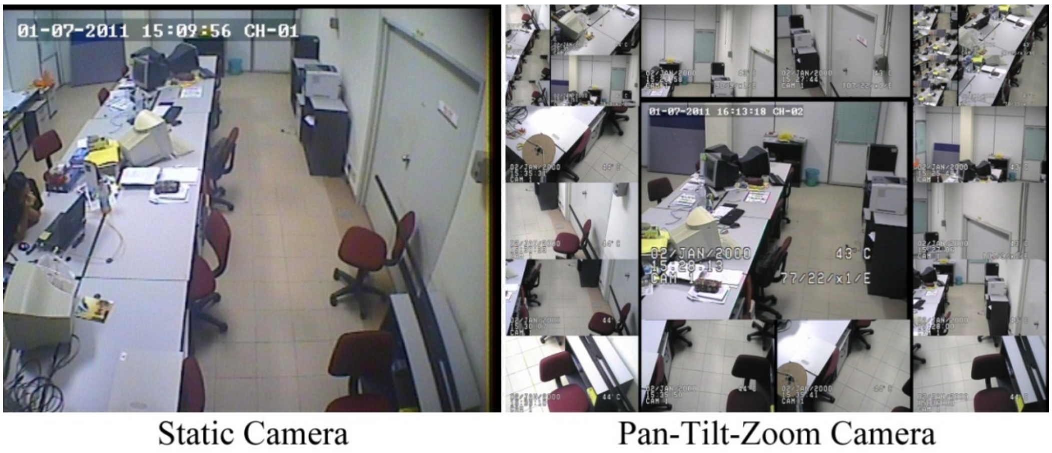
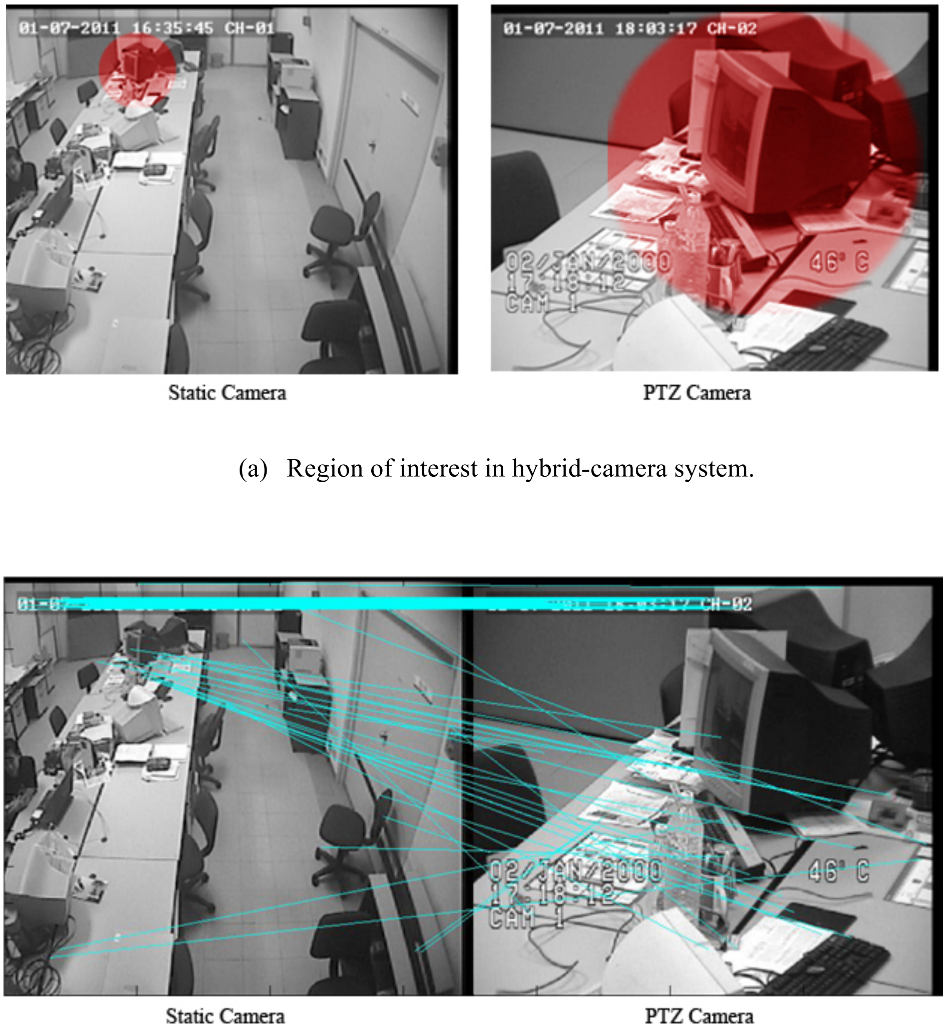


Figure 1.4: Relationship of camera network in hybrid-camera system.

iii. Figure 1.5(a) shows the current view of wide angle camera and PTZ camera in certain angles and focal-length. Figure 1.5(b) show the matching of regions of interest in the image from PTZ camera to the static image of the wide angle camera by using Scale Invariant Features Transform (SIFT) (Lowe D.G., 1999) feature matching method. It is apparent that the method resulted in significant mismatched pairs.



(a) Region of interest in hybrid-camera system.

(b) Result of image matching process in hybrid-camera system based on SIFT algorithm.

Figure 1.5: Significant mismatched pairs was detected in calibration process.

1.3 Objective

The main goal of the study is to design the implementation of semi-active stereo vision system and examine its effects on image calibration accuraccy and performance. This development is combination of a static wide angle camera and a Pan-Tilt-Zoom camera. From our study we believe that the conduction of the study might provide valuable information to the development of multiple camera calibration system as an alternative system. Multiple camera calibration is crucially important to semi-active stereo vision because images are obtained from two different types of optics and sensor resolutions, and there is a need to correct the distortion especially on the wide-angle images which are acquired by the static camera. The background appearance also changes dynamically as the PTZ camera pans, tilts and zooms. (X. Zhou, R. Collins, T. Kanade & P. Metes, 2003). Besides, we also hopes that the research of implementation of hybrid camera system will contribute to set up a camera calibration method for more complex camera networks.

In hybrid camera system, there are at least three concerns to be addressed in semi-active stereo vision calibration:

i. Image stitching process based on SIFT algorithm: PTZ camera performs dynamic-view-angle and dynamic-scales where each orientations and zoom levels may have differences caused by occlusions of objects, specular reflections which move independently of the surfaces of objects and sensor noise from different perspectives.

ii. Hybrid-camera calibration process: each camera has different type of optics and sensor resolution and although they may capture similar information, but all obtain images with different distortions.

iii. Automated calibration process: instead of calibrating the hybrid-camera manually, we have developed a test bed system with algorithm and mechanism that automatically perform calibration under different environments and conditions. The challenge is how we can improve the accuracy of stitching process in the automated system.

1.4 Contribution

The main contribution of this thesis is the development and evaluation of a hybrid camera system capable of recognizing and stitching a panoramic image automatically. We evaluate a fully automated 2D panoramic stitching method. This method has the following advantages over previous cameras calibration approaches:

i. Enhanced robustness to image zoom, rotation and exposure changes in images especially those from PTZ camera, due to the use of invariant features.

ii. Automatic detection of matching images, using a sequence model for image matching.

iii. Rendering of panoramic images based on the mean values of the pixels.

A popular filtering technique, Random Sample Consensus (RANSac) has also been studied. It was incorporated into the system to process data associated to inliers and outliers to improve the accuracy of correspondence process that estimates the position of Field of View (FoV)

Finally, we implemented and evaluated the calibration method on hybrid camera system based on SIFT features on our test bed.

1.5 Outline of Thesis

The remainder of the thesis is organized as follows. Chapter 2 reviews the literature on image information and background study of multi-view geometry. Scale Invariant Features Transform (SIFT) and analysis capable of features detection and matching among correspondence is introduced. In Chapter 3, a detailed analysis of PTZ camera controls and panoramic image stitching methods is presented. This chapter extends the image calibration of hybrid camera system to estimating the position of current view in panoramic image. A test bed for data collection was set up to perform automatic panoramic image stitching in Chapter 4. In chapter 5, conclusion and suggestions for future work are presented.

CHAPTER 2

LITERATURE & BACKGROUND STUDIES

2.1 Introduction

In recent technology development, computer vision concepts, theories and models have been widely applied in many commercial and industrial systems e.g. modeling of objects or environment, process control, object detection, object recognition, etc. One of the important tool to reduce the crime rate is to apply the computer vision into the monitoring system. Thus, camera networks have been implemented to monitor the behaviour and activities of human.

In this chapter, the fundamentals of basic image information and perspective image formation model of Aghajan, H. et al. (2009) which accurately reflect the phenomena observed in image taken by real cameras will be presented. On pictures taken by cameras it is possible that the rectangular or circular shapes look like ellipses. Such situation happens due to two perpendicular radii of a circle being stretched by different amounts by perspective projection. Angles, distances, ratios of distances are the key features that need to be preserved in projective geometry. Throughout the chapter, we represent object points by $X = (x, y, z)$ and image plane by $u = (x, y)$.

Computer vision is a field that involves the processing, analysis and understanding of images, which are the extracted properties of the 3-dimensional (3-D) world in order to achieve results and effects similar to human vision. Human vision is 3-D but the objects captured by camera are converted from 3-D to 2-D image. Conversion from 3-D to 2-D involves information loss due to the perspective views which is illustrated in Figure 2.1. In digital world, we need to understand an image in digital perception. An image may be defined as a 2-dimensional function, *f*(*x,y*), where *x* and *y* are spatial coordinates, and the amplitude of *f* at any pair of coordinates (*x,y*) is called intensity of the image at that point (Gonzalez, R.C. and Woods, R.E., 2007). On the other hand, human vision possesses the ability to perceive depth. Since human eyes are separated in space, each receives a slightly different image, and the different positions of corresponding points in these images which can be used to judge and perceive depth. The process to obtain accurate and fast estimate between the disparities of two different views of scene is known as stereo vision. Stereo vision is also defined as two different perspectives of human eyes that lead to slight relative displacement of objects in the two monocular views of scene.

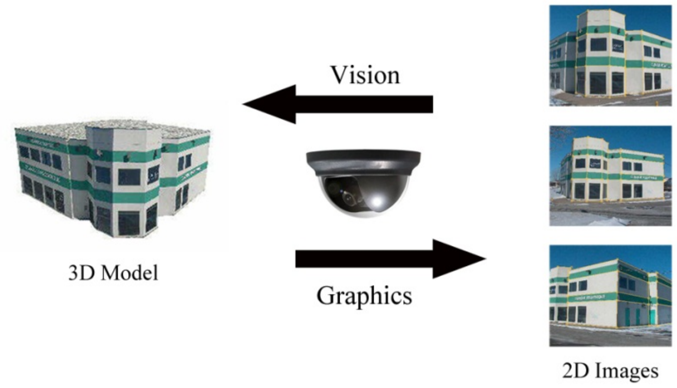


Figure 2.1: Differences between 2D Modeling and 3D Modeling.

The basic principle in stereo vision is to find out the correspondence of the images among the coordinates and intensity values by comparing the information of the images from two vantage points and it can be used in 3-D reconstruction, scene analysis and other depth related usage (Forsyth, D. et al., 2003; Wan, D. et al., 2008). Several approaches to find image correspondences have been proposed, for example, contours-based object detection (Shotton, J. et al., 2005; Yokoyama, M. et al., 2005).

In this work, a feature-based approach in Liu, J. and Hubbold, R. (2006) has been studied and adopted for image stitching and matching. Features detection in computer vision has been widely studied. Object detection based on low-level feature such as Canny Edge by Canny, J. (1986) and Harry Corner by Harris, C. and Stephens, M.J. (1988) were widely used in machine learning and image processing. In the past 10 years, extensive research has been carried out on edges and corners. Based on the research on edges and corners, Lowe D.G (1999) has developed a more complex and distinctive image features algorithm called Scale Invariant Feature Transform (SIFT).

2.2 Image Information with Geometry Perspective

We consider an ideal situation in which an imaging device, "*pinhole*" cameracan capture accurately the geometry of perspective projection as shown in Figure 2.2. We represent the coordinate of an object point *X* in the system as $X_o = (x_o, y_o, z_o)$. Lights enter the camera thru an extremely small aperture. The intersection of the lights with the image plane will form the image of the object, this is called *perspective projection*.

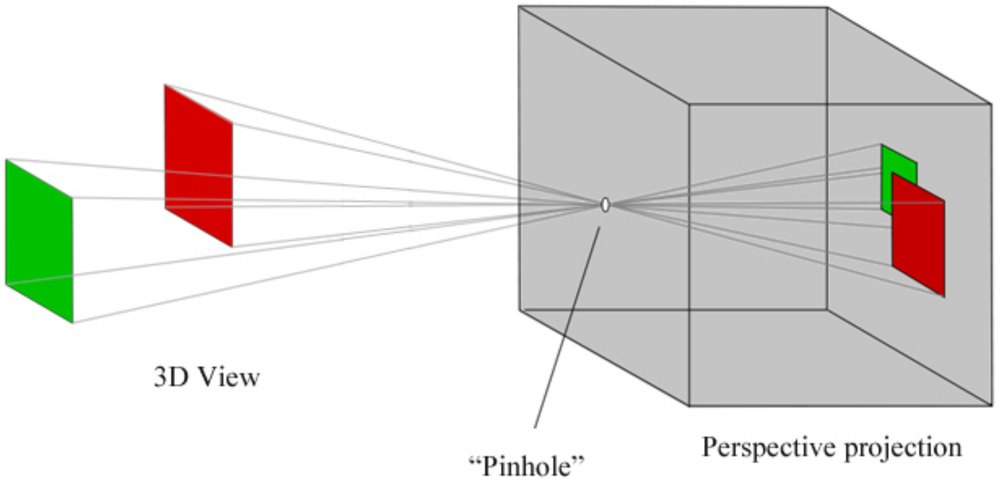


Figure 2.2: The basic geometry of a pinhole camera.

Generally, each camera coordinate has an associated image plane located in the camera coordinate system which is not aligned with the surrounding world. The image plane inherits a natural orientation with two-dimensional projection.

For real cameras, the relationship between the information of image points and the information of the object points is more complicated. To simplify the derivation of the perspective projection equations, a few assumptions as follows have been made:

a. The camera axis (optical axis) is aligned with the world's z-axis
b. No image inversion assuming that the image plane is in front of the center of projection.

c. Object points have the same information regardless of the viewing angle and information of an image point is the same as the information of a single corresponding object point. However the information of each point is different because of factors in a real imaging system.

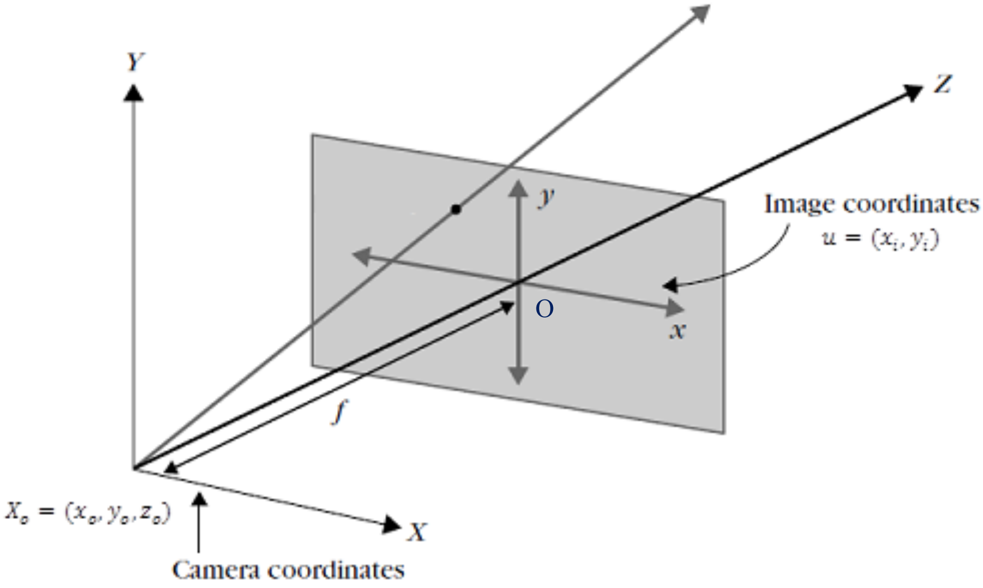


Figure 2.3: A modern camera projects 3-D world into 2-D image plane through perspective projection.

An object point $X_o = (x_o, y_o, z_o)$ is projected onto the image plane P at the point $u = (x_i, y_i)$ in Figure 2.3. The model consists of a plane (image plane) and a 3-D point O *(center of projection)*. Focal length is the distance f between the image plane and the center of the projection O, for example the distance between the lens and the CCD array. Optical axis is the line through O and perpendicular to the image plane. The intersection of the optical axis with the image planeis called the *principal point* or *image center*.

The equation of perspectives projection is given below

$$X = \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = R \left(\begin{bmatrix} x_o \\ y_o \\ z_o \end{bmatrix} - C \right) \qquad \text{[Eq. 2.1]}$$

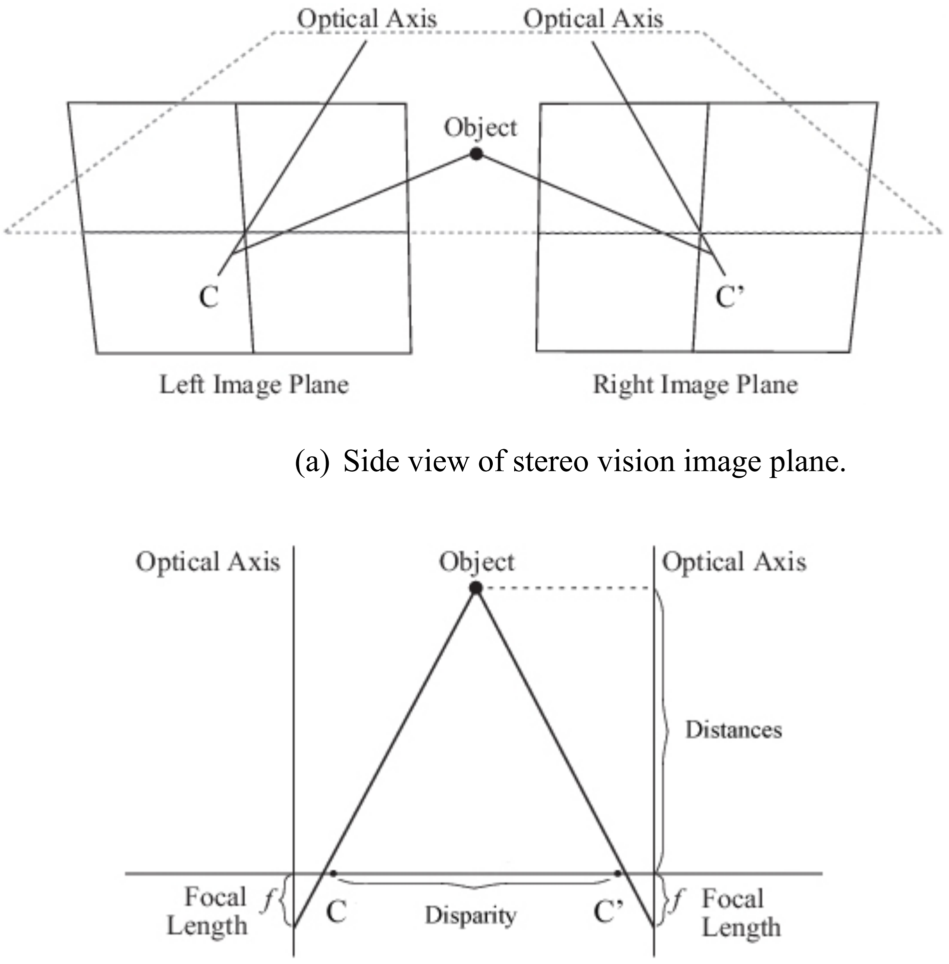
where C is the image center projection, R is the orientation matrix and f is the focal length.

A scene point $X = (X_c, Y_c, Z_c)$ is projected onto the image plane P at the point $u = (x_i, y_i)$ by the perspective projection equations

$$x_i = f\frac{X_c}{Z_c} \qquad y_i = f\frac{Y_c}{Z_c} \qquad \text{[Eq. 2.2]}$$

2.3 Stereo Camera Network Calibration

In camera networks, stereo vision is one of the most challenging parts of computer vision. The focus is on geometric models of perspectives cameras. Generally, we analyse the image relationships from the same static scene by two cameras *C* and *C'*, as shown in Figure 2.4.



(a) Side view of stereo vision image plane.

(b) Top view of stereo image plane

Figure 2.4: Concept of stereo vision image plane (Lee, C.W. et al., 2009)

In camera calibration we acquire images from two or more cameras from different perspective views. One of the most popular camera setup is stereo vision which has always been an important issue of interest in computer vision. “Stereo vision is the two different perspectives of our two eyes that lead to slight relative displacement of objects in the two monocular views of scene. Among the advantages of stereo vision is 3D modeling of scene and depth estimation (Lee, C.W. et al., 2009)”. “Stereo cameras calibration allows transferring 3D object locations onto the camera image plane and therefore using this information to steer the pan tilt and zoom into the appropriate direction (Bimbo, A.D., et al., 2010)". In recent years, many methods have been proposed in stereo vision calibration (Baker P. et al., 2000; Krumm J. et al., 2000; Senior, A.W., et al., 2005; Liu J. et al., 2006; Xing Y.J. et al., 2007; Gonzalez, R.C. et al., 2007; Wan, D. et al., 2008; Liu R. et al., 2009; Memont, Q. et al., 2011).

In video surveillance system, it is mainly about identifying the location and identity of people in the room. Krumm J. et al. (2000) and his research group have developed an intelligent system called EasyLiving. In EasyLiving room, multiple people can be tracked using stereo cameras rather than single static camera and this makes it easier to segment the overlapping people in the room.Two method of making measurement was proposed. First method is an interactive program which allowed user to establish correspondences and ground plane points by click on points in images from two cameras. Another method records, from each camera, the 2D ground plane locations of a person's path when the person moves around in the room, the translation and rotation will be calculated by a calibration program to give the best overlap between the two paths.By knowing the cameras' relative positions and orientations in the ground plane, it creates a platform to integrate more vision-based tracking for example to find the point by walking from place to place with whole-body tracking.

Camera calibration is challenging because handling nonlinearity requires good initial estimates on stereo vision application. Neural networks can be applied in many scientific disciplines to solve variety of problems in pattern recognition, prediction, and optimization. In the work of Xing, Y.J. et al. (2011), the researchers proposed multilayer artificial neural networks (ANNs) model for the training needed to correspond a variety of stereo-pair images and 3-D world coordinates. This approach is taken because camera calibration is a nonlinear problem and cannot be solved with a single layer neural network. They evaluated the experiment with no fixed rules for an ideal network model. The approach requires the training of the neural network for a set of matched image points of which the correspondences are known. This approach is different from conventional camera calibration techniques in the notice that no extrinsic or intrinsic parameters are required. Instead, the system is trained such that it learns to directly find the correspondences of the objects. The advantage of this approach lies mainly in its simplicity and generality. It works for any type of camera modeling. The results from Memont Q. et al. (2011) as shown in Figure 2.5 and 2.6 claimed that error became less than 5% after 40 000 epochs of training. After 100 000 epochs of training, the mean error of a point became 4.33%.

