A Method for Real Time Counting Passersby utilizing Space-time Imagery

Doctoral thesis approved by the Graduate School of Advanced Technology and science for the degree of

Doctor of Engineering
in
Department of System Innovation Engineering

By

Ahmed Ibrahim Elhossany Elmarhomy

March, 2014
The University of Tokushima

ABSTRACT

People counting is a process used to measure the number of people passing by a certain passage, or any marked off area, per unit of time; and it can also measure the direction in which they travel. Manual counting is one of the conventional methods for counting people. People can simply count the number of people passing a confined area by using counter. Although people can count accurately within a short period of time, human labors have limited attention span and reliability when large amount of data has to be analyzed over a long period of time, especially in crowded conditions. Every day, a large number of people move around in all directions in buildings, on roads, railway platforms, and stations, etc. Understanding the number of people and the flow of the persons is useful for efficient promotion of the institution managements and company's sales improvements.

On the other hand, most of the deficiencies of manual counting could be handled through automatic people counting systems. In such systems, counting is performed through many approaches among which are a real-time image processing approach. Where a video camera is used to capture video sequences of crossing people and export them to a software package for being processed and interpreted. Counting people and track objects in a video are an important application of computer vision. Computer vision is the science and application of obtaining models, meaning and control information from visual data. It enables artificial intelligence systems to effectively interact with the real world.

Counting people approaches using fixed cameras in image processing techniques can be separated into two types. The first one is an overhead camera that counts the number of people crossing a pre-determined area. The second is count people based detection and crowd segmentation algorithms. In the overhead camera scenario, many difficulties that arise with traditional side-view surveillance systems are rarely present. Overhead views of crowds are more easily segmented compared with a side-view angle camera that can segment as one continuous object.

This thesis proposes a method to automatically count passersby by recording images using virtual, vertical measurement lines. The process of recognizing a passerby is performed using an image sequence obtained from a USB camera placed in a side-view position. While different types of cameras work from three different viewpoints (overhead, front, and side views), the earlier proposed methods were not applicable to the widely installed side-view cameras selected for this work. This new approach uses a side view camera that faced and solved new challenges: (1) two passersby walking in close proximity to each other, at the same time, and in the same direction; (2) two passersby moving simultaneously in opposite directions; (3) a passerby moving in a line followed by another, or more, in quick succession.

This thesis introduces an automated method for counting passerby using virtual-vertical measurement lines. The process of recognizing a passerby is carried out using an image sequence obtained from the USB camera. Space-time images are representing the time as a pixel distance which is used to support the algorithm to achieve the accurate counting. The human regions treated using the passerby segmentation process based on the lookup table and labeling. The shape of the human region in space-time images indicates how many people passed by.

To handle the problem of mismatching, different color space are used to perform the template matching which chose automatically the best matching. The system using different color spaces to perform the template matching, and automatically select the optimal matching accurately counts passersby with an error rate of approximately 3%, lower than earlier proposed methods. The passersby direction are 100% accuracy determined based on the proposed optimal match.

This work uses five characteristics: detecting the position of a person’s head, the center of gravity, the human-pixel area, speed of passerby, and the distance between people. These five characteristics enable accurate counting of passersby. The proposed method does not involve optical flow or other algorithms at this level. Instead, human images are extracted and tracked using background subtraction and time-space images. Moreover, a relation between passerby speed and

the human-pixel area is used to distinguish one or two passersby.

In the experiment, the camera is fixed at the entrance door of the hall in a side viewing position. PC is connected to the camera with a frame rate 17 frame per second. The proposed method was evaluated with 50 cases in each situation. In addition to more than 40 short captured video about 5 minutes series of video sequences of various scenarios and also evaluated with 9 long captured video. Finally, experimental results have verified the effectiveness of the presented method by correctly detecting and successfully counting them in order to direction with accuracy of 97%.

ACKNOWLEDGEMENT

First and most of all, I am very grateful to my Creator Almighty ALLAH, the most Merciful and beneficent, for giving me the strength and patience throughout the course of my study at Tokushima University and all my life.