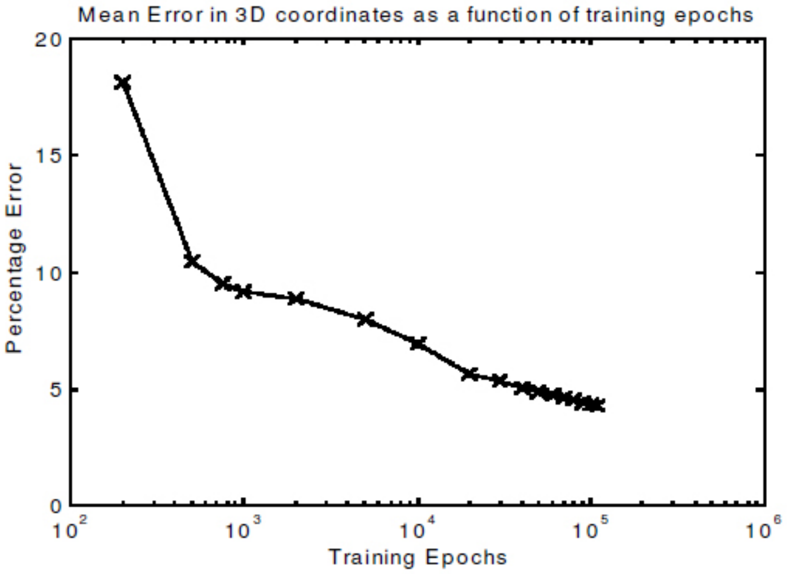


Figure 2.5: Mean percentage error in computing 3D coordinates as a function of the number of epochs (Memont Q. et al., 2011)

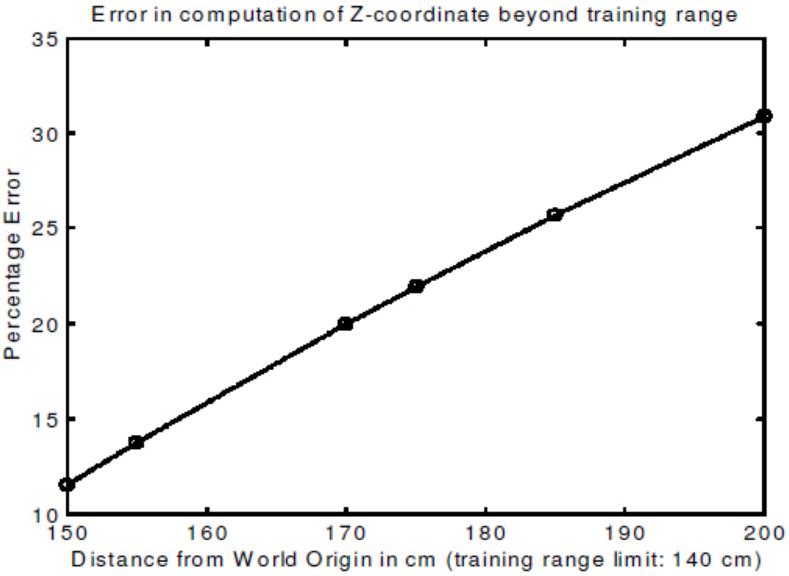


Figure 2.6: Percentage error in the computation of the Z coordinate beyond the training range (results taken after training the network for 50 000 epochs). (Memont Q. et al., 2011)

Besides that, several methods have been published in the literature to perform calibration of PTZ cameras. Calibration was carried out by estimating the homography among cameras in a home position taking into account the effects of pan and tilts controls in Senior, A.W. et al (2005). Firstly, they applied manual registration at each stage and set up a corresponding look-up table. Based on the initial system setup, they implemented an auto-learning mechanism of homography between the cameras in a home position. The system generally supports arbitrary combinations of multi-camera. However, the method lacks noise filtering to increase the accuracy on homography between two views.

In Chen, I.H et al. (2007) a method for the calibration of multiple cameras by estimating the tilt angle and the altitude of the each PTZ camera based on observation of some clues revealed in the captured images was proposed. The technique can simply place a few simple patterns on horizontal plane for example A4 paper, books, boxes, etc. on the table to calibrate multiple PTZ cameras. A specific calibration patterns is not required in the system setup. The advantage of the method is the comprehensible sense of camera pose. The tilt, altitude, and orientation of the camera offer a more direct physical sense about the camera pose in the 3D space, especially when the PTZ cameras are under panning, tilting and zooming operations from time to time.

Among others, the solutions proposed in Liu, J. et al. (2006), Wan. D. et al. (2008), Liu. R. et al. (2009), and Bimbo, A.D. et al. (2010) do not require any calibration patterns. An algorithm, “Scale Invariant Features Transform (SIFT) (Lowe D.G., 1999) can address the matching problems with translation, rotation and affine transformation among different images”. SIFT is a good method to extract feature point with representative feature descriptor and more stable features matching ability for images which are captured from random different angles. In Bimbo, A. D. et al. (2010) images from non-calibrated PTZ camera at different values of pan, tilt and zoom are acquired to build a panoramic image by using SIFT. Finding the right correspondence between the images helps to evaluate homography and localize the camera with respect to the scene. SIFT enables us to transform image data into scale-invariant coordinates relative to local features and store them in database.

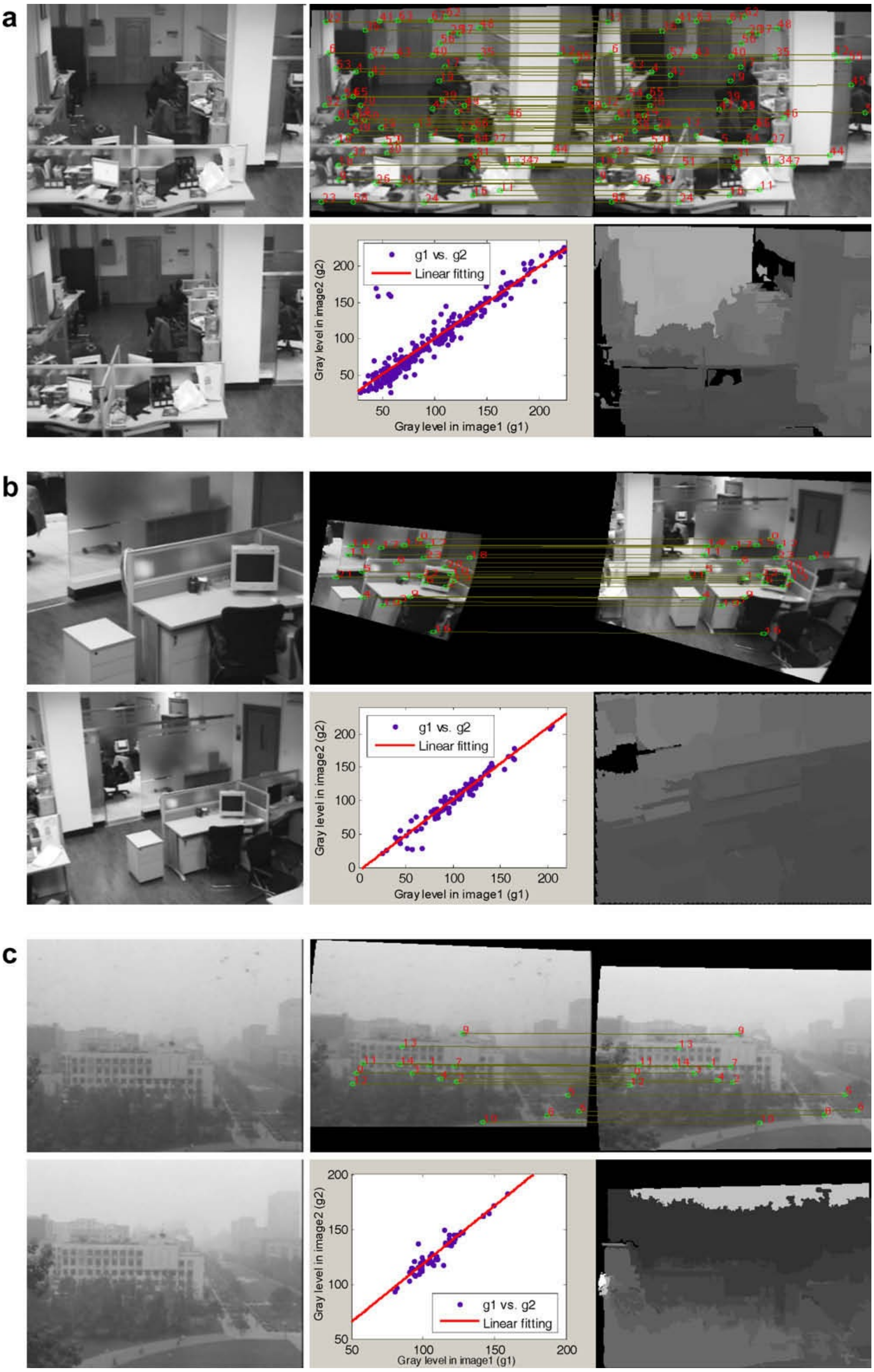


Figure 2.7: Stereo vision calibration by using two PTZ cameras with different pan, tilt and zoom parameter. (Bimbo, A. D. et al., 2010)

In figure 2.7, three sets of experiment have been tested by using two PTZ cameras with random pan, tilt and zoom parameter. SIFT algorithm behaves as a local image operator which transforms an image into a collection of local features (Lowe, D.G., 2009). To find corresponding features between the two images, different feature matching approaches can be used. According to the nearest neighbor algorithm, an image feature, F_2^i searches for the corresponding feature in model image feature, F_1^i. The corresponding key points consist of the smallest Euclidean distance (Gower, J.C., 1982) between featureF_1^i and F_2^i or match M (F_1^i,F_2^i). It can perform rapidly to find out the correct match of the keypoints descriptors with good proximity in large database of features. However, in a cluttered image, many features may affect the accuracy of matching and give rise to false matches of key points. The determination of these consistent clusters can be accomplish by using an effective hash table feat of the generalized Hough transform (P.V.C., 1962). Consistent interpretation using each feature from clusters of features determined through the Hough transform, to vote for entire objects poses the feature consistency. Theinterpretation probability for being correct is much more higher compare to the single feature during clusters of features vote for the same pose of an object.

In P. Baker et al. (2000) a calibration procedure was proposed for a multi-camera rig. The technique forms a multi-frame structure from motion algorithm based on point correspondences constructed with accuracy to within a pixel. It is not necessary for all the cameras to focus on same common field of view. As long as every camera is connected to each other a large set of correspondences can be constructed even in low light environment. A large, non-linear eigenvalue minimization routine will require the correspondences based on the epipolar constraint. All points' correspondences between every pair of cameras will be encapsulate by the eigenvalue matrix in a way that minimizes the smallest eigenvalue results in the projection matrices within single perspective transformation. It was claimed that the method was extremely accurate with the accuracy of the re-projection of the reconstructed points within a pixel uncertainty, which was the measurement error of the location of the LED.

2.4 Scale Invariant Features Transform (SIFT)

Scale Invariant Features Transform (SIFT) (Lowe D.G., 1999) is an approach to transform an image into a large collection of local feature vectors, each of which is invariant to image translation, scaling, and rotation, and partially invariant to illumination changes and affine or 3D projection. The scale-invariant features are efficiently identified by using a staged filtering approach. The following are the major steps that used to determine the image features:

a) **Scale-space extrema detection:** The first step is to search the over all scales and image locations. In this step, Difference-of-Gaussian is being used to identify potential interest points that are invariant to scale and orientation.

b) **Key point localization:** At each potential interest point, a detailed model is fitted to determine the scale and location. Meanwhilte, key points are selected based on their stabilities measurement.

c) **Orientation assignment:** Each key point location will be assigned with orientations to the local image gradient direction. By having invariance to the image transformations, operations on image data performed relative transformation to the assigned orientation, location for each feature and scales.

d) **Key-point descriptor:** At the vicinity of key point, the local image gradients at selected scale are measured and transformed into permissible local shape distortion and illumination change representational format.

An significant aspect of this approach is that it creates large numbers of features that densely cover the figure over the total range of scales and locations. SIFT features are first extracted from a set of reference images and stored in a database. A typical image of size 480x360 pixels will detect at least 1500 or more stable features depending on the parameters that have been set. The quality of features is particularly important for object recognition, where the ability to detect small objects in cluttered backgrounds requires at least three features be correctly matched from an object for reliable identification.

In image matching and recognition, comparison is performed between the current and previous image in term of image features based on the database of feature vectors of previous images. Nearest neighbor algorithm is a features search technique. This technique match features of current image to those of previous images by defining the key points with minimum Euclidean distance (Gower J.C., 1982) from the given descriptor vector. An accurate probability is defined by the ratio of distance between the targeted neighbour with another second closest target. It can operate rapidly to determine the correct match of the key point descriptors with good proximity in large database of features. However, in a cluttered image, many features may affect the accuracy of correct matches and rise false matches of the key points.

2.4.1 Scale-space extrema detection

In SIFT, the first stage of the key point detection is to identify candidate locations and scales that are being assigned repeatedly under different views of the same object. Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as Gaussian scale space of Lindeberg, T. (1994).

Therefore, the scale space of an image is defined as a function, $L(x, y, \sigma)$, that is produced from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \,, \qquad \text{[Eq. 2.3]}$$

where * is the convolution operation in *x* and *y*, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \, e^{-(x^2+y^2)/2\sigma^2} \qquad \text{[Eq. 2.4]}$$

The scale-space extrema in the difference-of-Gaussian function, $D(x, y, \sigma)$, is then convolved with the image to detect the stable keypoint locations in scale space efficiently:

$$D(x, y, \sigma) = \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) * I(x, y)$$

$$= L(x, y, k\sigma) - L(x, y, \sigma) \qquad \text{[Eq. 2.5]}$$

where the two nearby scales are separated by a constant multiplicative factor *k*.

The reasons to apply this function is to get a smoothed image *L*, which needs to be performed for scale space feature description, and *D* can therefore be performed by simple image subtraction.

Additionly, the difference-of-Gaussian function provides a precise approximation to the scale-normalized Laplacian of Gaussian, $\sigma^2\nabla^2 G$. It produces the most stable image features compared to a range of other possible image functions, such as gradient, Hessian, or Harris corner function.

The relationship between *D* and $\sigma^2\nabla^2 G$ can be understood from the heat diffusion equation:

$$\frac{\partial G}{\partial \sigma} = \sigma^2\nabla^2 G \qquad \text{[Eq. 2.6]}$$

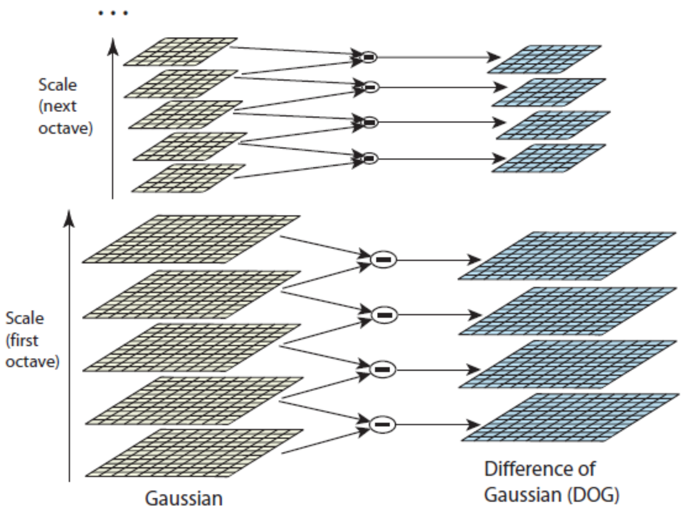
From this, $\nabla^2 G$ can be computed from the finite difference approximation to $\partial G / \partial\sigma$, using the difference of nearby sales at $k\sigma$ and σ:

$$\sigma^2\nabla^2 G = \frac{\partial G}{\partial\sigma} \approx \frac{G(x,y,k\sigma)-G(x,y,\sigma)}{k\sigma-\sigma} \qquad \text{[Eq. 2.7]}$$

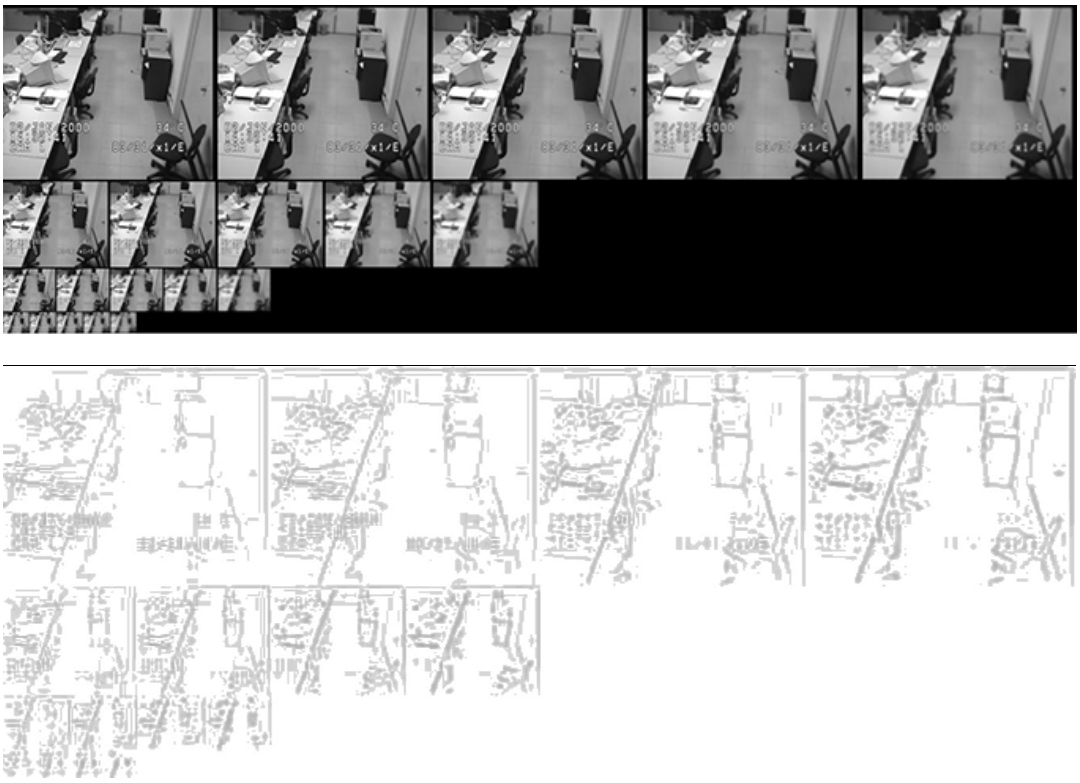
and therefore,

$$G(x,y,k\sigma) - G(x,y,\sigma) \approx (k-1)\sigma^2\nabla^2 G \qquad \text{[Eq. 2.8]}$$

The factor (k-1) in the equation is a constant over all scales. Therefore, it does not influence the extrema location. Besides, the approximation error will reach to zero while k comes to 1. According to Lowe, D.G. (2004) $k = \sqrt{2}$ has the less impact on the stability of extrema detection or localization for even signification differences in scale. An efficient approach to the construction of $D(x,y,\sigma)$ is shown in Figure 2.8.



(a) Different-of-Gaussian.



(b) Examples of DoG to find out edges of the image.

Figure 2.8: Image is convolved with DoG. For each octave, the Gaussian image is down-sampled by a factor of 2, and the process repeated. (Lowe, D.G., 2004)

In order to detect the minima and maxima of $D(x, y, \sigma)$, each sample point is compared to its eight neighbors in the current image and nine neighbors in the scale above and below as shown in Figure 2.9. It is selected if it is larger than all of these neighbors or smaller than all of them. The cost of this check is low due to the fact that most sample points will be eliminated following the first few checks.

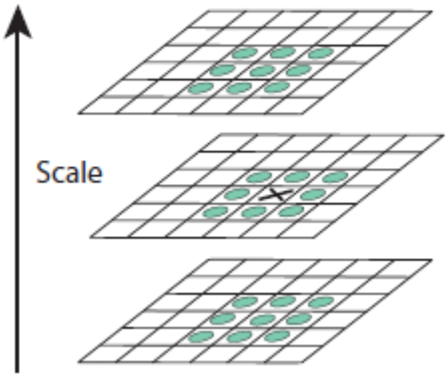
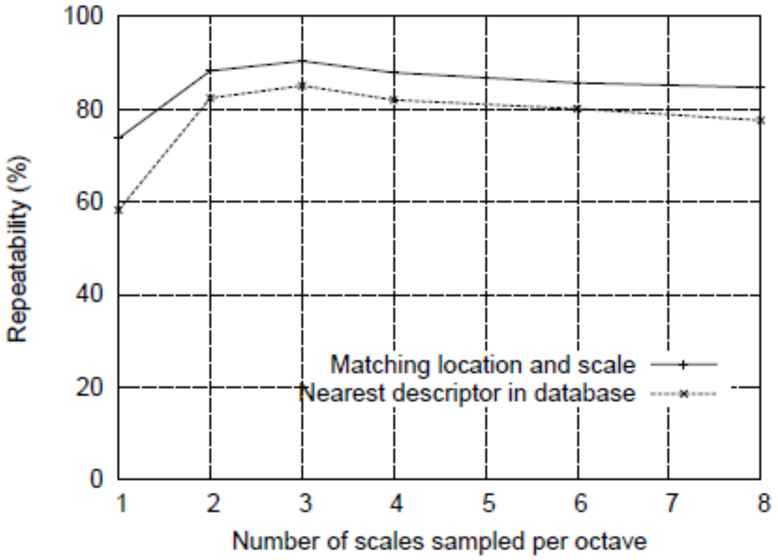
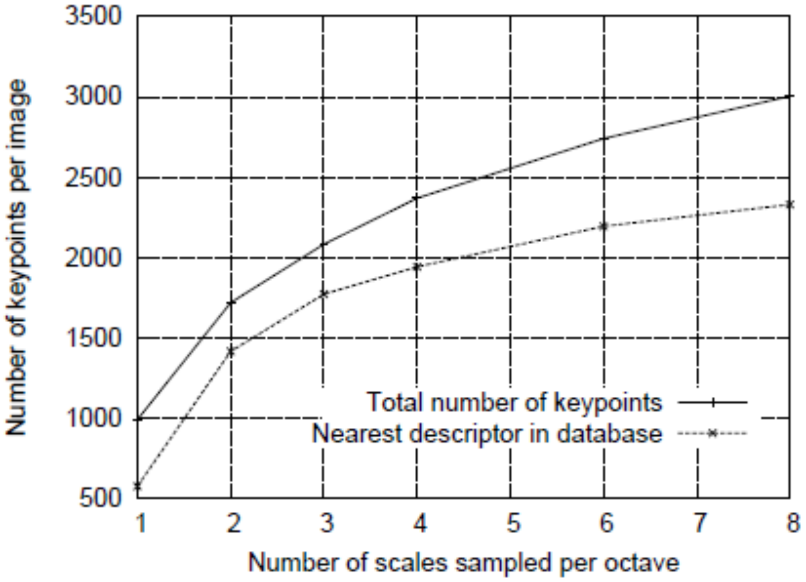


Figure 2.9: Detection of maxima and minima by comparing the pixel (x) to its 26 neighbors in 3x3 regions (o). (D.G. Lowe, 2004)

One of the significant issue is to determine the frequency of sampling in the image and scale domains that are required to dependablydiscover the extrema. Unfortunately, it turns out that there is no minimum spacing of samples that will detect all extrema, as the extrema can be arbitrarily close together. This can be seen by considering a white circle on a black background, which will have a single scale space maximum where the circular positive central region of the difference-of-Gaussian function matches the size and location of the circle. For a very elongated ellipse, there will be two maxima near each end of the ellipse. As the locations of maxima are a continuous function of the image, for some ellipse with intermediate elongation there will be a transition from a single maximum to two, with the maxima arbitrarily close to each other near the transition.



(a)



(b)

Figure 2.10: Experimental determination of sampling frequency that maximizes extrema stability. (Lowe D.G., 2004)

In Figure 2.10(a), the top line of the first graph shows the percentage of key points that are repeatedly detected at the same location and scale in a transformed image as a function of the number of scales sampled per octave.

The lower line shows the percentage of key points that have their descriptors correctly matched to a large database. Figure 2.10 (b) shows the total number of key points detected in a typical image as a function of the number of scale samples.

2.4.2 Key Point Localization

The candidates of maxima and minima are defined as key points. The next step is to perform a detailed fit to the nearby data for location, scale and ratio of principal curvatures. This information allows points that have low contrast or are poorly localized along an edge to be rejected. In the implementation of M. Brown and Lowe, D.G. (2002), Taylor expansion of the scale-space function, $D(x, y, \sigma)$ shifted such that the origin is at the sample point:

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{a x^2} x \qquad \text{[Eq. 2.9]}$$

where *D* and its derivatives are evaluated at the sample point and x $=$ $(x, y, \sigma)^T$ is the offset from this point. The location of the extremas, $\hat{x}$ is determined by taking the derivative of this function with respect to x and setting it to zero,

$$\hat{x} = -\frac{\partial^2 D}{\partial x^2}^{-1} \frac{\partial D}{\partial x} \qquad \text{[Eq. 2.10]}$$

The derivative of *D* is approximated by the differences of neighboring sample points.

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\,\hat{x} \qquad \text{[Eq. 2.11]}$$

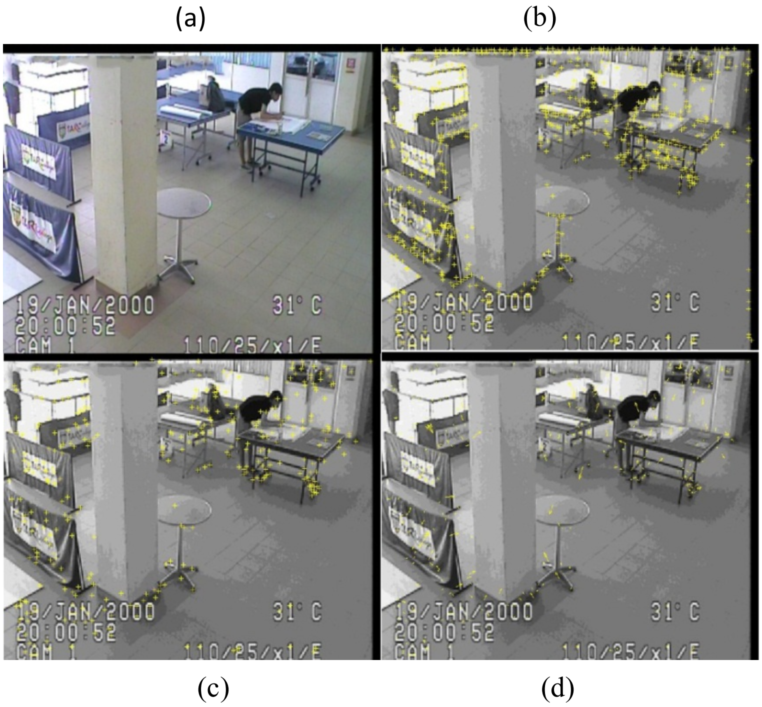


Figure 2.11: Process of keypoints localization.

Figure 2.11(a) shows the original image while Figure 2.11(b) shows the potential interest points after scale-space extrema detection. Figure 2.11(c) shows the result after low contrast filtering. Figure 2.11(d) shows the final result after removing edge points using principal curvature filtering.

In order to reject unstable extrema, difference-of-Gaussian function will have to generate strong response along edges addressing noise along the edge. A poorly defined peak in the difference-of-Gaussian function will have a large principal curvature across the edge but a small one in the perpendicular direction. The principle curvatures can be computed from a Hessian matrix by Harris and Stephens (1988), computed at the location and scale of the key points. The derivatives are estimated by taking differences of neighboring sample points. The transition from Figure 2.11(c) to (d) shows the effects of this operation.

2.4.3 Orientation Assignment

Based on local image properties, each key point may be assigned a consistent orientation so that the key point descriptor can be defined relative to the orientation and is invariant to image rotation. Each image property is based on rotationally invariant measure. One of the known disadvantages of this approach is that it limits the descriptors that can be used to discard image information by not requiring all measures to be based on a consistent rotation. According to Lowe, D.G. (2004), the scale of the key point is used to select the Gaussian smoothed image, L, with the closest scale, so that all computations are performed in a scale-invariant manner. For each image sample, $L = (x, y)$ at this scale, the magnitude, $m(x, y)$, and orientation, $\theta(x, y)$ is pre-computed using pixel differences:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \qquad \text{[Eq. 2.12]}$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \qquad \text{[Eq. 2.13]}$$

Gradient orientations of sample points within a region around the key point forms an orientation histogram. The orientation histogram consists of 36 bins to cover the 360-degree range of orientations. Samples added to the histogram will be weighted by its gradient magnitude and a Gaussian-weighted circular window with an σ which is 1.5 times of the scale of the key point.

Peaks in the histogram will be matched to determine directions of local gradients. a key point with that orientation will be created by detecting the highest peak in the histogram, and any other local peak that is within 80% of the highest peak. For that reason, locations with multiple peaks of similar magnitude, multiple key points will be created at the same location and scale, but with different orientations. Therefore, for locations with multiple peaks of similar magnitude, there will be multiple key points created at the same location and scale but different orientations. Only about 15% of points are assigned multiple orientations, but these contribute significantly to the stability of matching. Finally, a parabola is fitted to the 3 histogram values closest to each peak to interpolate the peak position for better accuracy. Figure 2.12 shows the experimental stability of location, scale, and orientation assignment under differrent amounts of image noise.

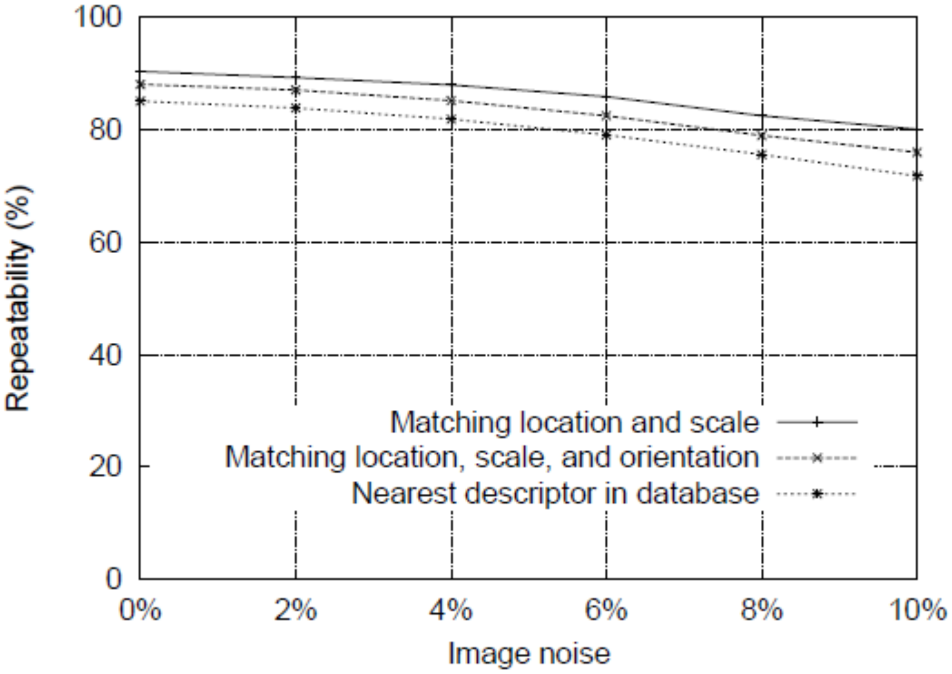


Figure 2.12: Stability of location under differing amounts of image noise. (Lowe, D.G., 2004)

2.4.4 Key Point Descriptor

The previous operations have assigned an image location, scale, and orientation to each key point. These parameters impose a repeatable local 2D coordinate system in which to describe the local image region, and therefore provide invariance to these parameters. The next step is to compute a descriptor for the local image region that is highly distinctive yet is as invariant as possible to remaining variations, such as change in illumination or 3D viewpoint. The local image intensities around the key point will match using a normalize correlation measure. However, the simple correlation of image is highly sensitive to changes that cause mis-registration of samples, such as affine or 3D viewpoint change or non-rigid deformations. Figure 2.13 illustrates the computation of the key point descriptor.

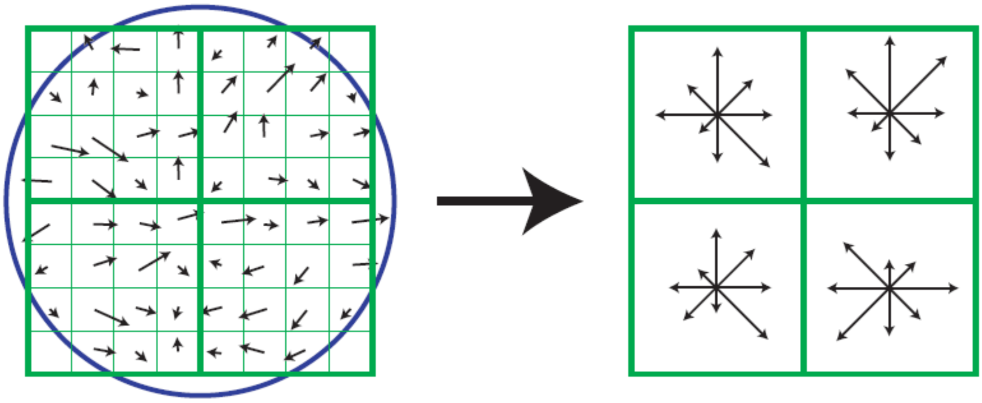


Figure 2.13: Gradient magnitude and orientation at each image sample point in a region around the keypoint location. (Lowe D.G., 2004)

In Figure 2.13 the image gradient magnitudes and orientations are sampled in a region around the key point location, as shown on the left. These are weighed by using the scales of the key point to select the level of Gaussian blur for the image in the overlaid circle. In order to achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the key point orientation. Each sample location is marked with small arrows. The descriptors are formed from a vector containing the values of all the orientation histogram entries, corresponding to the lengths of the arrows on the right side of Figure 2.13. The figure shows a 2x2 array of orientation histograms, but a 4x4 array of histograms with 8 orientation bins in each can achieve the best result. Therefore, experiments in this project utilised a 128 element feature vector (4x4x8) for each key point in the dataset.

Lastly, to reduce the effects of illumination change, the feature vector is normalised to unit length. Vector normalisation will cancel the change in image contrast in which each pixel value is multiplied by a constant will multiply gradients by the same constant..

2.5 Hough Transformation

A typical image contains thousands or more features which may come from foreground and background clutter. To improve the performance of object matching, many well-known robust fitting methods, such as RANSAC or Least median of Squares, perform poorly when the percentage of inliers is below 50%. However, "Hough transform (Hough, 1962) achieves better performance by clustering features in pose space".

Hough Transform uses a consistent interpretation of each feature in voting for all object poses that are consistent with the feature to define the feature clustering. The corrected probability of interpretation will be higher than for any single feature if any of the clusters of features were found to vote the same object's pose. Each key point specifies 4 parameters: 2D location, scale and orientation and each matched key point in the database has a record of the key point's parameters relative to the training image in which it was found. Therefore, Hough transform creates prediction on the model location, orientation and scale from the match hypothesis. In the implementation of the Hough transform, a multi-dimensional array is used to represent the bins. However, many of the potential bins remain empty, and it is difficult to compute the range of possible bin values due to their mutual dependence. These problems can be avoided by using a pseudo-random hash function of the bin values to insert votes into a one-dimensional hash table, in which collisions are easily detected.

The Hough transform is used to identify all clusters with at least three entries in a bin. Each cluster is then subjected to a geometric verification procedure in which a least-squares solution is performed for the best affine projection parameters relating the training image to the new image (Lowe, D.G., 2004). When high number of votes fall in the right bin, the Hough transform will be efficient, the bin can easily detected among the background noise. The bin must not be too small to be detected, or it will reduce the visibility of the main bin when some votes fall in the neighbouring bin. The Hough transform must be used with attention to detect anything other than lines or circles, when number of parameters is large, the average number of votes cast in a single bin will be low, and those bins matching to a real figure in the image might not appear to have a higher number of votes than their neighbours.

In a nutshell, the quality of the input data heavily influences the efficiency of the Hough Transform. In order for the Hough Transform to be efficient, edges of the images must be detected well. Using noisy image on Hough Transform needs a great deal of attention, and normally, image must go thru a de-noising stage before used in Hough Transform.

2.6 Robust Homography Estimation using RANSAC

RANSAC (Random Sample Consenses) is a general parameter estimation approach and also a resampling techniquethat generates candidate solutions which designed to deal with a large scale of outliers in the input data. It required mininum number of data points to estimate the underlying model parameters. Meanwhile, the common rubust estimation techniquesrequired much of the data as possible to obtain an initial solution to prune outliers. For instance, techniquesM-estimators and least-median squares. RANSAC uses a minimal set of randomly sampled correspondences to estimate image transformation parameters, and finds a solution that has the best consensus with the data.

For each pair of potentially matching images, there is a set of feature matches that are geometrically consistent (RANSAC inliers) and a set of features that are inside the area of overlap but not consistent (RANSAC outliers). The idea of this verification is to compare the probabilities of inliers and outliers was generated correctly.

The number of iterations, *N*, is chosen high enough to ensure that the probability p (usually set to 0.99) that at least one of the sets of random samples does not include an outlier. Let u represent the probability that any selected data point is an inlier and $v = 1 - u$ the probability of observing an outlier. *N* iterations of the minimum number of point's denoted m are required, where

$$1 - p = (1 - u^m)^N \qquad \text{[Eq. 2.14]}$$

And thus with some manipulation,

$$N = \frac{\log(1-p)}{\log(1-(1-v)^m)} \qquad \text{[Eq. 2.15]}$$

An advantage of RANSAC is to perform robust estimation of the model parameter.For example, when a remarkable outliers are detected in the data set, it has high accuracy to estimate the parameters. In the other hand, the disadvantage of RANSAC is that there is no limit on the time it takes to compute these parameters. The results may not be optimum when the number of iterations computed is limited, and it may not even be the one that fits the data in a good way. Another disadvantage is that it requires the setting of problem-specific thresholds which is a common disadvantage in most of the current image processing solutions. Only one model for a particular data set can be estimated by RANSAC. So, it may fail to find the model when there are two ore more model instances exist. However, the Hough Transform is an alternative robust estimation technique which is useful when more than one model instance is present in the data set.

2.7 Affine Transformation

Hough transform had been selected to perform robust estimation in dataset. Before proceeding to image stitching process, affine transformation is an important class of linear 2-D geometric transformation which maps variables, for example “pixel intensity values located at position (*x,y*) in an input image into new variables in an output image by applying a linear combination of translation, rotation, scaling and/or shearing operations (R.Fisher et al., 2000)”. Perspective irregularities introduces geometric distortion that subjects for image acquisition where in the PTZ camera position with respect to the scene that alters the apparent dimensions of the scene geometry. An uniformly distorted image can be corrected by applying an affine transformation for a range of perspective distortion with the measurements transformation from the outstanding coordinates to those which in used actually.

An affine transformation is any transformation that preserves collinearity and ratios of distances. In this sense, affine indicates a special class of projective transformations that do not move any objects from the affine space to the plane at infinity or conversely. An affine transformation is also called an affinity. The problem of perspective can be overcome if we construct a shape description which is invariant to perspective projection. Many interesting tasks within model based computer vision can be accomplished without recourse to Euclidean shape descriptions and, employ descriptions involving relative measurements. For instance, those which rely only upon the configuration's intrinsic geometric relations. From the images, these relative measurements can be determined directly. Figure 2.14 shows a hierarchy of planar transformations which are important to computer vision.

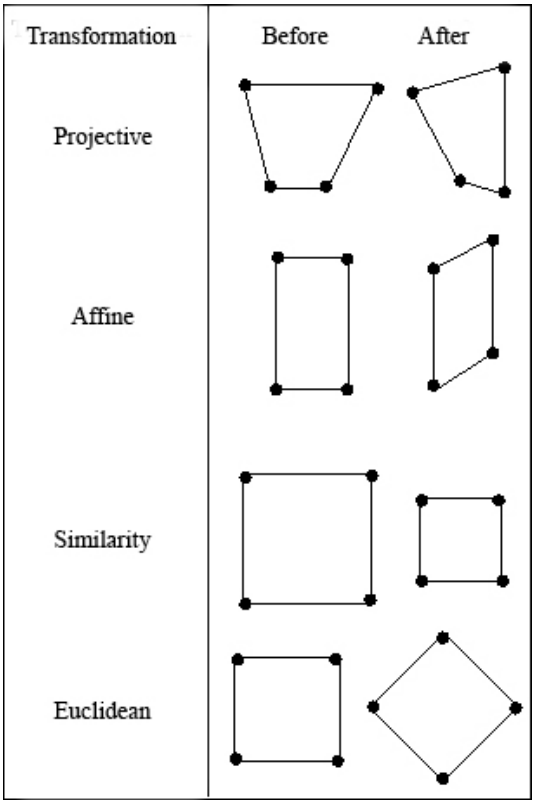


Figure 2.14: Hierarchy of plane to plane transformation from Euclidean to Projective.

Under orthographic projection,an affine transformation correctly accounts for 3D rotation of a planar surface, but the approximation for 3D rotation of non-planar objects is well as planar objects. However, a fundamental matrix solution requires at least 7 points matches as compared to only 3 for the affine solution and in practice requires even more match for good stability. A more general approach in Brown, M. and Lowe, D.G. (2002), initial solution is based on a similarity transform, which then progress to solution for the fundamental matrix in those cases in which a sufficient number of matches are found.

The affine transformation of a model point $[x\ y]^T$to an image point $[u\ v]^T$ can be written as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} m1 & m2 \\ m3 & m4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \qquad \text{[Eq. 2.16]}$$

where the model translation is $\left[t_x t_y\right]^T$. Then, the affine rotation, stretch, and scale are represented by the m_i parameters. As a result, the equation above can be rewritten to gather the unknowns into a column vector:

$$\begin{bmatrix} x & y & 0 & 0 & 1 & 0 \\ 0 & 0 & x & y & 0 & 1 \\ & & \dots & \dots & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u \\ v \\ \dots \end{bmatrix} \qquad \text{[Eq. 2.17]}$$

Any number of further matches can be added although this equation shows a single match. Each match is contributing two more to the first and last matrix. Yet, at least three matches are needed to provide a solution.

2.8 Geometry Transformation based on ImTrans (MATLAB)

ImTran (Appendix A) applies a geometric transform to an image with 3x3 homogeneous transformation matrixes which is affine transformation values. The region of interest in panoramic may invert a perspective transformation of a plane and vanishing line of the plane lies within the image. Attempts to transform any part of vanishing line will position at infinity. Therefore, we should specify a region that excludes any part of vanishing line.

CHAPTER 3

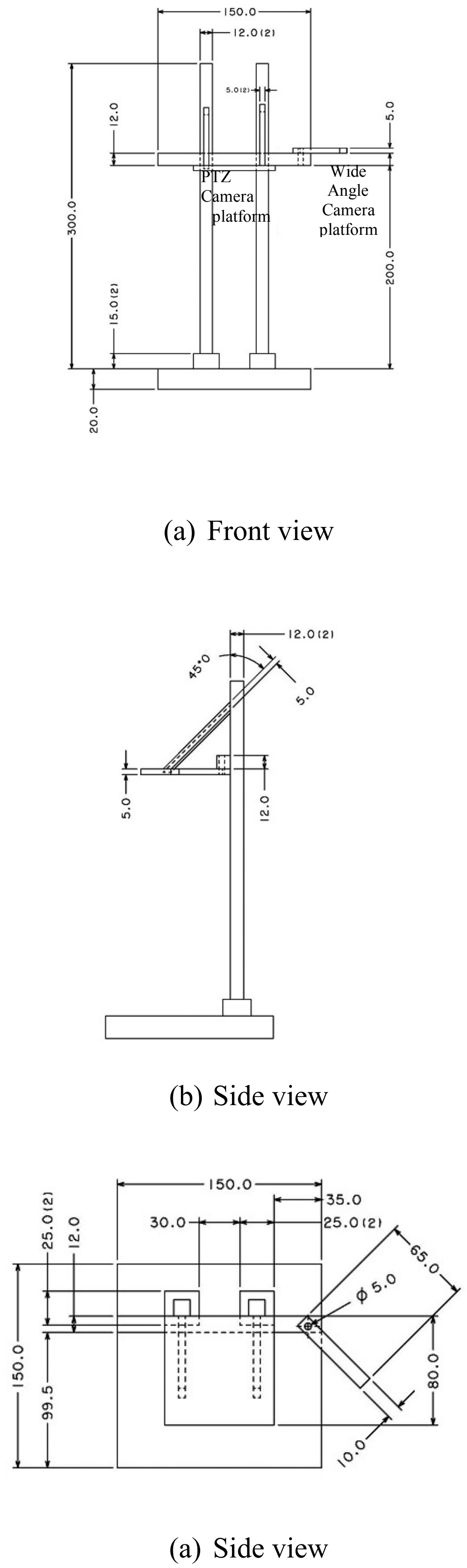
HYBRID CAMERA SYSTEM CONCEPT AND DESIGN

3.1 Introduction

This chapter will cover how the hybrid camera system is being set up in this project. A testbed was being setup to collect the dataset by using both cameras, wide angle camera and PTZ camera, an aluminum rack was constructed with two mounting brackets. Both cameras were mounted side by side with a distance *x*, where $0.1 < x < 1.0$ meter. Greater value distance, *x* between both cameras may increase the disparity of the images. The purpose of this rack is to allow us to run experiment in different environment (e.g. indoors and outdoors) with ease of mobility.

3.2 Equipment Setup

The camera rack is made of aluminum as shown in Figure 3.1 with 3.0 meters height and 2.0 meters width. PTZ camera and wide angle camera are mounted at around 2.5 meters height. With this setup, it creates a good position to acquire information of the environment for object detection and object recognition process for features enhancement.



(a) Front view

(b) Side view

(a) Side view

Figure 3.1: Layout of camera rack design.

3.3 Pan-Tilt-Zoom Camera

Pan-tilt-zoom camera is a typical and the simplest active camera which can be fully controlled by specifying pan, tilt and zoom parameters. Additionally, PTZ cameras can rotate 360 degrees spinelessly and view an object directly below the camera. It minimizes the area of blind spots and fully covered the view of environment. "PTZ cameras are able to obtain multi-view-angle and multi-resolution information (Wan D. and Zhou, J., 2008)". Typically, a camera's motion is remotely controlled with a keyboard and joystick. Users are allowed to pan, tilt and zoom into a specific area with a push of a button. Recent years, they have been recognized and received more and more attention in the research and development of surveillance system. They are more than one camera to monitor the environment. Supervised and unsupervised calibrations of camera networks are needed.

In computer vision research, PTZ cameras create more possibilities for improvement in technology compared with static camera. One instinct advantage PTZ cameras have is the ability to zoom. It can zoom in on any objects. Most common research topics are automation monitoring based on motion tracking, behavioral analysis, human detection and recognition and etc.

At the same time, these cameras have become cheaper and already deployed in many real applications with features integrated, for instance: PTZ cameras are integrated with devices such as magnetic door contacts, and alarms systems.

When a certain device is triggered, the camera can move to view the predetermined specified location. The PTZ camera system used in this project has 10X optical zoom and 360 degrees spineless camera as shown in Figure 3.2.

Figure 3.2: PTZ Camera consists 10x optical zoom and 360º of pan and 90º of tilt.

3.4 Pelco-D Protocol

Pelco-D is the standard protocol that is widely used in CCTV industry. In this project, PTZ camera was remotely controlled by sending/receiving the Pelco-D messages. Pelco-D consists of 7 hexadecimal bytes. In this session, all byte data are in hexadecimal format. Table 3.1 shows the format of the message.

Table 3.1: Format protocol for message the PTZ camera.

Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6	Byte 7
Sync byte	Camera Address	Command 1	Command 2	Data 1	Data 2	Checksum

- Byte 1 (Sync) – The synchronization byte is always $FF.
- Byte 2 (Camera address) – The address is the logical address of the receiver/driver being controlled.
- Byte 3 & 4 (Command 1 & 2) – The command to send are shown in Table 3.2.
- Byte 5 (Data 1) – Pan Speed.
- Byte 6 (Data 2) – Tilt Speed.
- Byte 7 (Checksum) – The checksum.

Command 1 and 2 are as follows:

Table 3.2: 8 bits format in Command 1 and Command 2.

	Bit 7	Bit 6	Bit 5	Bit 4	Bit 3	Bit 2	Bit 1	Bit 0
Command 1	Sense	Reserved	Reserved	Auto/ Manual Scan	Camera On/ Off	Iris Close	Iris Open	Focus Near
Command 2	Focus Far	Zoom Wide	Zoom Tele	Down	Up	Left	Right	(Always 0)

The sense bit acts as an indicator to determine the value in bits 4 and 3 (Protocol Manual, 2011). When the sense bit is on, and bit 4 and 3 are on, the command will enable auto-scan and turn on the camera. However, bit 4 and 3 are on the command will enable manual scan and turn off the camera when the sense bit is off. If either bit 4 or 3 are off then no action will be taken for any features.

Bit 6 and 5 are reserved bit. They will always set to 0.

Furthermore, byte 5 controls the pan speed. The pan speed is in the range 00 (Stop) to 3F (Medium Speed) while FF is maximum speed. In this system, maximum setting is not recommended because it might be not a smooth step from an angle to another.

Byte 6 controls the tilt speed. The tilt speed is in the range from 00 (Stop) to 3F (Maximum Speed).

Byte 7 is the check sum for the command. It is the 8bit sum of the payload bytes in the message.

In addition, there are more control commands shown in appendix B. Users are allow to access more advanced features by customize the command. But, the device being queried can only be used in a point to point architecture. Else it will respond to any address. Therefore, you need multiple devices transmitting if there are more than one device listens to this command at the same time.

The response to one of these commands is seven bytes long, for an example: Set Preset command message.

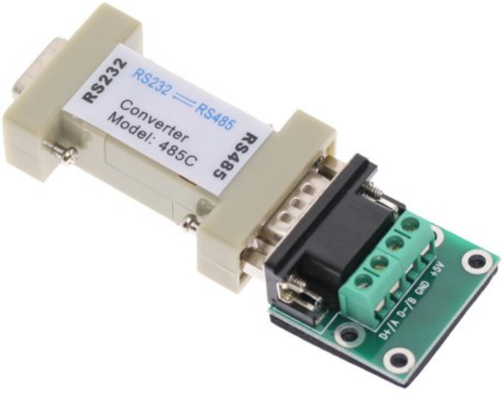
Table 3.3: Check sum is the summation of byte 2,3,4,5, and 6 in format hexadecimal.

Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6	Byte 7
FF	01	00	03	00	01	05

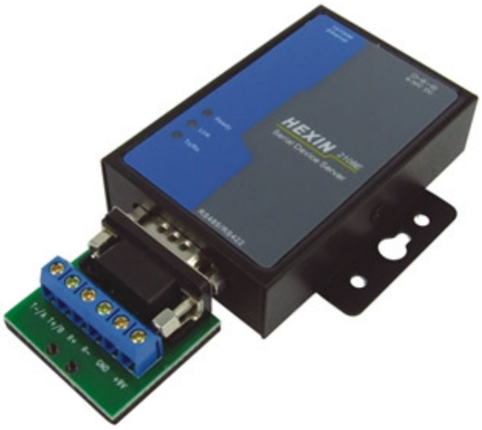
3.5 RS 485 Transmitter

In hybrid-camera system, controller and PTZ camera has to run on the same protocol and interface before they can “talk to each other”. Pelco-D protocol with RS485 interface on a single twisted pair cable is the most common way to interface with and control a PTZ camera.

Command protocol has to be transmitted from the system via an interface cable to the PTZ camera. RS 485 is a one way interface on a single twisted pair. The difference between RS 485 and RS 422 is that RS422 has two way interface on a double twisted pair. Another alternative is to use RS232 for very short distance. RS485 interface able to transmit the data over a few kilometers and is common used by most PTZ cameras. In this project development, PC based DVR had been used to communicate with PTZ camera through “com-port”, or serial port. In fact, an interface converter is needed to convert the serial RS232 port to RS 485 port. For short distance, the passive converter can be use since that it does not require external power source which shown in Figure 3.3(a). For long distance, up to a few kilometers, the converter should be powered up to maintain the signals, Figure 3.3(b).



(a) Without external power source needed.



(b) Require external power to maximum the safe distance for data transmission

Figure 3.3: Both devices ability to convert serial RS232 port to RS485 ports.

Computer sends command to the PTZ camera through RS485 cable pairs. Therefore, Tx+ and Tx- (Transmit + and -) are being used on the computer side. Then, the twisted pair to Rx+ and Rx- (Receive + and -) is connecting on the PTZ side.

Procedure to connecting a PTZ camera and static camera on hybrid-camera system:

i. Firstly, the PC is installed with a DVR card. Then, a coaxial cable is used to connect the PC to cameras for video acquisition. Also, a twisted pair is used for the RS485 command interface.

ii. Appropriate protocol need to be set for PTZ camera command interface, for examples baud rate is 2400bps for pelco 'D' and camera ID, eg. 1 for static camera, 2 for PTZ camera,etc.

iii. Matched up the setting with the setting those are configured in the camera. At this point, the hybrid-camera system is able to control the PTZ camera and also to acquire the images.

CHAPTER 4

PANORAMIC IMAGE STITCHING BASED ON SIFT

4.1 Introduction

In this chapter, the objective is to implement the SIFT algorithm to perform panoramic image stitching. The SIFT has several advantages over several other approaches such as PCA-SIFT by Ke, Y. and Sukthankar, R. (2004) and SURF by H. Bay et al. (2006). In the previous work, such features matching method based on Harris corner is lack of invariant properties to increase the reliability of image matching and stitching. Firstly, SIFT performs well while matching panoramic image sequence despite rotation, zoom and illumination change in the input images. Secondly, the matching relationships between images may be discovered and recognized from datasets. Thirdly, all the connected sets of images can be stitched to form a panoramic image. However, the main objective to match images is to identify overlapping portions of images in order to get a good solution for the image geometry.

In hybrid-camera system, all cameras have overlapping fields of view and thus share geometric information. In this case, static camera's views are defined as model images while PTZ camera can serve to capture high resolution images from the viewpoints. Figure 4.1 shows the basic geometry of the hybrid-camera system and the relationship of the cameras. Instead of capturing an image from static camera, PTZ camera will capture the image of the scene with different values of pan, tilt, and zoom to produce more details of the environment.

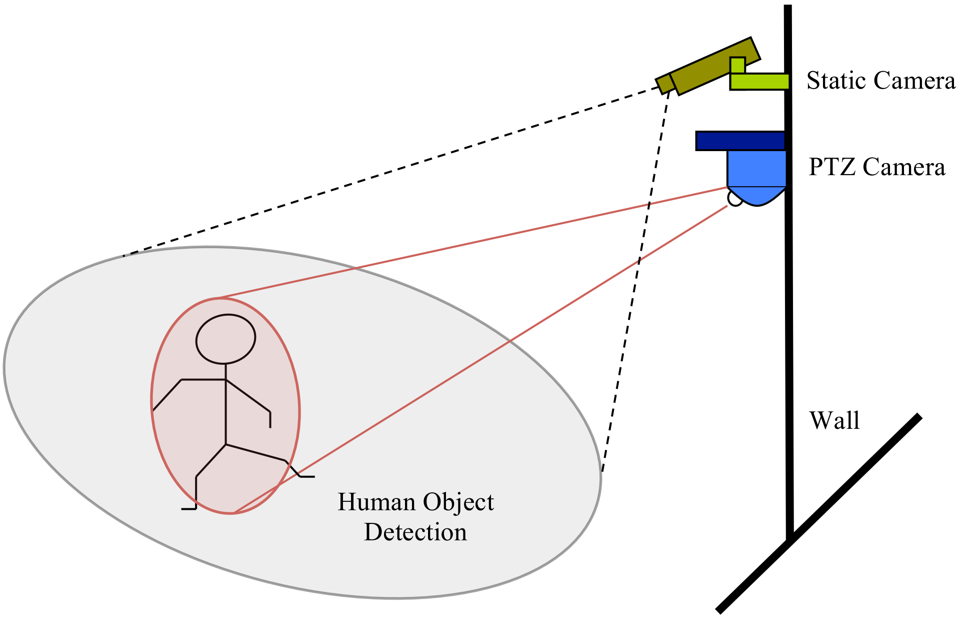


Figure 4.1: Geometry relationship of Hybrid cameras system in testbed.

Based on Figure 4.1, the relationship between the static camera and PTZ camera are defined as

- Image PTZ camera, I_{PTZ}
- Image static camera, I_{static}

$$I_{PTZ} \subset I_{static} \qquad \text{[Eq. 4.1]}$$

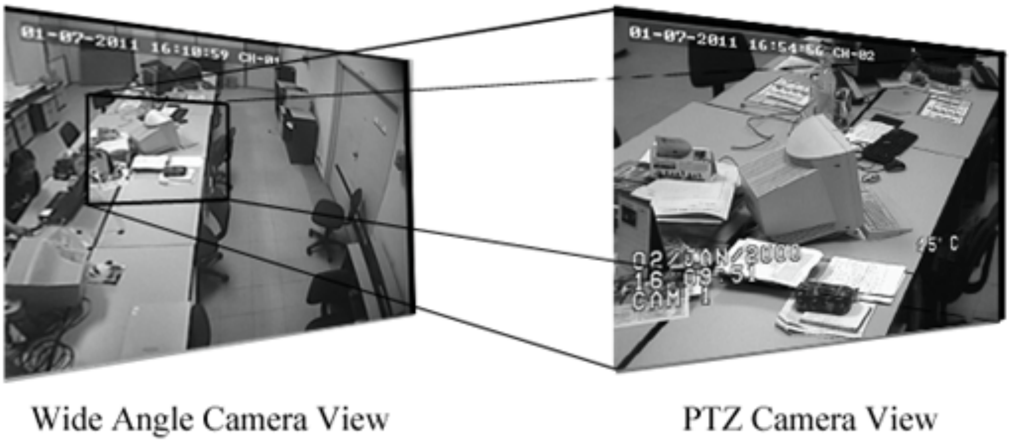


Figure 4.2: Sample image to show that I_{PTZ} is subset of image, I_{static}.

Image, I_{PTZ} is subset of image, I_{static}. In hybrid camera system, there are more than one images, I_{PTZ} will be acquired by PTZ camera to be compared with static camera. Each $I_{PTZ(n)}$, *n* = 1, 2, 3… , has intersections with others.

$$I_{PTZ(n)} \cap I_{PTZ(n+1)} \text{ , } n = 1, 2, 3.... \qquad \text{[Eq. 4.2]}$$

The union of images, $I_{PTZ(n)}$ become image panoramic, $I_{panoramic}$.

$$\left[\sum_n I_{PTZ(n)} \cup I_{PTZ(n+1)}\right] = I_{panoramic} \qquad \text{[Eq. 4.3]}$$

$I_{panoramic}$, provides wider of view and more information of the environment. The following assumption is made

$$I_{static} \subset I_{panoramic} \qquad \text{[Eq. 4.4]}$$

4.2 PTZ camera Geometry

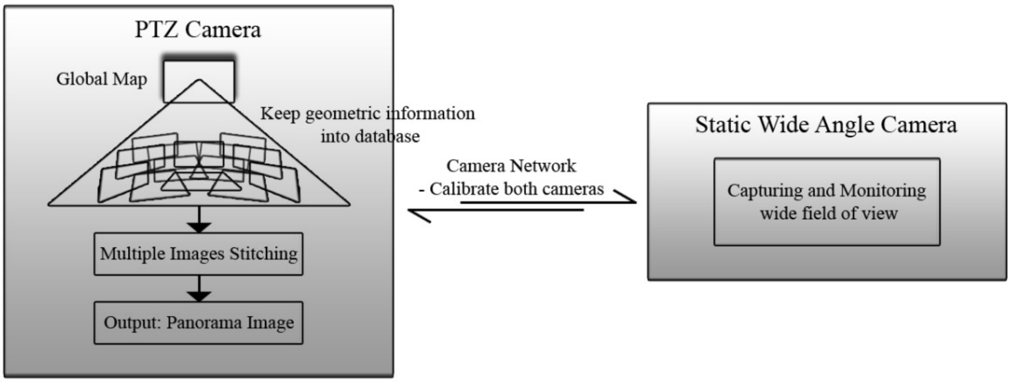

PTZ Camera
Global Map
Keep geometric information into database
Multiple Images Stitching
Output: Panorama Image
Camera Network
- Calibrate both cameras
Static Wide Angle Camera
Capturing and Monitoring wide field of view

Figure 4.3: Architectural and concept overviews of hybrid camera system.

When cameras are used in large and wide environment, the deviation of the image nodal point is negligible compared to narrow environment. Based on this, we fixed the PTZ camera routine and assume that the camera rotates within the coverage of static camera image. In this process, un-calibrated images are acquired; geometric information of each image is kept in a database. In order to have an overview of those images, we propose the similar approach as Gonzalez R.C. and Woods, R.E. (2007), which is to build a panoramic image of the scene from PTZ camera views before we performing image calibration between both cameras. The scene is decoupled from frame-to-frame positioning of camera and focal length so that panoramic images are created at different set of zooms.

In order to estimate the intensity mapping between images, we adopted the method Scale Invariant Feature Transform (SIFT) features as the first layer to extract the key points from each PTZ's image. We implemented the image stitching method as proposed by Chen I.H and Wang S.J. (2007) to stitch them to become a panorama image. Since a panorama image has a much wider field of view, we proceed to a SIFT-matching process to estimate the master camera pose of the images composing the panoramas. The work flows and architecture are shown in Figure 4.4.

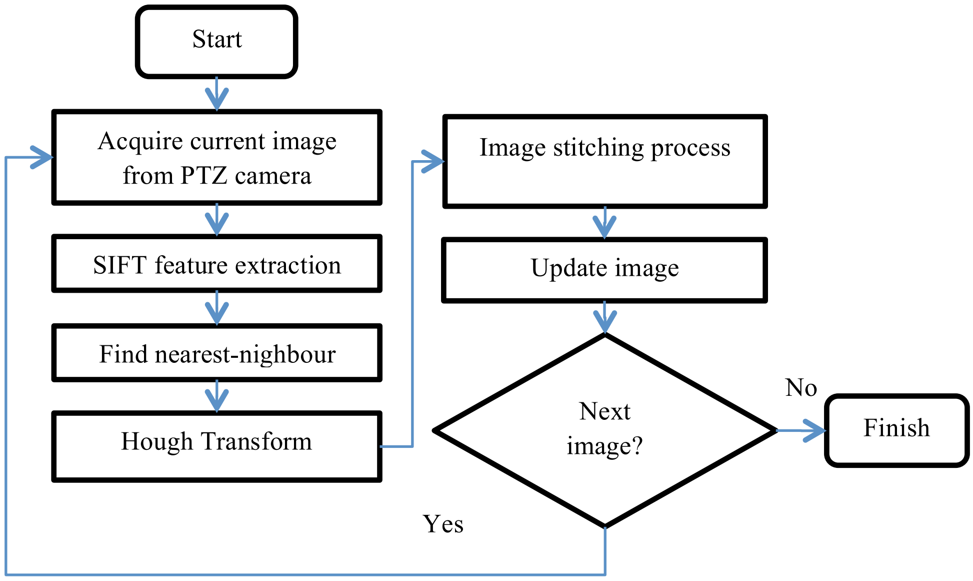


Figure 4.4: Work flow of multiple images stitching

In the stitching process, we know there are overlapping regions between an image with another. So, one of the images will be selected as the reference image. Then, the images acquired from the PTZ camera can be mapped to this image according to the homorgraphy or corresponding key points between the reference image and other images. Therefore, the panorama can be treated as a very large single image. Then images information is stored in a database. Due to the ability limitation of the matching technique, panoramas are generated at different zoom parameters to cover a large range of scale change. For example figure 4.5 and figure 4.6 shows the data collection of images of different zoom parameters. We set a limit and do not create panoramas for very large zoom although PTZ camera can achieve ten times optical zooms. This procedure is because the number of images needed to cover the whole homology is too large and thus making the process slow. The robustness of the image will cause failure in matching while calibrating the master and slave cameras because of the texture and scale of the object might have a large different index. To ensure the calibration process smoothly, the image was acquired according to a fixed direction lookup table.

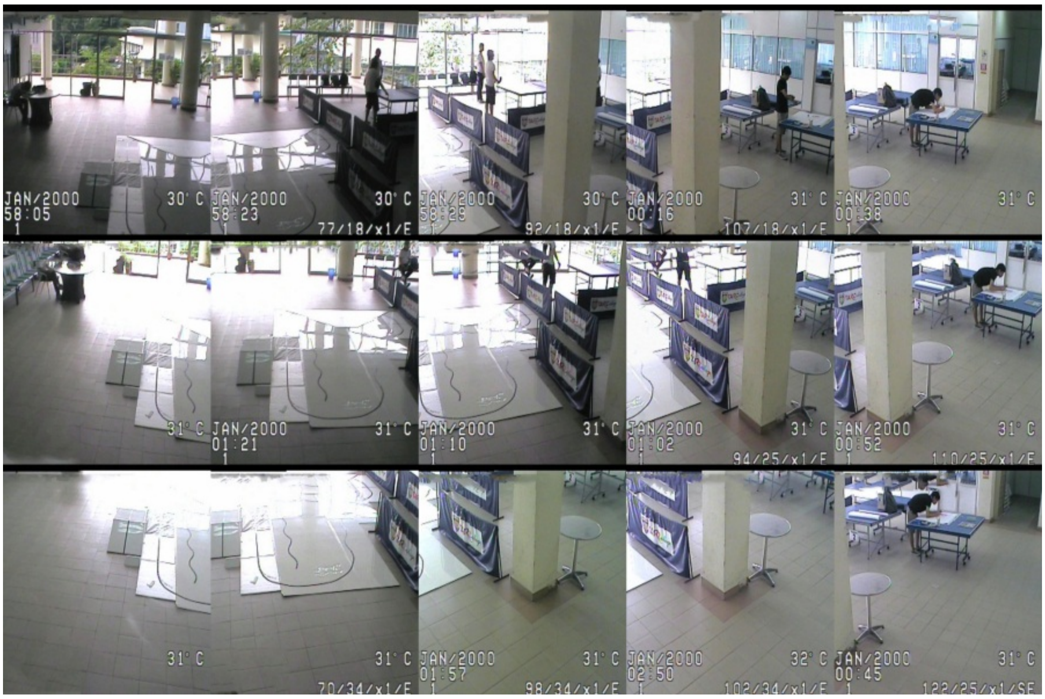


Figure 4.5: Data collection under different light illumination.

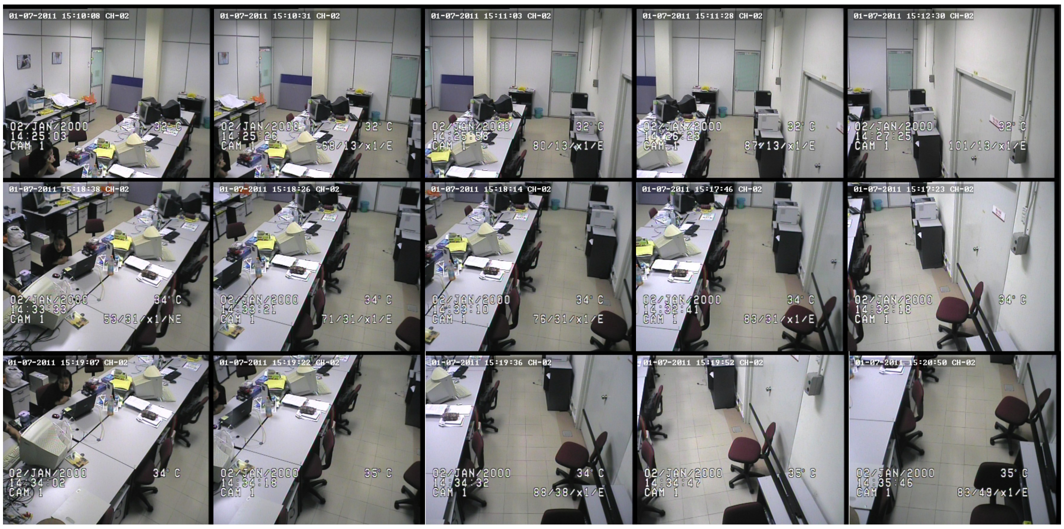


Figure 4.6: Data collection under constant light illumination.

4.3 Images Matching

As explained earlier in Section 2.4,SIFT algorithmcan identify and generate large numbers of keypoints in an image. However, many features of an image will not have any correct match in the database of training images due to the cluttered background. Therefore, the best matching candidate for each keypoint is found by identifying its nearest neighbor in the database. The nearest neighbor is determined as the keypoint with the minimum Euclidean distance for the invariant descriptor vector. A global threshold on the distance has to be set to discard the keypoints which do not have good match to the database. By comparing the distance of the closest neighbor to that of the second-closest neighbor, it able to obtain more effective measurement results. This measure performs well because correct matches need to achieve reliable matching. For a false match, there are likely a number of other false matches withquite similar distancesdue to the high dimensionalityof the feature space.

According to Lowe D.G. (2004), at least three nearest features identified as reliable key points. A typical image contains thousands of features which may come from different objects as well as background clutter. The nearest-neighbor process discards false matches arising from background clutter but does not identify matches from other valid objects and thus further matching process is needed to identify correct subsets of matches. In this project, well-known robust fitting methods such as RANSAC and Hough transform have been implemented in panoramic image stitching process.

For image matching and recognition, a new image is compared with the database based on their feature vectors. In the Nearest Neighbour process, a search on the database for a feature in model image, F_1^i which corresponds to an image feature, F_2^i is carried out. The correspondences are termed as key points and each key point consists of the smallest Euclidean distance between feature F_1^i and F_2^i or match M (F_1^i,F_2^i). This can be performed rapidly to identify the correct match key point descriptors with good proximity in the database of features. However, in a cluttered image, many features may affect the accuracy of correct matches and results in false matches of the key points.

4.4 Experiments and Results

The experiment was conducted to show how the different effects, such as illumination or texture on an image affect the image stitching process. The experiment has been evaluated on the testbed setup in two different kinds of environment. The first location was in the hallway while the second location is in the laboratory. The testbed was placed in proximity 2.5 meters height in both environment.

4.4.1 Experiment I – SIFT algorithm detection

In this experiment, three different kinds of images was selected to test through the capability of SIFT detection. These images are classified in three catories.

a. High texture with high contrast as shown in figure 4.7.
b. Low texture with high contrast as shown in figure 4.8.
c. Medium texture with consistent illumination as shown in figure 4.9.

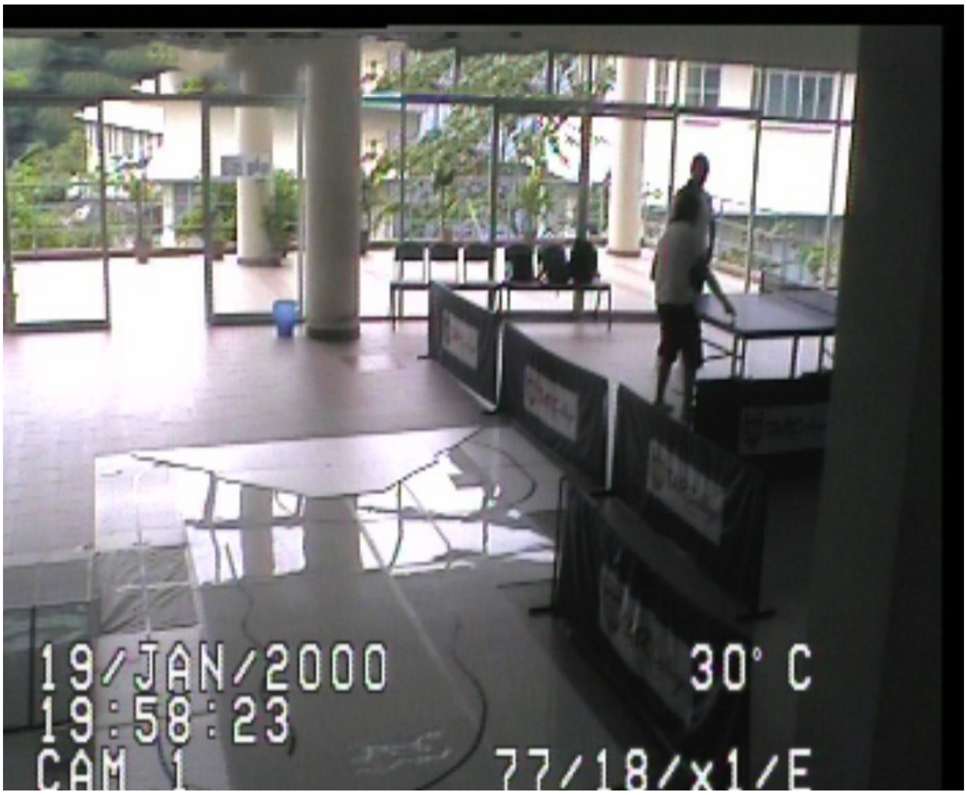


Figure 4.7: High texture with high contrast.

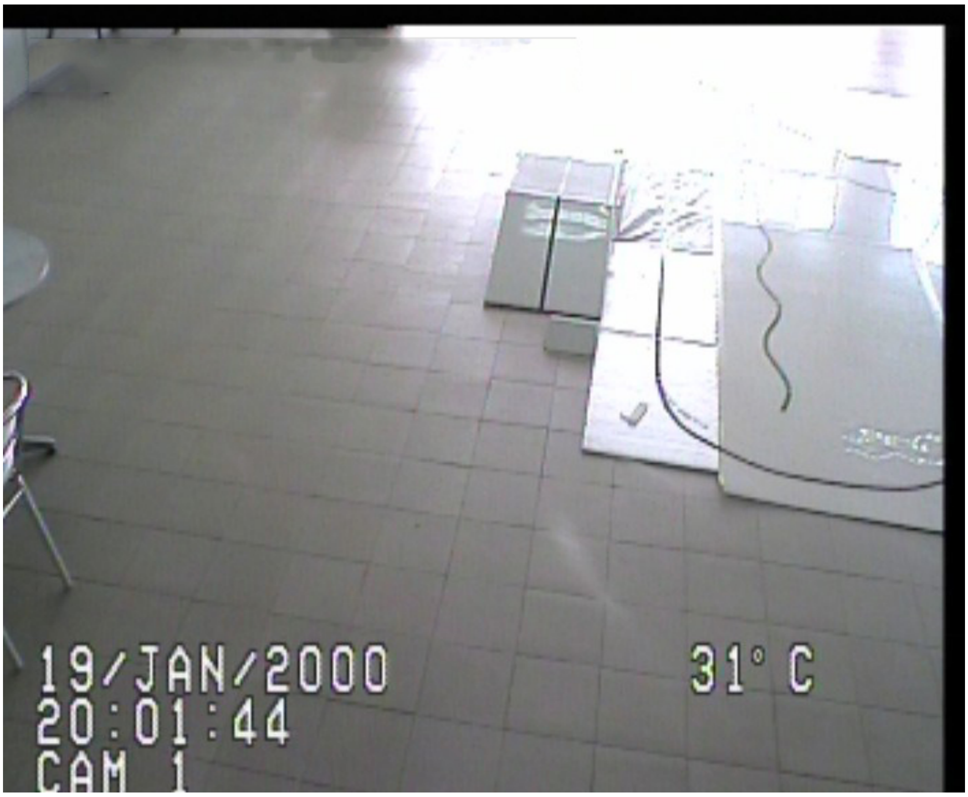


Figure 4.8: Low texture with high contrast.