My deep thanks go to Professor Kenji Terada, my supervisor, for his guidance, his kind support, continuous encouragement and help throughout the accomplishment of these studies. I appreciate his in-exhaustive efforts, unending cooperation and advice, his deep insights that always found solutions when problems supervened and very creative criticism. I will never forget his kindness and patience in dealing with my style and characters, which are very different from Japanese style. I also thank him for introducing me to the image processing fields of Information science and intelligent system. I also wish to thank Dr. Stephen Karungaru, for his guidance, his kind support. My thanks and appreciation goes to all my colleagues and friends at the Department, especially to Dr. Akinori Tsuji for educational utilities, help and friendship.

I am also grateful to my family my beloved Wife Mona Harras, my beloved son Mostafa Elmarhomy, and my lovely daughter Sondos Elmarhomy. I could not complete this work without the unwavering love, patience, prayer, supplication (Doa'a) of them. My deep thanks to all of them for their patience and understanding during the many hours of working and preparing the Information retrieval studies and for the many years of living as foreigners outside our home country Egypt.

I would like to make a deep thank my Parents, my beloved Father Ibrahim Elmarhomy and my beloved Mother Fadyah El-marhomy, for her supported me at all by what I need from her and by her doa'a. I wish to express my deep appreciation to my brother in law Prof. Elsayed Atalm, for his help, his guidance, his kind support and my sisters and my brother, for supporting me all the time, also, all my Muslim friends in Japan and especially in Tokushima for their concern and encouragement.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1. Introduction to People Counting

People counting is a process used to measure the number of people passing by a certain passage, or any marked off area, per unit of time; and it can also measure the direction in which they travel. Every day, a large number of people move around in all directions in buildings, on roads, railway platforms, and stations, etc. the people flow within a confined region could indicate the number of people crossing this region within a specified period of time. There are various reasons for counting people. In retail stores, counting is done as a form of intelligence-gathering. The use of people counting systems in the retail environment is necessary to calculate the conversion rate, i.e., it can help to collect the statistical information on the people flow at different periods of time over certain places. This is the key performance indicator of a store's performance and is superior to traditional methods, which only take into account sales data [1].

Accurate visitor counting is also useful in the process of optimizing staff shifts; Staff requirements are often directly related to density of visitor traffic and services such as cleaning and maintenance are typically done when traffic is at its lowest. More advanced People Counting technology can also be used for queue management and customer tracking.

The People Counting System is a real time people traffic measurement and analysis system for business intelligence solutions. The people counting system provides precise data on people entry and exit activities that allows users to make strategic decisions necessary to improve business performance. It helps managers to understand factors affecting human traffic and thus plan and optimize resources effectively. These factors may include new tenants, special

promotional activities, market research, and customer advertising campaign, new competition and renovation. Shopping mall marketing professionals rely on visitor statistics to measure their marketing. Often, shopping mall owners measure marketing effectiveness with sales, and also use visitor statistics to scientifically measure marketing effectiveness.

Recently, real-time people flow estimation can be very useful information for several applications like security or people management such as pedestrian traffic management or tourists flow estimation. To analyze store performance correctly, people counting must be accurate. It is a ‘false economy’ to select a people counting system on the basis of cost alone. As management consultant Peter Drucker once said: “If you can’t measure it, you can’t manage it.” [2]. Moreover, many violent crimes have increased and become serious problems for many institutions and commercial areas [3]. Many of such measurements are still carried out on manually [4]. The use of video cameras to track and count peoples increased considerably in the past few years due to the advancement of image processing algorithms and computers' technology. Furthermore, tracking and counting people movements are important for the office security or the marketing research. Many of such measurements are still carried out on manual works of persons [2].

1.1.1. Manual Counting

Manual counting is one of the conventional methods for counting people. People can simply count the number of people passing a confined area by using counter. Although people can count accurately within a short period of time, manual counting is labor-intensive and very expensive [5]. Human labors have limited attention span and reliability when large amount of data has to be analyzed over a long period of time, especially in crowded conditions. It is also difficult to provide manual results in real-time for on-line surveillance. Therefore, it is necessary to develop the automatic method of counting the passing people and this is not a simple task, there are some situations difficult to solve even with today's technology.