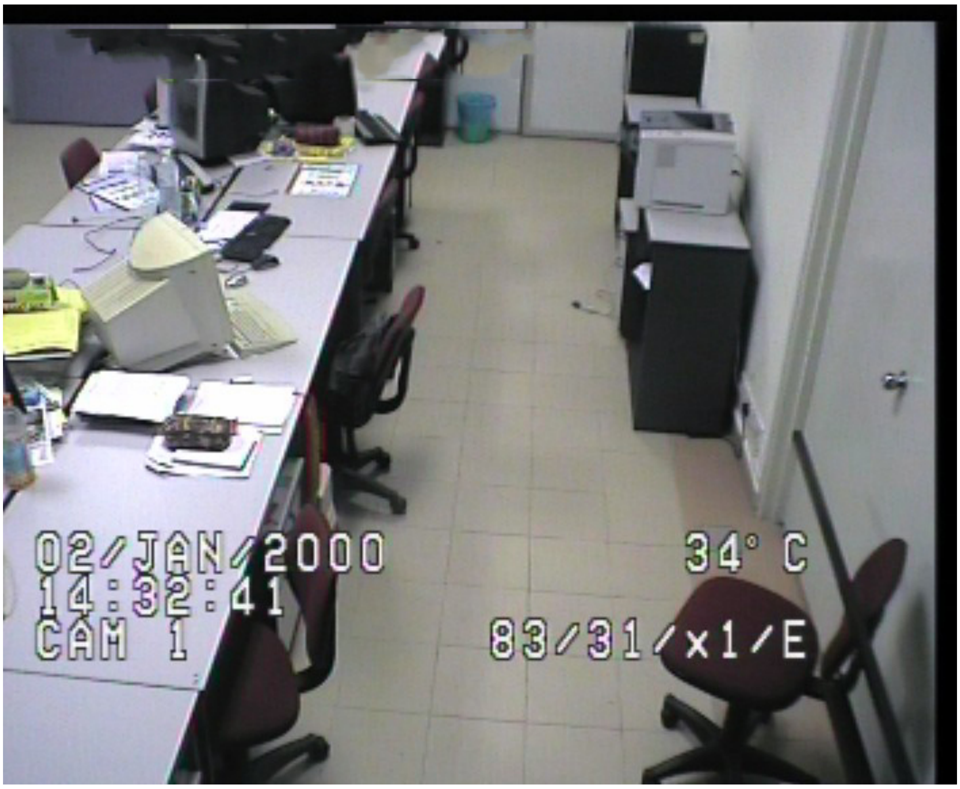


Figure 4.9: Medium texture with consistent contrast.

These images were subjected to SIFT algorithm to identify the correspondence key points among them. The original resolution of the images are 704x576 each. The resolution of the images shrinked 10 percent in each experiment while the smallest resolution is 70x58. These images are shown in appendix C.

When the dataset followed through the SIFT algorithm, the key points in were identified in each images. Figure 4.10 shows the example of SIFT algorithm process, the total number of key points are detected by going through the process below:

a. detecting and locating raw key points,
b. eliminating low contrast key points,
c. eliminating edge key points,
d. optimizing the number of key points.

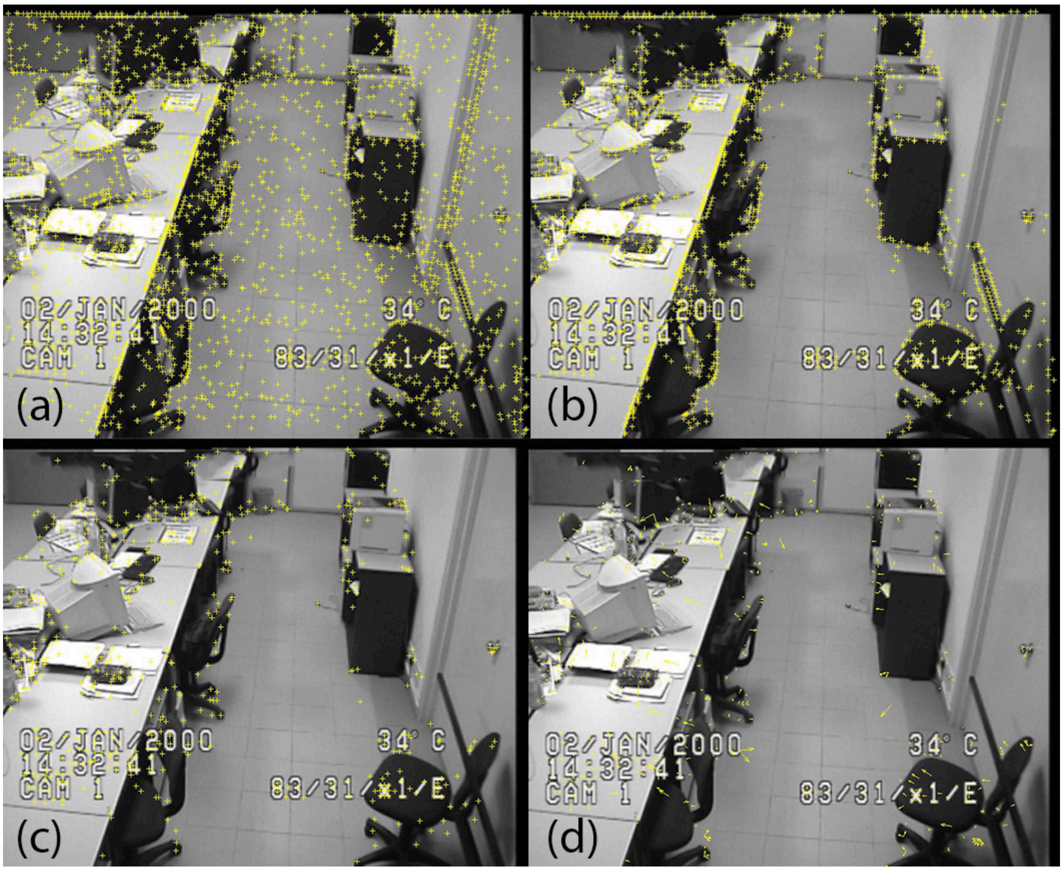


Figure 4.10: Example of SIFT algorithm process.

4.4.2 Results Experiment I(a)(b)(c)

In this experiment, the red colour lines shows the output of each key points detected in each image. As below, figure 4.11 is the SIFT experiment results of figure 4.7, 4.8 and 4.9.

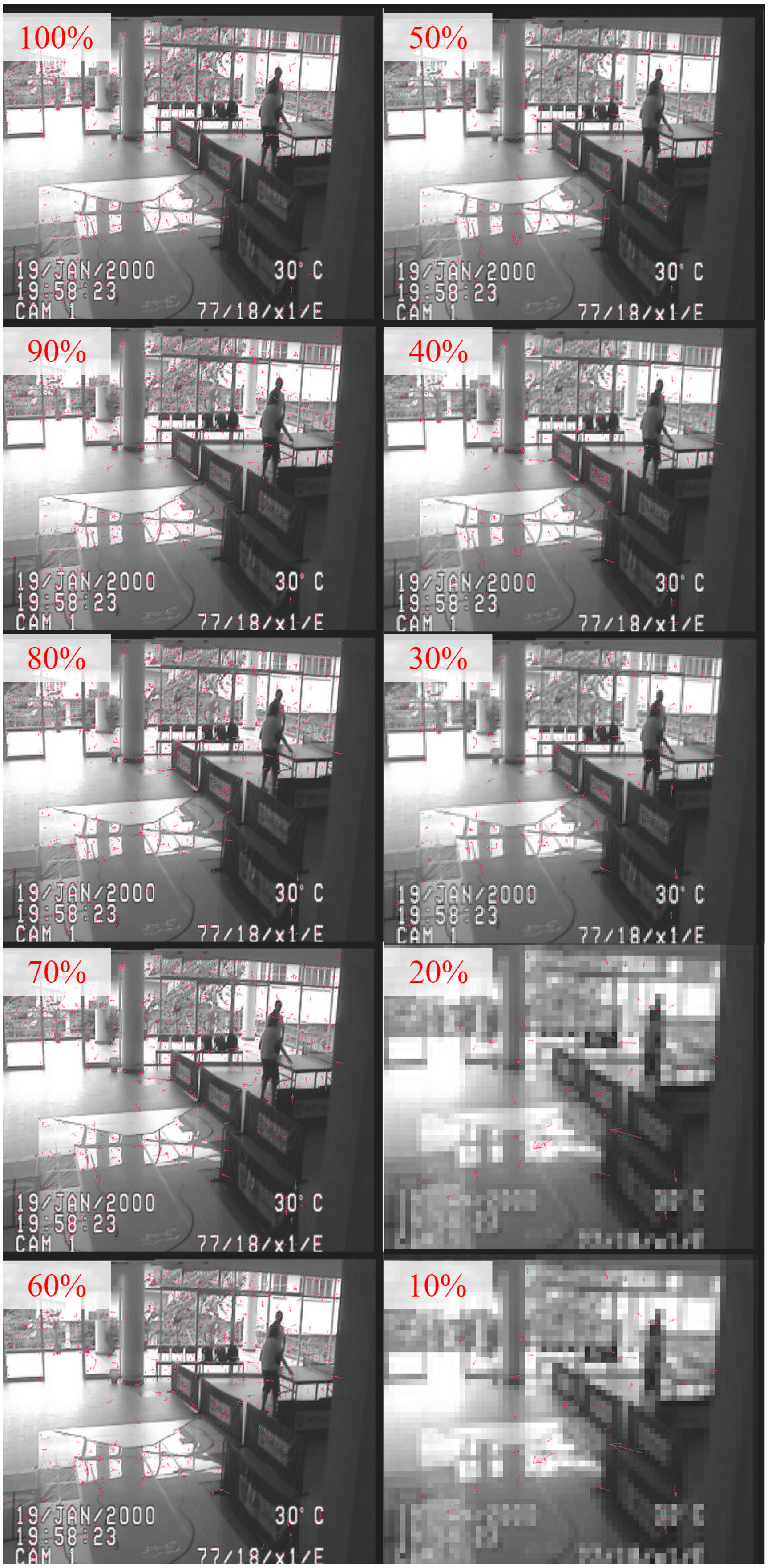


Figure 4.11 (a) – Results of high texture with high contrast

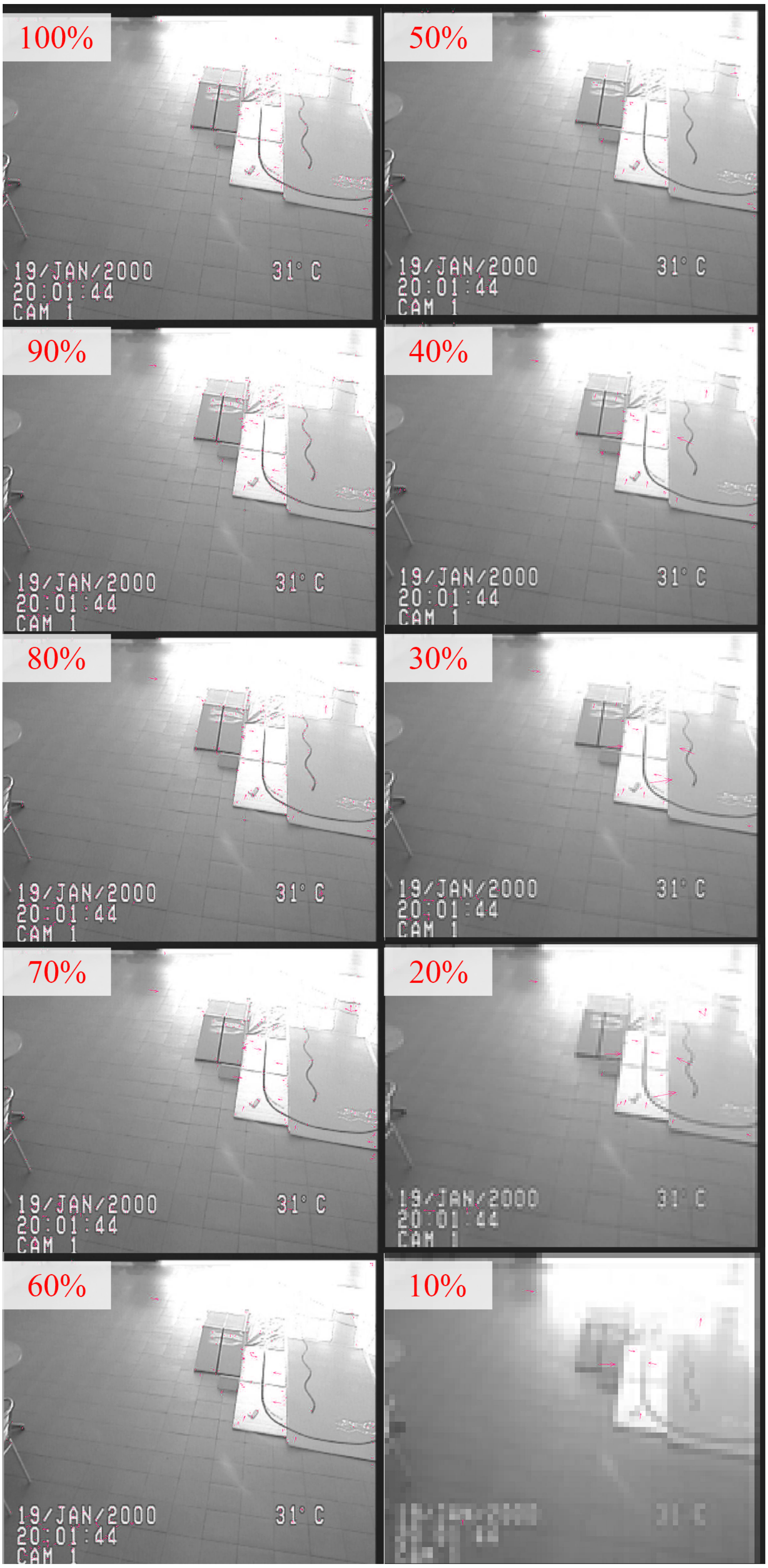
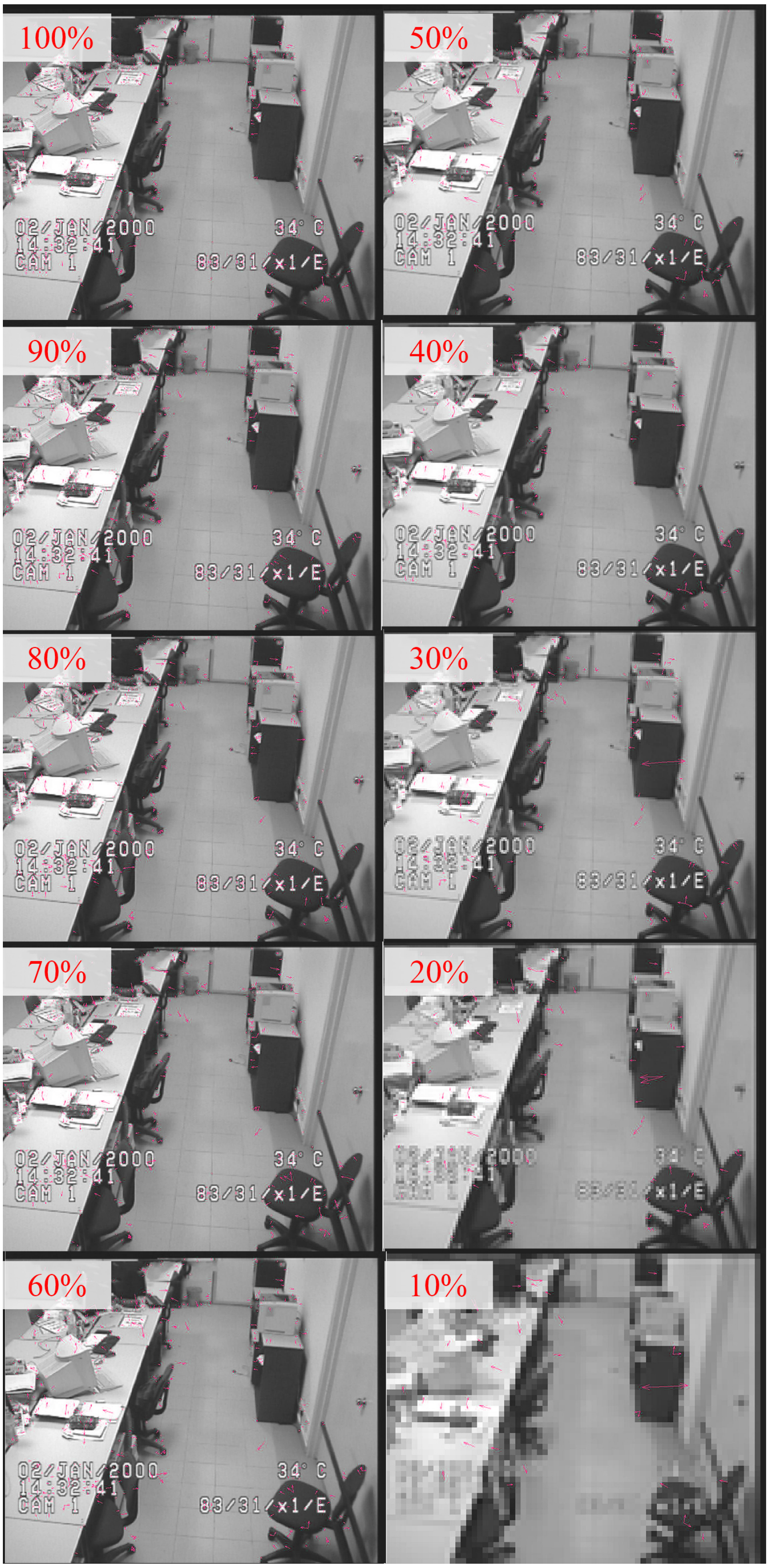


Figure 4.11 (b) - Results of low texture with high contrast



(c) Results of medium texture with consistent contrast

Figure 4.11: The output of the keypoints detected.

Table 4.1: Number of key points detected through the SIFT algorithm of Figure 4.11

Resolution	Pixels	Image(a)	Image(b)	Image(c)
0	0	0	0	0
70 x 58	4060	45	7	36
141 x 115	16215	152	29	108
211 x 173	36503	299	41	208
282 x 230	64860	552	89	335
352 x 288	101376	794	162	470
422 x 346	146012	1144	237	665
493 x 403	198679	1531	355	817
563 x461	259543	1850	474	1030
634 x 518	328412	2074	595	1244
704 x 576	405504	2113	622	1241

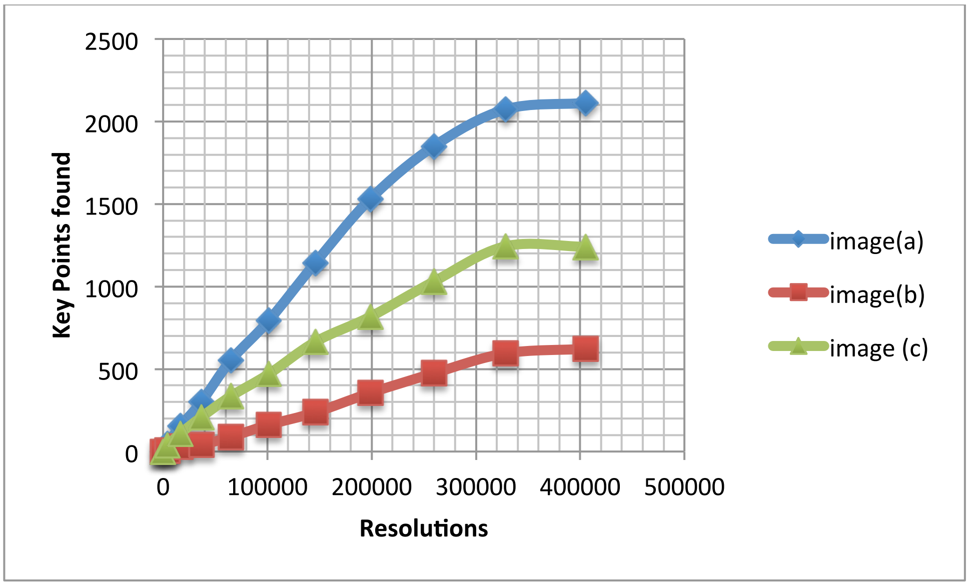


Figure 4.12: Key points found in the images with different resolutions

For comparison, the graph Figure 4.12 according to the observation in Table 4.1. From the graph above, we determined that the number of key points initially increases with image resolution. As we know, to perform reliable SIFT algorithm, features extracted from the sample image is important. It can be detectable under changes in noise, image scale and illumination. Most of the points normallyrely on high-contrast part of image, such as object edges.In figure 4.11, these images had shown that output key points are concentrated at high contrast edges. From figure 4.12, it also proved significantly that image (a) which is high texture and high contrast, found the most number of key points. In the other had, image (b) which is the low texture and high contrast, found the less key points.The number of key points in image (a), (b) and (c) will be nearly constant after an optimum resolution is,for example data image (c). It is because the higher resolution of the image, the more noise will be detected. These noises in the image will form low contrast key points. SIFT algorithm will rejects the low contrast key points.

Table 4.2: Time elapsed through the SIFT algorithm of Figure 4.11

Resolution	Pixels	Image(a)	Image(b)	Image(c)
0	0	0	0	0
70 x 58	4060	2.59	2.96	1.92
141 x 115	16215	6.69	3.09	5.58
211 x 173	36503	12.98	5.73	11.13
282 x 230	64860	22.98	10.26	17.27
352 x 288	101376	35.24	15.5	24.77
422 x 346	146012	50.27	23.66	36.52
493 x 403	198679	73.30	33.77	47.94
563 x461	259543	90.03	46.52	64.55
634 x 518	328412	104.64	64.45	79.38
704 x 576	405504	117.66	71.94	93.62

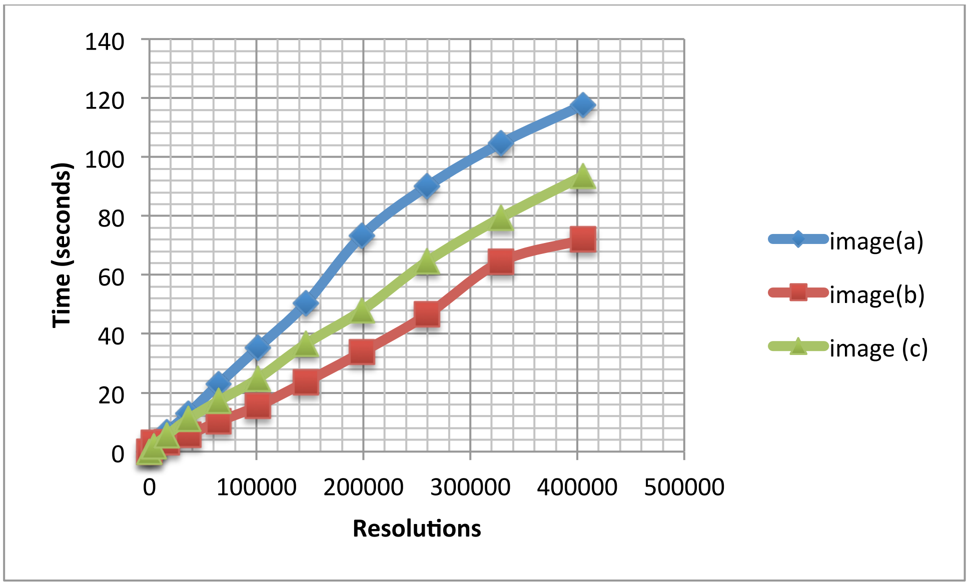


Figure 4.13: Time elapsed in SIFT algorithm process against resolutions.

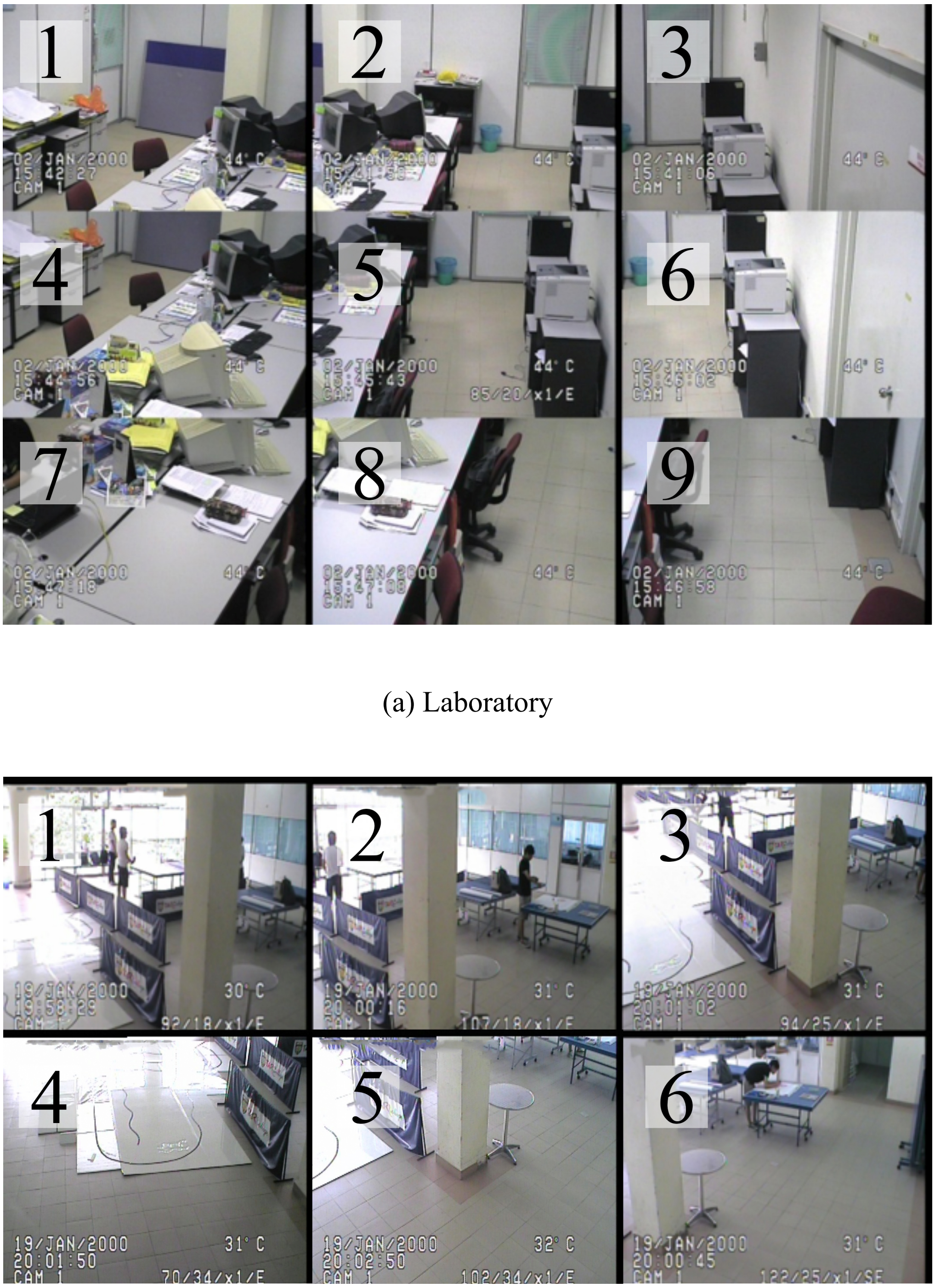
Figure 4.13 was plotted according to results of Table 4.2. It had shown that when image resolutions increased, the processing time increased linearly.

As a conclusion, we can determine that the processing times in Table 4.2 and plotted in Figure 4.13 for different resolutions as listed in Table 4.1 and plotted in Figure 4.12. The number of output key points is affected by size of resolution and also texture of the image. But the number of output keypoints are nearly constant when it hits around 328 412 pixels or 0.3 megapixel. This shows that larger resolution image does not increase the key point detection, but increase detection of false key points and also processing time. So, we decided to optimise the image resolution to 50% of the original image or which is 352 x 288 as our resolution parameter in following experiment.

4.4.3 Experiment II - Image stitching based on Hough Transformation

In this section, the performance evaluation of the proposed improvement of the Lowe's SIFT feature matching algorithm is presented. The goal is to increase the number of correct matches and minimize the number of false matches for an image pair from the wide-angle camera and PTZ camera. As mentioned in Sub-section 4.3, two SIFT features F_1^i and F_2^i are matched when SIFT descriptor of the feature F_2^i has the smallest distance to the descriptor of feature F_1^i as compared to all other extracted features. If the ratio between the Euclidian distances to the nearest neighbour and to the second nearest neighbour is below a threshold, τ then the match is labeled as positive, otherwise it will be labeled as negative. Among positive and negative labeled matches, correct matches as well as false matches can be found.

In experiment I, the SIFT algorithm was conducted in hallway and laboratory. Both location consisted a set of database each, which contained the nine most distinct images was acquired by the testbed as shown in Figure 4.14 below. The quality of the clarity under the laboratory on images showed the good consistency of illumination. However, the image quality in the hallway is over contrast and shows inconsistcy of illumination. These images were subjected to SIFT algorithm to identify the correspondence key points among them. The process results show in appendix D.



(a) Laboratory

(b) Hallway

Figure 4.14: Dataset which acquired by the testbed.

Firstly, we proceed with the laboratory dataset which had better image quality compare to hallway. Table 4.3 shows the number of key points detected in laboratory dataset. Then, hough transformation identified the key points positions of the arbitrary shapes, and classified them in to inliers and outliers as shown in figure 4.15. The yellow and blue "+" symbol are the nearest-neighbour key points. The green "o" symbols are inliers, which are positive, and red "o" symbols are outliers that are negative or false key point. The results reflexed in Table 4.4. For example figure 4.16, images set (a) has been processed by SIFT algorithm. in image 1, it detected 625 key points and 377 key points in image 2. Then, the hough transformation performed robust affine alignment on image 1 based on nearest-neighbor found in image 2. Next, we performed image stitching process according to 189 inliers and 25 outliers that have been recognised. Finally, figure 4.18 as below shows the decent final output of overall stitched image.

Table 4.3: Key points found in laboratory dataset

Image	1	2	3	4	5	6	7	8	9
Key points	625	377	270	878	506	408	693	371	284

Table 4.4: Numbers of inliers and outliers found in Figure 4.14

Set	Inliers (o)	Outliers (o)
a	189	25
b	203	89
c	287	66
d	160	90
e	90	14
f	56	8

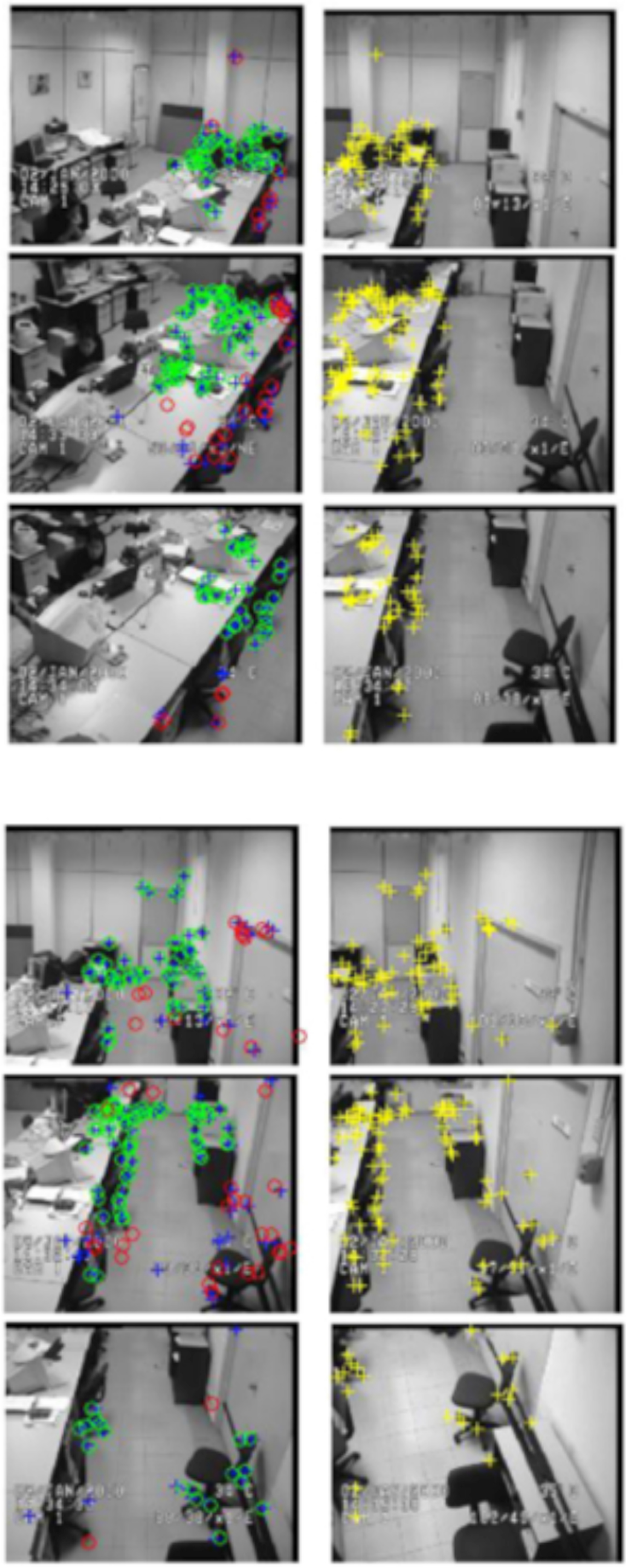


Figure 4.15: Inliers and outliers detection in laboratory dataset.

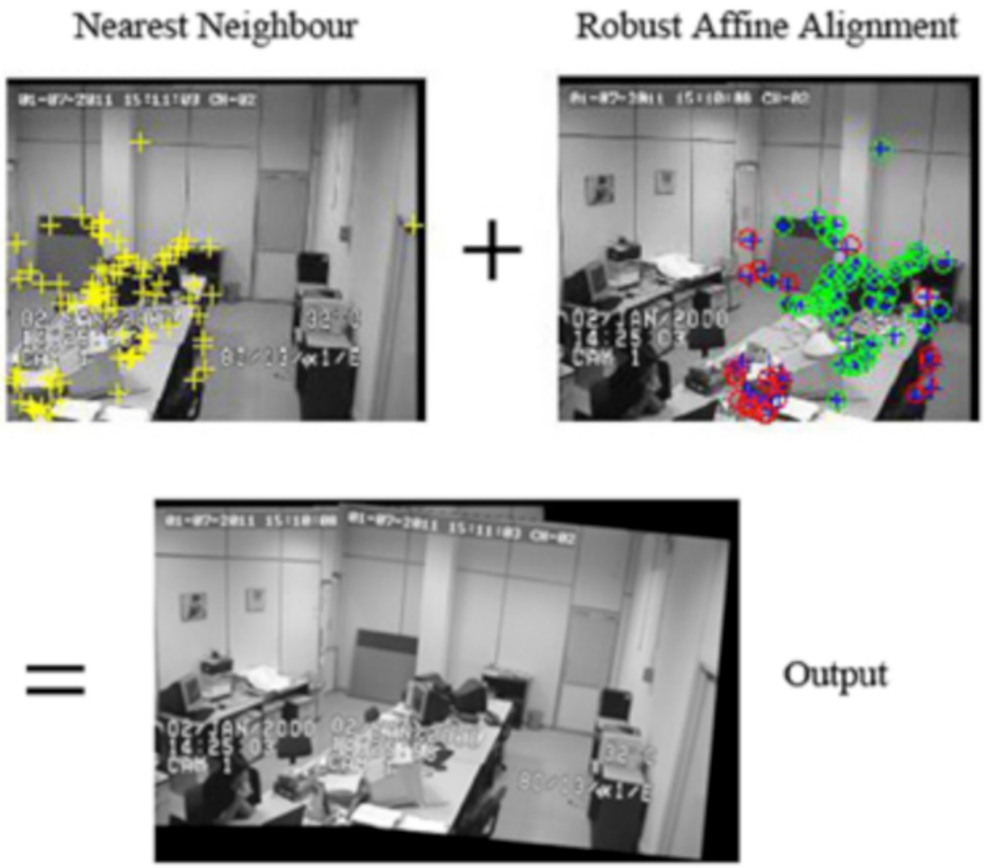


Figure 4.16: Image stitching process and robust estimation based on hough transformation between image 1 and image 2.

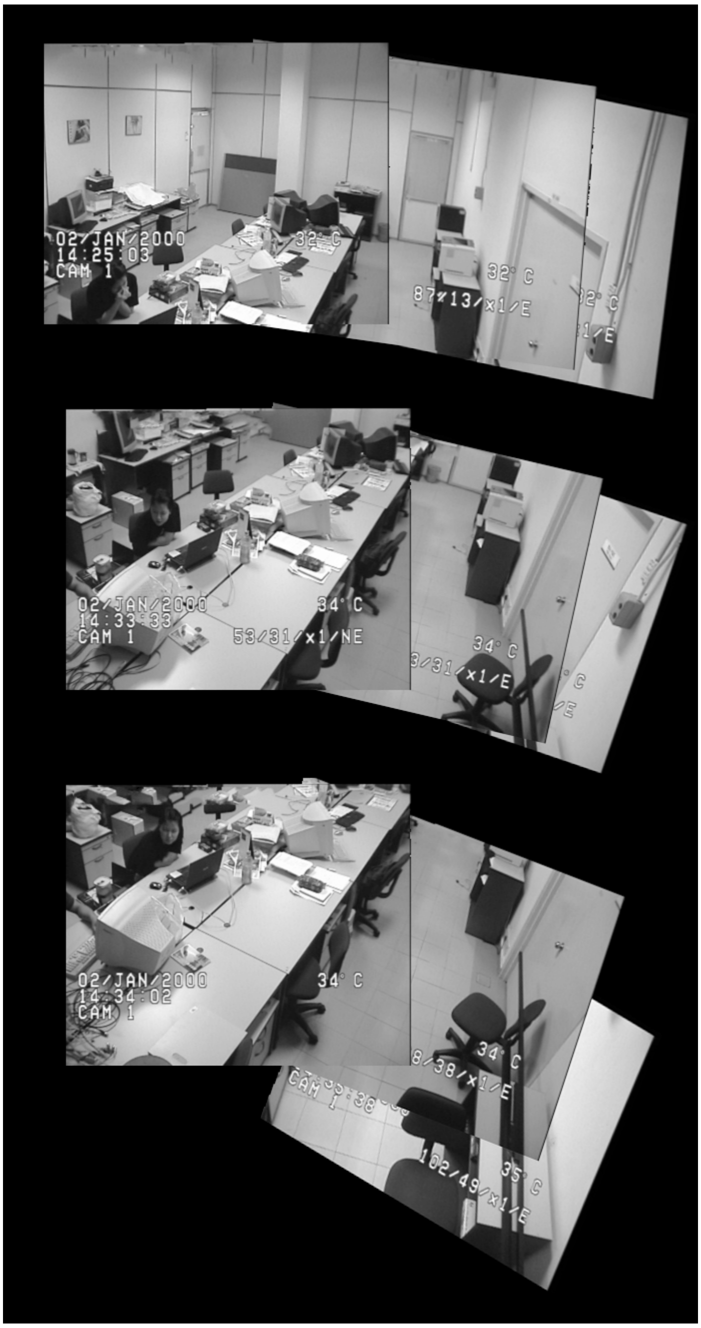


Figure 4.17: Stitch the images portion by portion.

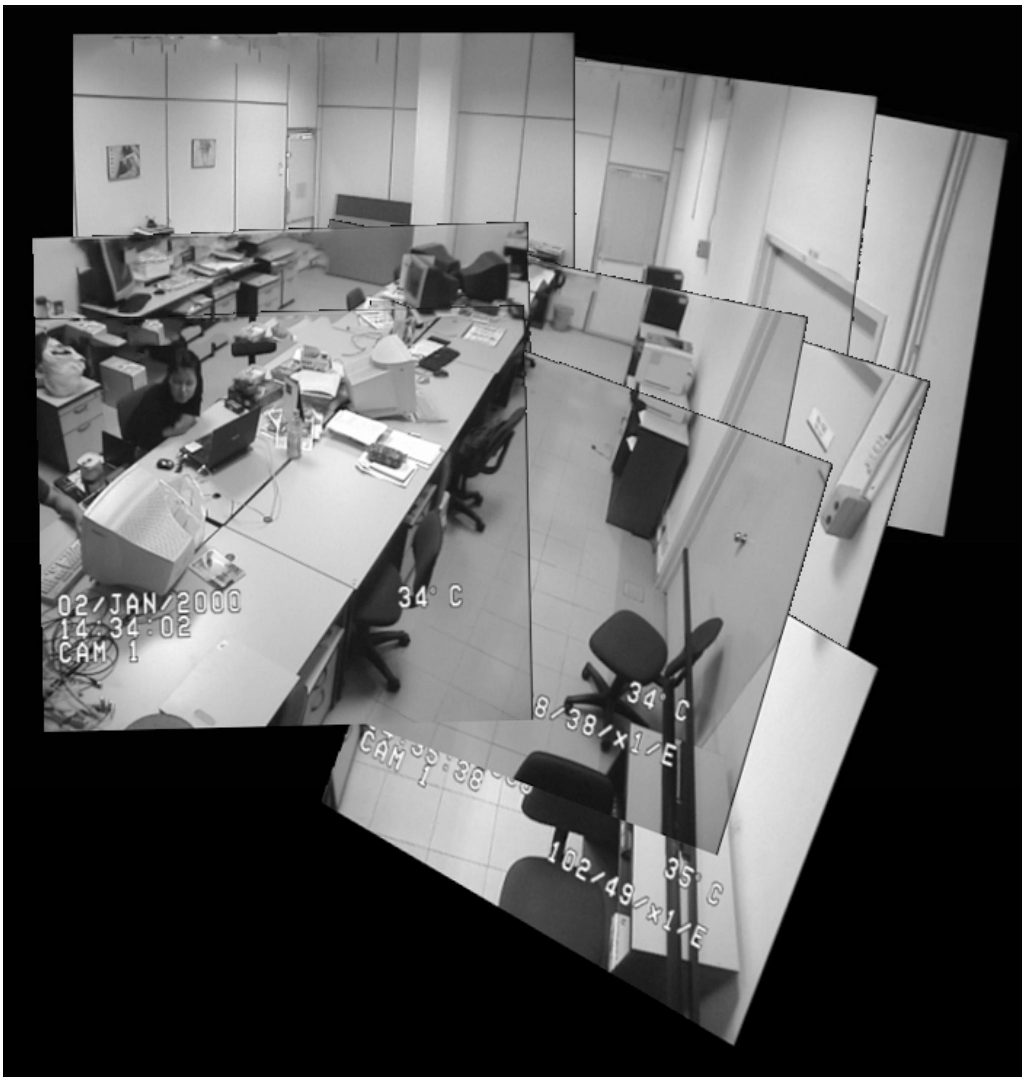
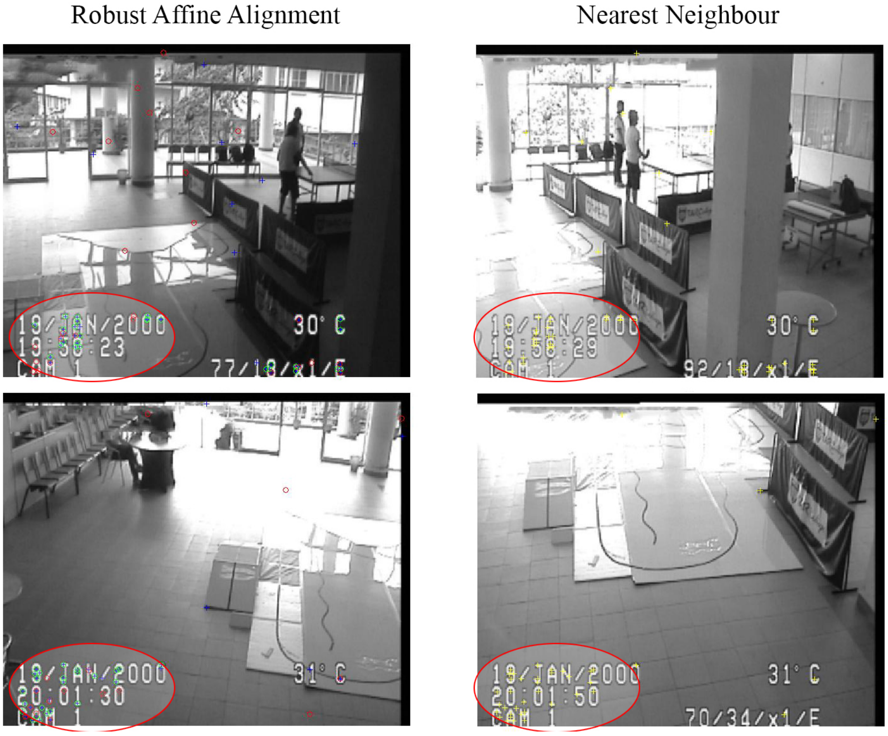


Figure 4.18: Output of image stitching.

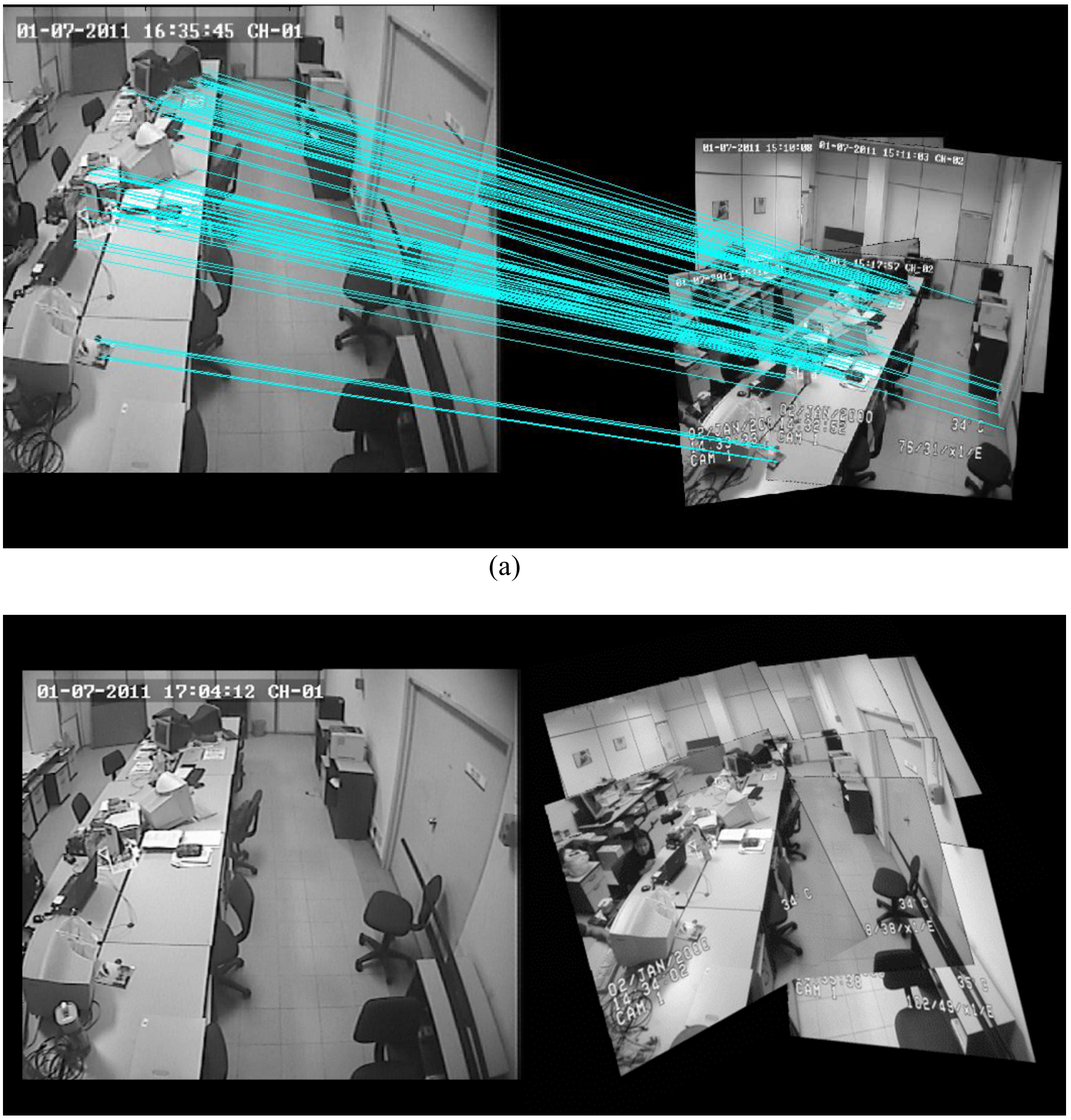
Secondly, hallway dataset which had poor illumination and poor texture shown in figure 4.14(b). As we know, hough transformation identified the key points positions of the arbitrary shapes, the robust affine alignment was effected by poor texture information and over exposed image quality. The images were aligned with the timestamp printed on the image because of the high intensity key points and high similarity shape detected. Figure 4.19 shows the reason results that it failed to continue the stitching process.