1.1.2. Automated Counting

Counting people is a challenging scientific problem and has many practical applications such as monitoring the number of people sitting in front of a television set, counting people in the elevator, railway stations and trains, counting the number of people passing through security doors in malls and counting the number of people working in the laboratory. One of the automated methods of counting people is the Turnstiles. Turnstiles are also commonly used for counting the ridership of trains, but the mechanical contacts of turnstiles are inconvenient and uncomfortable to the people. The installations of the gates slow down the normal flow of the people when people flow. Also ultrasonic sensors can be used to count people, ultrasonic receivers can count the number of the people when it detects the echo bouncing off from the people within the detection zone. The accuracy of counting degraded when many objects walk across the detection region, especially person in front of the sensors blocked the detection of other people. Also microwave sensors and weight-sensitive sensors are one of the devices that can be used to count people. In fact, thanks to the fast evolution of computing, it is possible to count people using computer-vision even if the process is extremely costly in terms of computing operations and resources. In general, counting people is important in surveillance based applications, market research and people management. People detection by means of artificial vision algorithms is an active research field. Three main research lines can be noted according to the distance of capturing people, thus limiting the number of people given in a captured image [6].

1.2. Computer Vision and Application

Computer vision is the science and application of obtaining models, meaning and control information from visual data. It enables artificial intelligence systems to effectively interact with the real world. The advance of technology has made video acquisition devices better and less costly, thereby increasing the number of applications that can effectively utilize digital video such as:

- Traffic Management: To extract statistics about the traffic

information from the cameras and automatically direct traffic flow based on the statistics [7]

• Automobile driver assistance: In lane departure warning systems for trucks and cars that monitor position on the road [8].

• Sports analysis: To track sports action for enhanced broadcasts and also to provide real-time graphics augmentation [9].

• Retail video mining: To track shoppers in retail stores and determining their trajectories for optimal product placement [10].

• Games and gesture recognition: To track human gestures for playing games or interacting with computers.

• Automated Video Annotation and Retrieval: To automatically annotate temporal sequences with object tags and index them for efficient content-based retrieval of videos [11]

An important application of computer vision is to track objects in a video. The primary aim of tracking is to establish the location and trajectory of the object movement in a captured video. The tracking can be online, during the video capture or offline, in the post processing stage of video analysis. Counting people is the important task of computer vision. With recent advances of computer technology automated visual surveillance has become a popular area for research and development. Surveillance cameras are installed in many public areas to improve safety, and computer-based image processing is a promising means to handle the vast amount of image data generated by large networks of cameras. A number of algorithms to track people in camera images can be found in the literature, yet so far little research has gone into building real time visual surveillance systems that integrate such a people tracker. The task of an integrated surveillance system is to warn an operator when it detects events which may require human intervention, for example to avoid possible accidents or vandalism. These warnings can only be reliable if the system can detect and understand human behavior, and for this it must locate and track people reliably.

1.3. Computer Vision Based People Counting

Computer vision based people counting offers an alternative to these other methods. The first and common problem of all computer-vision systems is, to separate people from a background scene, determine the foreground and the background. Many methods are proposed to resolve this problem. Several suggest counting people systems use multiple cameras to help with this process.

A sophisticated non-vision based system is presented by Li et al. in [12] . They collect input given by a photoelectric sensor and classify it using a BP neural network. Their system shows good results with a counting accuracy of up to 95%, but suffers from the typical difficulties like people walking in a row. The system of Laurent et al. [13] focuses on counting people in transport vehicles. They identify people by skin blob ellipses of their head, which they obtain by skin-color segmentation. Then these skin blobs are used for tracking and counted after they pass a counting line in the image. The counting accuracy of the system is about 85%. Problems are false counts from hands and the need for a robust skin-color model.

In [14], Septian et al. developed a system that counts people using an overhead mounted camera, which looks straight down to the floor. This prevents the problems arising from occluded persons. They segment and track foreground blobs. In their experiments they show a counting accuracy of 100%, but the database is very small. Problems with this approach are e.g., that it does not sophistically detect humans and the inability to use it in a normal security camera setup due to the untypical camera angle. Also, the overhead mounted camera cannot cover large areas, which is normally an advantage of vision-based systems.