Robust Affine Alignment

Nearest Neighbour

Figure 4.19: Robust affline alignment affected by timestamp in the image.

4.4.4 Robust estimation based on RANSAC and ImTrans



(a)

(b)

Figure 4.20: Fast matching process and applied geometric transform on laboratory dataset.

In figure 4.20, the image on the left is image acquired by a wide angle camera in laboratory while images on the right are the panoramic images after matrix affine transformation. In Figure 4.20 (a) fast matching between the wide angle image and panoramic images was performed and in Figure 4.20 (b) the result after geometry transformation using Imtrans algorithm in MATLAB is shown. Next, the panoramic image was transformed and calibrated with wide angle camera image. This procedure will enabled both images to have similar perspective ratio as shown in figure 4.21 below.

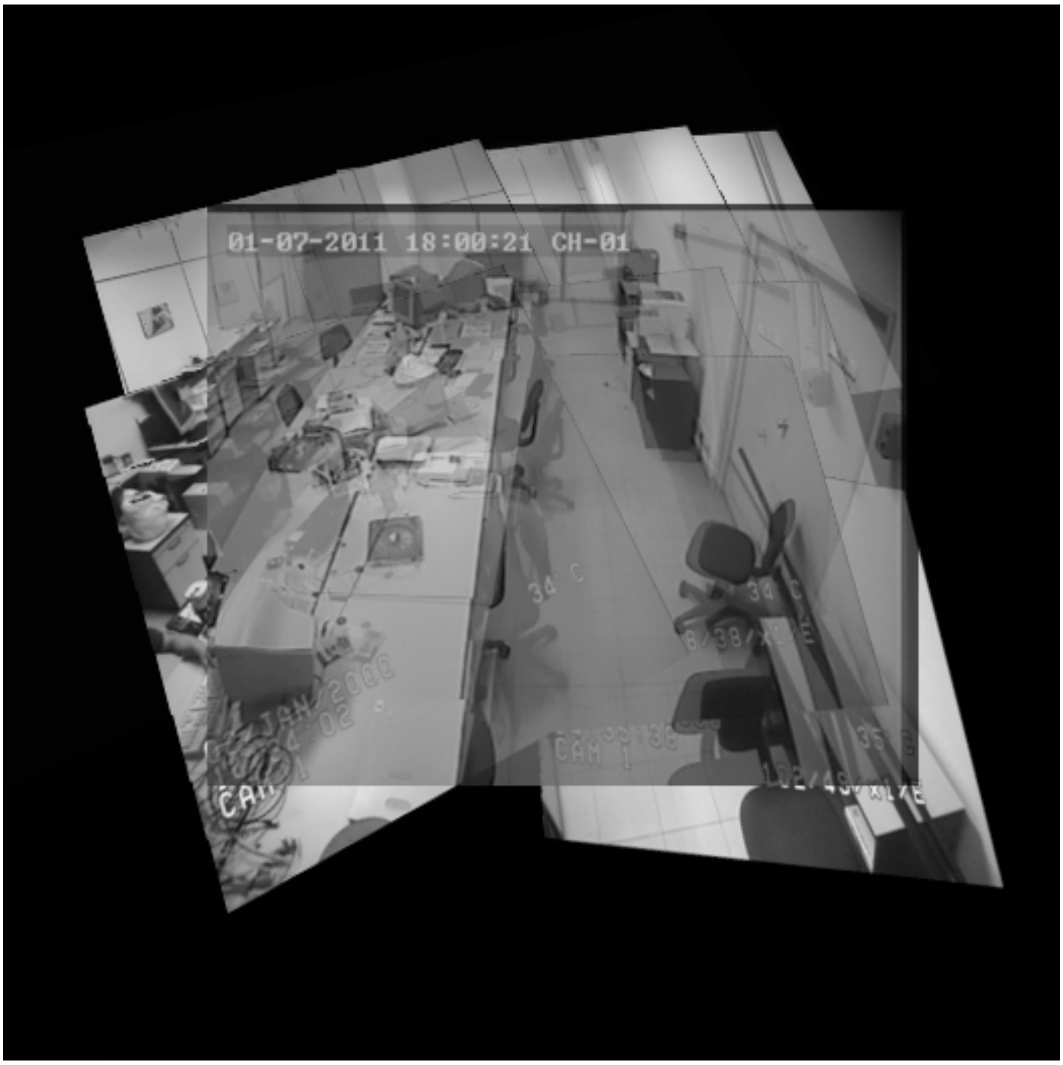


Figure 4.21: Similar perspective ratio between wide angle camera and PTZ camera.

4.4.5 Limitations and Solutions

SIFT transforms an image into a large collection of feature vectors, each of which is invariant to image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion. It requires features matching method, for example RANSAC to identify the matching keys from image A to image B. This technique relied on threshold of distance ration to identified the best candidate match for each keypoint found. The weakness of using this technique is, when threshold increases, the rate of the false key points increase as well. The detection rate of the image will be affected by the objects with high similarity which shown in figure 4.22 below.

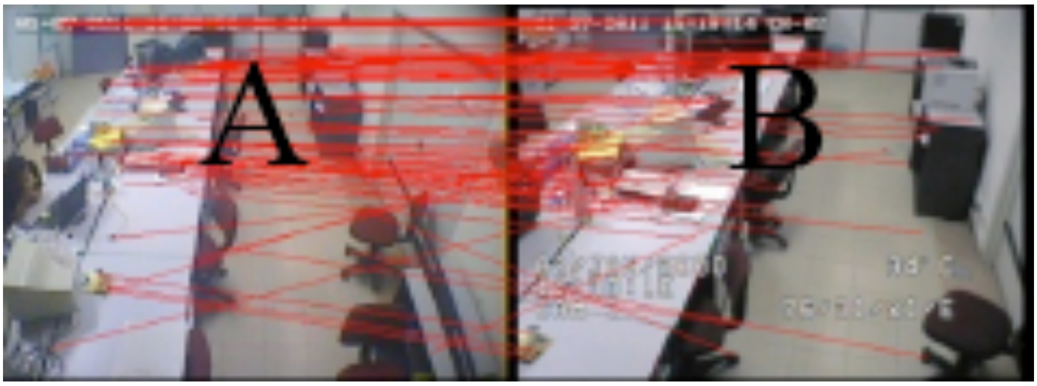


Figure 4.22: Mismatch pairs detected by SIFT algorithm.

In this experiment, we applied filtering method to minimize the false detection by using theorem trigonometry filter. By calculating the gradient of the pairs, we successfully rejected most of the unsed key points as in figure 4.23. The red lines show as positive pairs and blue lines show as floating pairs that are unable to confirm.

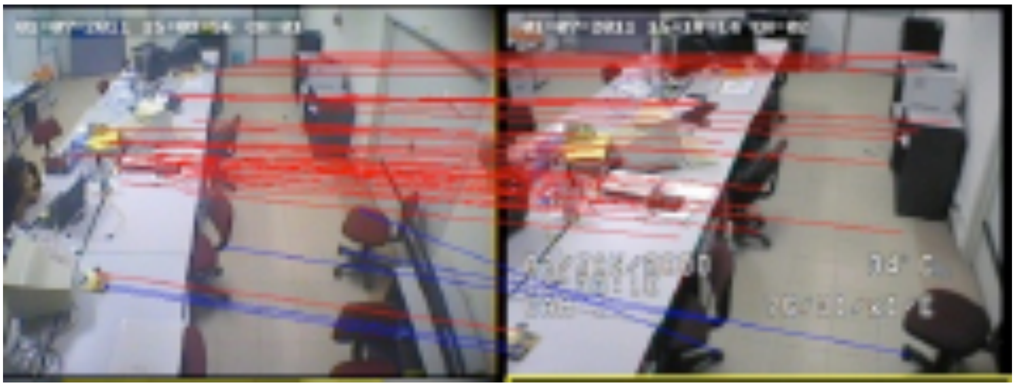


Figure 4.23: Reject the false key points.

Table 4.5: The accuracy(%) by using different threshold values

Matching threshold	**Positive Matching Pairs**	**Negative Matching Pairs**	**Without Filter (%)**	**With Filter (%)**
0.4	39	31	55.71	94.87
0.5	55	46	54.46	90.90
0.6	75	67	52.81	90.67

In table 4.5, different matching threshold value applied to a sample image. Through the process, the threshold value 0.4 computed 39 pairs of positive and 31 pairs of negative gradient keypoints that are found at matching locations. With the larger vote of positive pairs, we defined positive gradient consist more region of interest in the image. So, the calculation of matching percentage is 55.71%. Apparently, the rate of accuracy is increase from 55.71% to 94.87% after applied the theorem trigonometry filter.

Hough transform is limited in efficiency if a higher number of votes fall in the correct bin and the bin can be detected easily amid the background noise. The scaling of the bin if in too small, votes will be fall into neighbouring bins and this causes the visibility of the main bin will be decreased.This is the reason which caused hallway dataset image stitching to fail. The bin around the timestamp on the images get high number of votes compared to other bin which lack of visibility.To resolve the problem, we decided to remove the timestamp on the images. Then, we continued the whole process again by using hallway dataset. The results show in figure 4.24, figure 4.25 and figure 4.26 below.

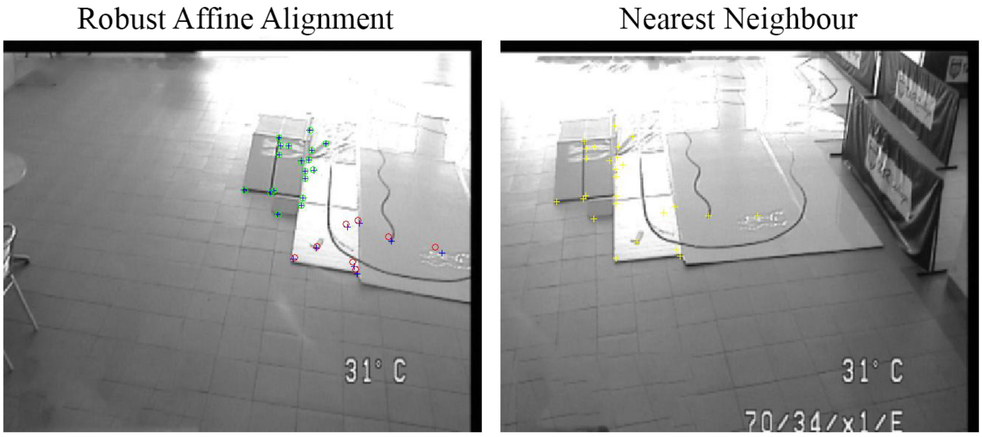


Figure 4.24: Fixed image stitching process by removing the timestamp on the images

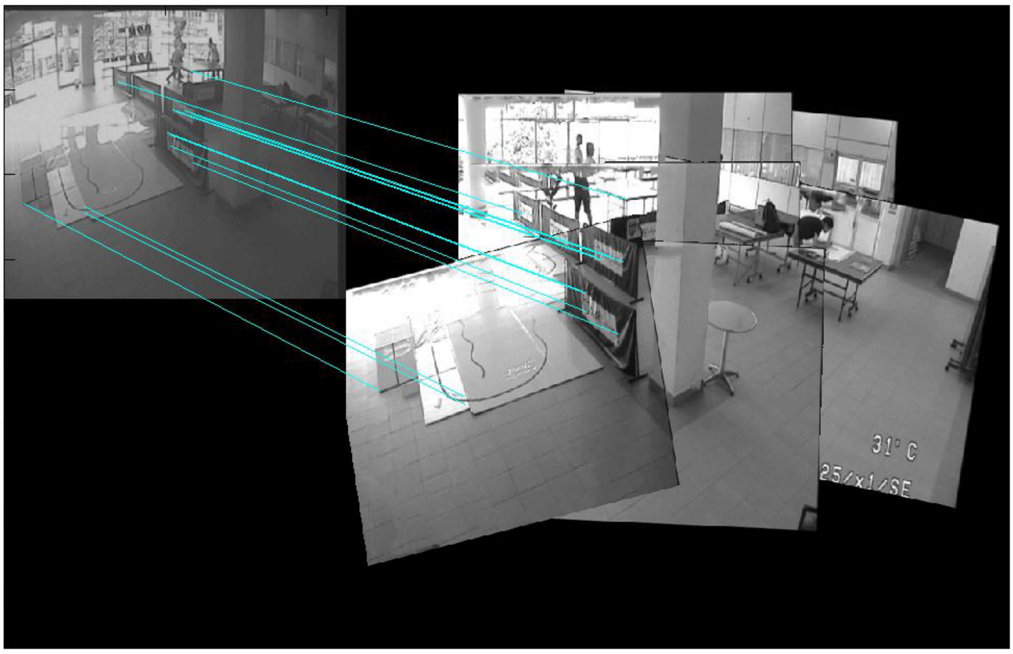


Figure 4.25: Fast matching process on hallway dataset.

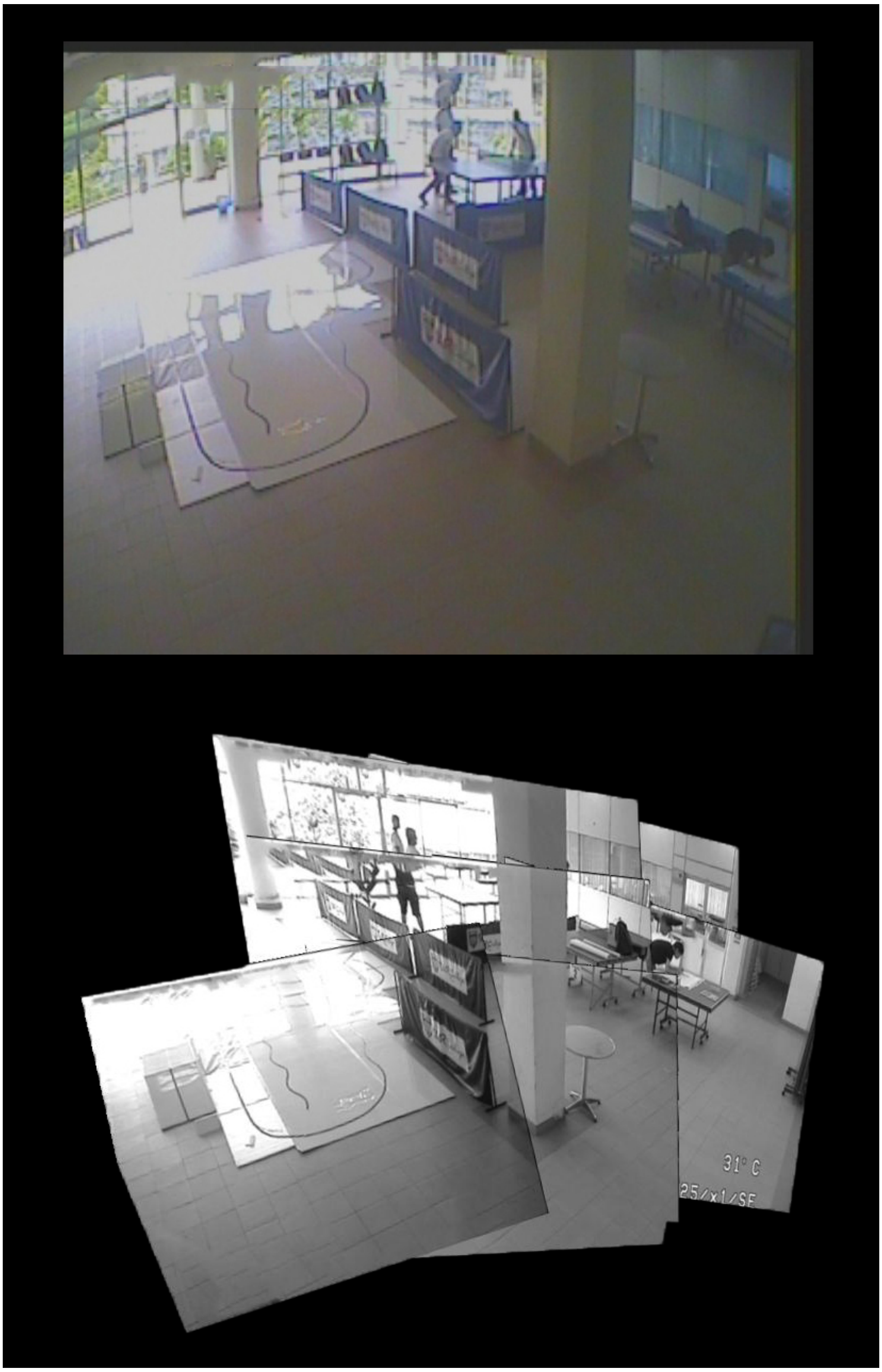


Figure 4.26: Comparison wide angle image with panoramic image after applied geometric transform on hallway dataset.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this project, a comprehensive technical and background study on hybrid-camera system has been carried out. To collect the information and dataset, we have constructed a prototype system that allowsobjects to be moved around from indoors to outdoors.

The major contribution of this project has been the developments of a novel image stitching algorithm, that can automatic recognize and stitch high quality image panoramas from image datasets. With that we are able to match panoramic image to image from wide-angle camera effectively. The prototype system was used to evaluate various multiple image matching methods and test the performance and tune the parameters of our stitching algorithm.

We carried out a survey and study on various relevant approaches and techniquesrelated to the scope of the project and have proposed the application of SIFT to identify the correspondence keypoints in image dataset. There are several advantages to such an approach. Firstly, by organizing the features of *n* images into a feature database,the complexity of matching images can be reduced. Secondly, the geometric constraints for multiple views are stronger than their pairwise counterparts and this allows more incorrect matches to be rejected. Finally, we can exploit the probabilistic nature of *n* images matching problem by using known incorrect matches in a data driven classifier.

Lastly, we have also carried out the research on how to improve the accuracy and robustness by rejecting unwanted keypoints based on trigonometry theorem. The rate of accuracy is increase from 55.71% to 94.87%. This has shown great improvement in terms of determination and detection rate in camera networks of Y.Q. Low et al. (2011).

5.2 Future Work

We conclude by identifying some avenues for future explorations:

In Chapter 3 and 4 we discussed automatic image stitching from multiple views. Illumination, sensors and optics can affect the key points found. These problems can be thought of within the general framework of computational photography, which refers to computational image capture, processing and manipulation techniques that enhance or extend the capabilities of digital photography. Typically, it is about merging multiple pictures of the same subject matter by using different exposure parameters. This is well handled by the illuminated images and object to focus. In the future, there may be no such "poor" photograph, because the shooting conditions will be completely reconfigurable after the event has been recorded.

Currently, the images need to be sorted out manually before stitching. With the advancement of digital photography, our ability to understand images has not kept up in pace with our ability to generate them. In the future, algorithms for searching and sorting in image database will become as fundamental and available as those for searching for text on the World-Wide Web.

In this project, further increase of the clarity of the image background, background learning process is needed to differentiate the foreground and the background objects.This technique will be able to solve the occlusion of objects while its moving. Flexible models to correspondence with similar metrics linked to human perception will be required for progress in this area.

REFERENCES

Aghajan, H. and Caavallaro, A., 2009. *Multi-Camera Networks: Principles and Applications*, Elsevier Inc.

Baker, P., and Aloimonos, Y., 2000. Complete Calibration of a Multi-camera Network. *DOI 10.1109/ OMNVIS.2000.853820*

Bay,H,. Tuytelaars, T., and Van Gool, L., 2006. SURF: Speeded Up Robust Features, *Computer Vision - European Conference on Computer Vision 2006*, Volume 3951, 2006, pp. 404-417

Brown M., and Lowe, D.G., 2002. Cardiff: Invariant features from interest point groups. *British Machine Vision Conference*, pp. 656-665.

Canny, J., 1986. A Computational approach to Edge Detection. *Pattern Analysis and Machine Intelligence*, PAMI-8(6):679-697.

Chen, I.H, and Wang, S.J., 2007. An Efficient Approach for the Calibration of Multiple PTZ cameras. *Automation Science And Engineering,* 4, pp. 286 - 293

Del Bimbo, A., Dini, F., Lisanti, G., and Pernici, F., 2010. Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks. *Computer vision and Image Understanding 114*, pp. 611-623.

Fischler M., and Bolles R., 1981. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM*, 24:381-395.

Fisher R., Perkins S., Walker A., and Wolfart, E., 2000, *Affine Transformation* [Online].
Available at: http://homepages.inf.ed.ac.uk/rbf/HIPR2/affine.htm [Accessed: 25 January 2011]

Forsyth, D., and Ponce, J., 2003. *Computer Vision: A Modern Approach*, Prentice Hall.

Gower, J.C., 1982, *Math. Scientist: Euclidean Distance Geometry*, 7, pp. 1-14.

Harris, C., and Stephens, M.J., 1988. A combined corner and Edge Detector. *Alvey Vision Conference*, pp. 147-152.

Henkel, R.D., 1997. Fast Stereovision by Coherence Detection. *Computer Analysis Of Image and Pattern*, Volume 1296, pp. 297-304.

Hough, P.V.C., 1962. Method and means for recognizing complex patterns. *U.S. Patent 3069654*.

Ke, Y., and Sukthankar, R., 2004. PCA-SIFT: “A More Distinctive Representation for Local Image Descriptors”. *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 511-517.

Krumm,J., Harris,S., Meyers, B., Brumitt, B., Hale M., and Shafer S., 2000. Multi-Camera Multi-Persom Tracking for EasyLiving. *Microsoft Research Vision Technology Group*, Third IEEE International Workshop on Visual Surveillance.

Lee, C.W., Sebastian, P., and Yap, V.V., 2009. Stereo Vision Tracking System. *Future Computer and Communication*, pp487-491.

Lindeberg T., 1994. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(2):pp. 224-270.

Liu, J., and Hubbold, R., 2006. Automatic Camera Calibration and Scene Reconstruction with Scale-Invariant Features. *International Symposium on Visual Computing,* pp. 558-568.

Liu, R., Zhang, H., Liu, M., Xia, X., and Hu, T., 2009. Stereo Cameras Self-Calibration Based on SIFT, *Measuring Technology and Mechatronics Automation,* pp. 352-355.

Low, Y.Q., Lee, S.W., Goi, B.M., and Ng, M.S., 2011. A New SIFT-based Camera Calibration Method for Hybrid Dual-Camera. *ICIEIS*, pp96-103, 2011.

Lowe, D.G., 1999. Object Recognition from Local Scale-Invariant Features. *Proc. Of the International Conference on Computer Vision*, pp. 1150-1157.

Lowe, D.G., 2001. Local feature view clustering for 3D object recognition, *IEEE Conferences on Computer Vision and Pattern Recognition*, pp. 682-688.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, pp.91-110.

Matlab Central [Online]. Available at: http://www.mathworks.com/matlabcentral/fileexchange/28760-2d-2d-projective-homography-3x3-estimation/content/homography.m [Accessed: 25 April 2011]

Memont ,Q., and Khant, S., 2011. Camera calibration and three-dimensional world reconstruction of stereo-vision using neural networks, *International Journal of System Science*, pp 1155-1159.

Protocol Manual, *“D” Protocol* [Online]. Available at: http://www.optovision.fr/telechargement/Pelco-D_protocol.pdf [Accessed: 01 April 2011]

Rafael, C., Gonzalez, and Woods, R.E., 2007. *Digtal Image Processing Third Edition*, Pearson Prentice Hall.

Senior, A.W., Hampapur, A., and Lu, M., 2005. Acquiring Multi-scale Images by Pan-Tilt-Zoom Controls and Automatic Multi-Camera Calibration. *Application of Computer Vision Volume 1*, pp. 433-438.

Shotton,J., Blake,A., and Cipolla, R., 2005. Contour-Based Learning for Object Detection. *International Conference of Computer Vision*, pp. 503-510.

University of Calgary, *Homogeneous Coordinates* [Online]. Available at: http://www.ece.uvic.ca/~bctill/20006/additional/homcoord/homog-coords.pdf [Accessed: 01 April 2012]

University of Nevada, Reno, *The Geometry of Perspective Projection* [Online]. Available at: http://www.cse.unr.edu/~bebis/CS791E/Notes/PerspectiveProjection.pdf [Accessed: 01 March 2012]

University of Oxford, *Introduction – a Tour of Multiple View Geometry.* [Online]. Available at: http://www.robots.ox.ac.uk/~vgg/hzbook/hzbook2/HZintroduction.pdf [Accessed: 06 January 2012

Wan, D., and Zhou, J., 2008. Stereo Vision using two PTZ cameras. *Computer vision and Image Understanding 112*, pp. 184-194.

Xing, Y.J., Xing, J., Sun, J., and Hu L., 2007. An Improved Neural Networks for Stereo-camera Calibration. *Journal of Achievements in Materials and Manufacturing Engineering*, volume 20, issues 1-2.

Yokoyama, M., and Poggio, T. 2005. A Contour-Based Moving Object Detection and Tracking. *International Conference of Computer Vision*, pp. 271-276.

Zhou, X., Collins, R., Kanade, T., and Metes, P., 2003. A master-slave system to acquire biometric imagery of humans at a distance. *ACM International Workshop on Video Surveillance*, pp. 113-120.

Appendix A

IMTRANS - Homogeneous transformation of an image

```
function newim = imTrans(im, T, region, sze);

im = double(im)/255;  % Make sure image is double

threeD = (ndims(im)==3);  % A colour image

if threeD    ## Transform red, green, blue components separately
  r = transformImage(im(:,:,1), T, region, sze);
  g = transformImage(im(:,:,2), T, region, sze);
  b = transformImage(im(:,:,3), T, region, sze);

  # Fix to correct the image - for some reason
  # it comes out mirrored left-to right
   r = fliplr(r);
  g = fliplr(g);
  b = fliplr(b);

  newim = repmat(uint8(0),[size(r),3]);
  newim(:,:,1) = uint8(round(r*255));
  newim(:,:,2) = uint8(round(g*255));
  newim(:,:,3) = uint8(round(b*255));

else               # Assume the image is greyscale
  newim = transformImage(im, T, region, sze);
  # Applying Fix
  newim = fliplr(newim);
end

#-----------------------------------------------------------

# The internal function that does all the work

function newim = transformImage(im, T, region, sze);

[rows, cols] = size(im);

# Determine default parameters if needed
if nargin == 2
  region = [0 rows 0 cols];
  sze = max(rows,cols);
elseif nargin == 3
  sze = max(rows,cols);
elseif nargin ~= 4
  error('Incorrect arguments to imtrans');
end

# Find where corners go - this sets the bounds on the final image
B = bounds(T,region);
nrows = B(2) - B(1);
ncols = B(4) - B(3);

# Determine any rescaling needed
s = sze/max(nrows,ncols);
```

```
S = [s 0 0       # Scaling matrix
     0 s 0
     0 0 1];

T = S*T;
Tinv = inv(T);

# Recalculate the bounds of the new (scaled) image to be generated
B = bounds(T,region);
nrows = B(2) - B(1);
ncols = B(4) - B(3);

newim = zeros(nrows,ncols);

[x,y] = meshgrid(1:ncols,1:nrows);   # All possible xy coords in the image.

# Transform these xy coords.
sxy = Trans(Tinv, [x(:)'+B(3) ; y(:)'+B(1) ; ones(1,ncols*nrows)]);

xi = reshape(sxy(1,:),nrows,ncols);
yi = reshape(sxy(2,:),nrows,ncols);
[x,y] = meshgrid(1:cols,1:rows);

newim = interp2(x,y,double(im),xi,yi);        # Interpolate values from source image


#----------------------------------------------------------------------
#
# Internal function to find where the corners of a region, R
# defined by [minrow maxrow mincol maxcol] are transformed to
# by transform T and returns the bounds, B in the form
# [minrow maxrow mincol maxcol]

function B = bounds(T, R)

P = [R(3) R(4) R(4) R(3)      # homogeneous coords of region corners
    R(1) R(1) R(2) R(2)
     1   1   1   1  ];

PT = round(Trans(T,P));

B = [min(PT(2,:)) max(PT(2,:)) min(PT(1,:)) max(PT(1,:))];
%      minrow        maxrow      mincol      maxcol
```

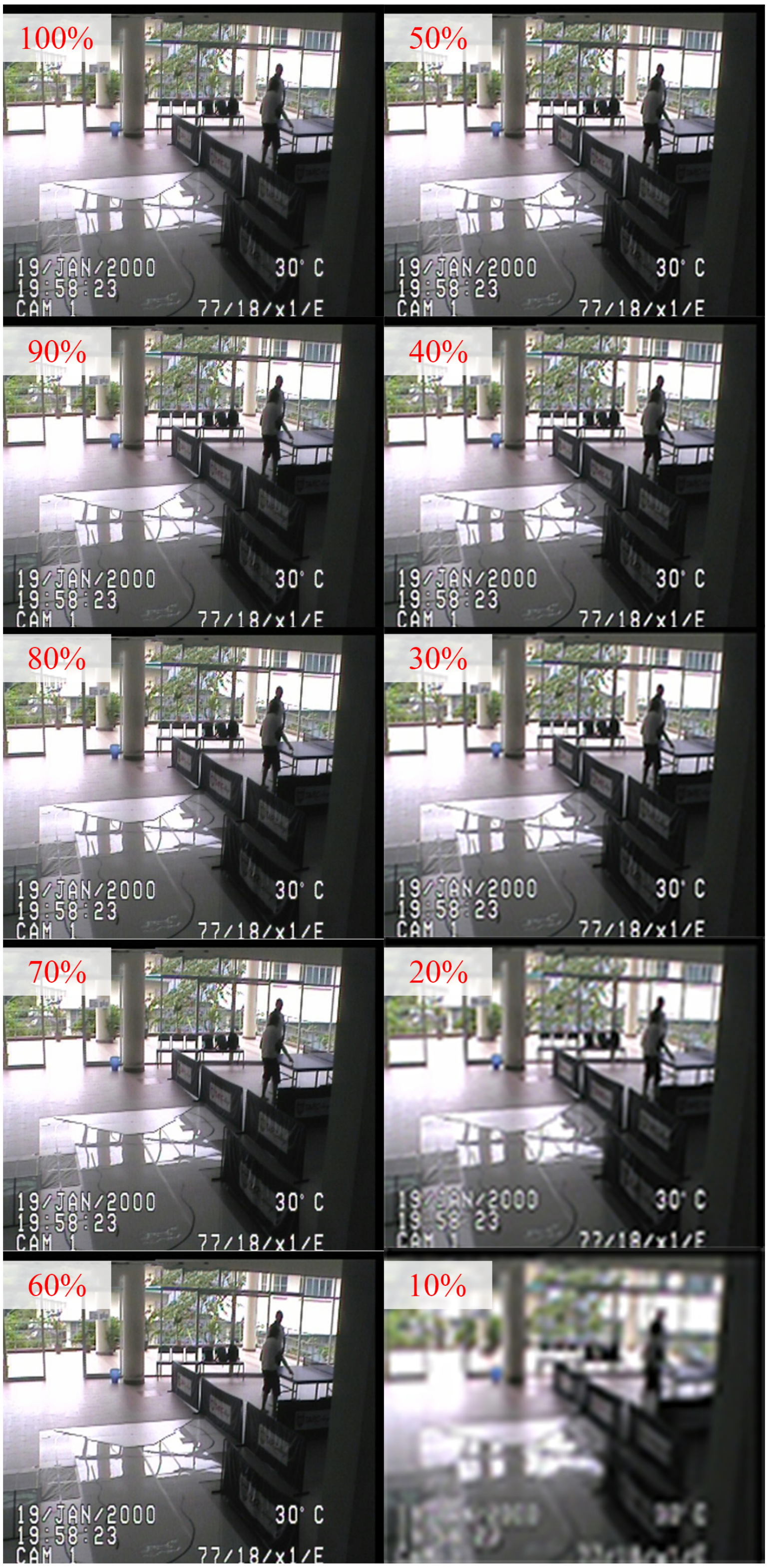
Appendix B

Sample cammads of the PTZ camera

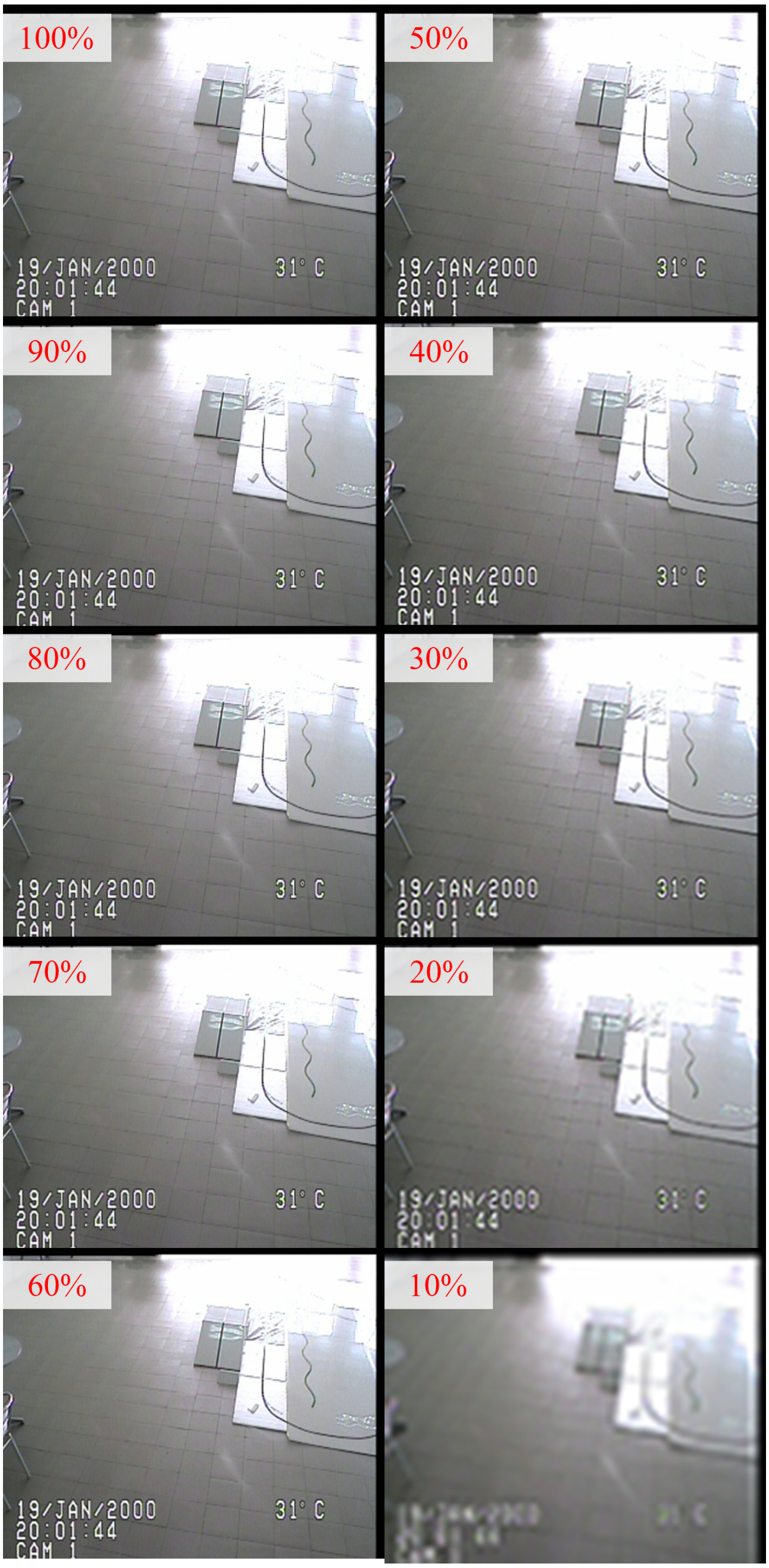
Command	Word 3	Word 4	Word 5	Word 6
Set Preset	00	03	00	01 to 20
Clear Preset	00	05	00	01 to 20
Go To Preset	00	07	00	01 to 20
Flip (180° about)	00	07	00	21
Go To Zero Pan	00	07	00	22
Set Auxiliary	00	09	00	01 to 08
Clear Auxiliary	00	0B	00	01 to 08
Remote Reset	00	0F	00	00
Set Zone Start	00	11	00	01 to 08
Set Zone End	00	13	00	01 to 08
Write Char. To Screen	00	15	X Position 00 to 28	ASCII Value
Clear Screen	00	17	00	00
Alarm Acknowledge	00	19	00	Alarm No.
Zone Scan On	00	1B	00	00
Zone Scan Off	00	1D	00	00
Set Pattern Start	00	1F	00	00
Set Pattern Stop	00	21	00	00
Run Pattern	00	23	00	00
Set Zoom Speed	00	25	00	00 to 03
Set Focus Speed	00	27	00	00 to 03
Reset Camera to defaults	00	29	00	00
Auto-focus auto/on/off	00	2B	00	00-02
Auto Iris auto/on/off	00	2D	00	00-02
AGC auto/on/off	00	2F	00	00-02
Backlight compensation on/off	00	31	00	01-02
Auto white balance on/off	00	33	00	01-02
Enable device phase delay mode	00	35	00	00
Set shutter speed	00	37	Any	Any
Adjust line lock phase delay	00-01	39	Any	Any
Adjust white balance (R-B)	00-01	3B	Any	Any
Adjust white balance (M-G)	00-01	3D	Any	Any
Adjust gain	00-01	3F	Any	Any
Adjust auto-iris level	00-01	41	Any	Any
Adjust auto-iris peak value	00-01	43	Any	Any
Query1	00	45	Any	Any

Appendix C

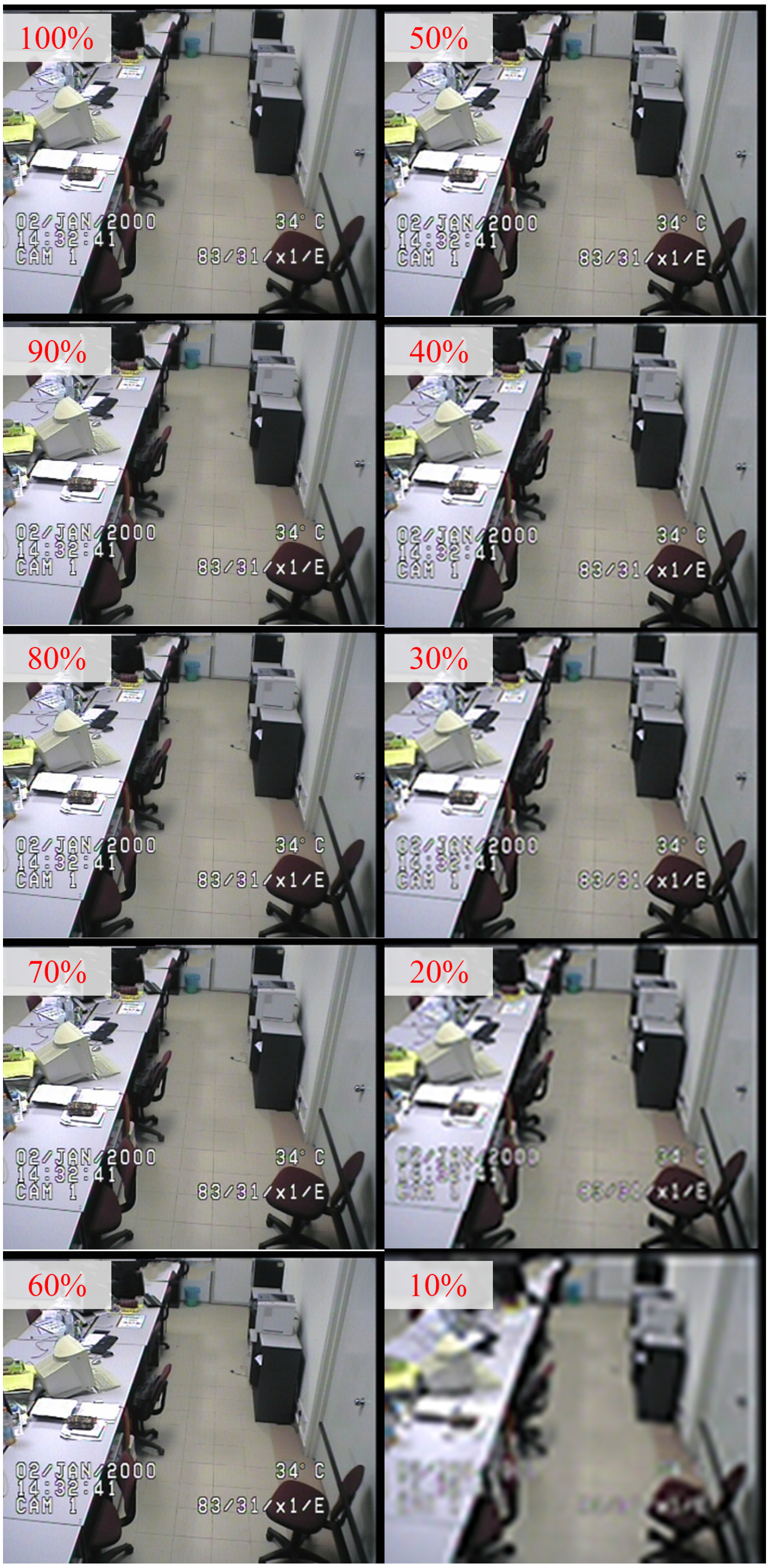
Different resolutions sample dataset.



(a)



(b)


100%
50%
90%
40%
80%
30%
70%
20%
60%
10%
02/JAN/2000
14:32:41
CAM 1
34° C
83/31/x1/E

(c)

Appendix D

SIFT algorithm results of the laboratory dataset.

Image 1

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	0.51	0.05	0.06	0.08	0.11	0.14	0.17	0.22	0.27	0.32
DoG pyramids	0.03	0.05	0.07	0.10	0.14	0.18	0.24	0.30	0.37	0.45
Locating Keypoints	0.38	1.52	3.56	6.50	10.84	15.82	23.38	29.76	38.19	48.83
Gradient	0.02	0.02	0.04	0.06	0.10	0.13	0.18	0.23	0.29	0.35
Orientation assignment	0.45	0.16	0.30	0.51	0.77	1.09	1.34	1.52	1.81	1.85
Features Descriptor	2.98	6.10	12.12	21.46	34.05	44.58	60.13	68.44	79.85	77.17
Total Processing time	4.37	7.90	16.15	28.71	46.01	61.94	85.44	100.47	120.78	128.97
Keypoints Found	47	113	224	396	625	809	1064	1229	1446	1383

Image 2

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	0.01	0.01	0.03	0.05	0.08	0.11	0.15	0.19	0.24	0.29
DoG pyramids	0.03	0.03	0.06	0.09	0.15	0.17	0.22	0.29	0.35	0.44
Locating Keypoints	0.34	1.34	3.19	5.85	9.77	14.52	20.71	27.41	35.58	45.72
Gradient	-	0.01	0.03	0.06	0.09	0.12	0.18	0.22	0.28	0.34
Orientation assignment	0.06	0.11	0.20	0.33	0.47	0.60	0.78	0.99	1.20	1.17
Features Descriptor	2.12	4.54	8.08	14.74	20.43	27.18	35.01	46.20	52.61	53.24
Total Processing time	2.56	6.04	11.59	21.12	30.99	42.70	57.05	75.30	90.26	101.20
Keypoints Found	39	84	150	267	377	493	618	832	960	957

Image 3

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.01	0.03	0.05	0.08	0.11	0.14	0.18	0.23	0.29
DoG pyramids	0.20	0.03	0.06	0.09	0.14	0.17	0.23	0.29	0.36	0.43
Locating Keypoints	0.30	1.24	3.03	5.60	9.68	14.25	20.36	26.92	35.30	45.03
Gradient	-	0.02	0.03	0.06	0.09	0.13	0.17	0.22	0.28	0.34
Orientation assignment	0.03	0.08	0.15	0.20	0.31	0.43	0.60	0.77	0.88	0.90
Features Descriptor	1.20	3.19	5.56	8.52	14.64	18.43	26.72	35.80	41.20	42.67
Total Processing time	1.73	4.57	8.86	14.52	24.94	33.52	48.22	64.18	78.25	89.66
Keypoints Found	22	59	103	157	270	337	482	644	749	769

Image 4

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.01	0.03	0.05	0.08	0.11	0.15	0.18	0.24	0.29
DoG pyramids	0.02	0.03	0.06	0.09	0.13	0.17	0.23	0.29	0.36	0.43
Locating Keypoints	0.42	1.73	4.03	7.18	11.58	17.29	23.94	30.98	39.59	50.43
Gradient	-	0.02	0.03	0.06	0.09	0.13	0.19	0.22	0.28	0.36
Orientation assignment	0.07	0.22	0.41	0.69	1.04	1.36	1.68	2.01	2.28	2.34
Features Descriptor	2.55	9.46	18.39	30.74	48.35	59.59	74.21	92.18	102.61	103.72
Total Processing time	3.06	11.47	22.95	38.81	61.27	78.65	100.40	125.86	145.36	157.57
Keypoints Found	47	176	335	566	878	1090	1338	1655	1857	1855

Image 5

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.01	0.03	0.05	0.08	0.11	0.15	0.18	0.23	0.29
DoG pyramids	0.02	0.03	0.06	0.09	0.13	0.18	0.24	0.29	0.35	0.44
Locating Keypoints	0.38	1.58	3.64	6.76	10.66	16.15	21.85	29.09	37.33	48.11
Gradient	-	0.02	0.03	0.06	0.09	0.13	0.17	0.23	0.28	0.35
Orientation assignment	0.05	0.15	0.30	0.44	0.61	0.85	1.07	1.28	1.51	1.48
Features Descriptor	2.36	6.47	11.79	20.58	27.52	38.98	47.36	59.79	72.32	67.66
Total Processing time	2.81	8.26	15.85	27.98	39.09	56.40	70.84	90.86	112.02	118.33
Keypoints Found	44	120	219	381	506	707	853	1079	1313	1216

Image 6

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.02	0.03	0.05	0.08	0.11	0.15	0.18	0.23	0.29
DoG pyramids	0.02	0.03	0.06	0.09	0.13	0.17	0.23	0.29	0.35	0.44
Locating Keypoints	0.35	1.48	3.49	6.31	10.28	15.65	21.45	28.51	37.37	47.93
Gradient	-	0.02	0.03	0.06	0.09	0.13	0.17	0.22	0.28	0.35
Orientation assignment	0.03	0.12	0.23	0.32	0.50	0.70	0.88	1.00	1.19	1.22
Features Descriptor	1.41	5.12	9.75	14.75	22.47	29.52	40.12	46.07	54.35	55.34
Total Processing time	1.81	6.79	13.59	21.58	33.55	46.28	63.00	76.27	93.77	105.57
Keypoints Found	26	95	181	273	408	535	722	834	990	984

Image 7

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.01	0.03	0.05	0.08	0.11	0.14	0.19	0.23	0.29
DoG pyramids	0.02	0.03	0.06	0.09	0.14	0.17	0.23	0.29	0.36	0.46
Locating Keypoints	0.41	1.66	3.80	6.82	11.09	16.43	22.52	30.21	38.25	48.93
Gradient	-	0.02	0.03	0.06	0.09	0.13	0.18	0.23	0.29	0.38
Orientation assignment	0.05	0.16	0.30	0.53	0.82	1.20	1.49	1.75	2.12	2.25
Features Descriptor	2.06	6.79	12.44	23.72	38.96	53.77	68.87	79.93	95.14	90.63
Total Processing time	2.54	8.67	16.66	31.27	51.18	71.81	93.43	112.60	136.39	142.94
Keypoints Found	38	126	231	433	693	972	1244	1441	1721	1616

Image 8

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.01	0.03	0.05	0.07	0.11	0.14	0.19	0.24	0.30
DoG pyramids	0.02	0.03	0.06	0.09	0.12	0.17	0.23	0.29	0.36	0.45
Locating Keypoints	0.38	1.53	3.50	6.33	10.33	15.32	21.19	28.11	36.21	46.81
Gradient	-	0.01	0.03	0.06	0.09	0.13	0.17	0.22	0.28	0.35
Orientation assignment	0.04	0.12	0.19	0.27	0.44	0.61	0.79	0.94	1.15	1.15
Features Descriptor	1.68	4.99	7.52	12.25	20.28	28.40	35.99	44.96	52.85	53.74
Total Processing time	2.12	6.69	11.33	19.05	31.33	44.74	58.51	74.71	91.09	102.80
Keypoints Found	31	89	138	227	371	511	650	811	958	971

Image 9

Resolution	71x58	141x116	212x173	282x231	352x288	423x346	493x404	564x461	634x519	704x576
Preprocessing	-	0.01	0.03	0.05	0.08	0.11	0.14	0.18	0.23	0.29
DoG pyramids	0.02	0.03	0.07	0.09	0.17	0.17	0.23	0.29	0.35	0.45
Locating Keypoints	0.37	1.61	3.78	6.30	10.23	15.33	21.27	28.46	36.81	47.41
Gradient	-	0.02	0.03	0.06	0.09	0.13	0.18	0.22	0.28	0.35
Orientation assignment	0.04	0.10	0.16	0.20	0.35	0.44	0.62	0.85	0.98	1.05
Features Descriptor	1.74	4.25	6.78	9.38	15.53	19.89	28.66	38.98	46.12	46.86
Total Processing time	2.17	6.02	10.85	16.08	26.45	36.07	51.10	68.98	84.77	96.41
Keypoints Found	32	78	126	173	284	362	519	700	842	840