Zhao et al. [15] track persons' faces using data coming from a face detector. The faces are tracked by a scale-invariant Kalman Filter. Instead of relying on a counting line or area like other approaches, they count people by classifying their trajectories. That way, a counting accuracy of 93% can be achieved. Drawbacks are the need of newly training this classifier for each camera and the need of persons moving towards the camera. An advantage is that by using a face

detector, they are among the first approaches which are able to reliably differentiate humans and non-humans.

In [16] the authors present a tracking based approach to segment moving objects in densely crowded videos. Their approach shows encouraging results for videos with up to 50 people. Briefly, they track a large number of low level features. These features are clustered considering the spatial distribution of the features over time. Afterwards, the clusters are counted, denoting the result of the segmentation. Because this approach tracks many low level features, it is resistant to some features getting lost. Features might get lost due to occlusions or unpredicted movement. That means, the approach neither depends on a sophisticated model to deal with occlusions nor requires a complex motion model. Due to its general formulation, it is applicable to many different scenarios. A drawback is, that because of the lack of an elaborate model characterizing human, it is likely that not only humans are counted by the program. Furthermore, the system can only estimate the number of people in the videos but not give any additional information about how many people went where. The system moreover cannot detect persons who are fully occluded.

1.4. Video Surveillance

Video surveillance systems are very important in our daily life. Video surveillance applications exist in airports, banks, offices and even our homes to keep us secure. Video surveillance systems currently are undergoing a transition where more and more traditional analog solutions are being replaced by digital ones. Compared with the traditional analog video surveillance system, a digital video surveillance offers much better flexibility in video content processing and transmission. At the same time, it, also, can easily implement advanced features such as motion detection, facial recognition and object tracking.

Video surveillance has been a popular security tool for years. And thanks to new breakthroughs in technology, security cameras are more effective than ever before. Banks, retail stores, and countless other end-users depend on the protection provided by video surveillance. Fortunately, advances in digital technology have made

video surveillance systems far more cost-effective, flexible, and simple to operate. Security systems using IP (Internet Protocol) cameras are easy to install and maintain, and can be customized and scaled to perfectly match your specific needs. Some surveillance applications strong points:

- Have a video security blanket across the district;
- Be able to monitor activities in any part of the district in real time and automatic alarms for suspicious events;
- Be able to playback the past events after the occurrence and to locate or trace vehicles, objects or people;
- Proactive actions (Preventive Surveillance);
- Verification (recording) of events;
- Additional tool against crime;
- Low operational cost and cost effectiveness for escalating system;
- Flexibility to escalate system to unlimited number of cameras.

1.4.1. Huge Amount of Information in Video Surveillance

Providing for the security of citizens at home and abroad have become top priorities for many other nations around the globe. To this aim, a huge amount of information needs to be captured, processed, interpreted and analyzed. Various computer based technologies can provide a great help in addressing these challenges. Traditional Close Circuits TeleVision (CCTV) networks are a well established off the shelf product with well defined specifications and a mature market [17][18]. However, this kind of surveillance brings with it the problem of managing the large volume of information that can be generated by such a network of cameras. The video streams are transmitted to a central location, displayed on one or several video monitors and recorded. Security personnel observe the video to determine if there is ongoing activity that warrants a response. Given that such events may occur infrequently, detection of salient events requires focused observation by the user for extended periods of time. Commercially available video surveillance systems attempt to reduce the burden on the user by employing video motion detectors to detect changes in a

given scene [19]. Video motion detectors can be programmed to signal alarms for a variety of reasonably complex situations, but the false alarm rate for most systems in typical environments is unacceptable yet.

Ideally, a video surveillance system should only require the user to specify the objectives of the surveillance mission and the context necessary to interpret the video in a simple, intuitive manner. For many scenarios real-time interpretation is required for the information produced by the system to be valuable. Therefore the challenge is to provide robust real-time video surveillance systems that are easy to use and are composed of inexpensive, commercial off-the-shelf hardware for sensing and computation. Given the capability to interpret activity in video streams in real-time, the utility of a video surveillance system increases dramatically and extends to a larger spectrum of missions. With such a system, a single user can observe the environment using a much larger collection of sensors. In addition, continuous, focused observation of activity for extended periods of time becomes possible. As such capabilities mature, the roles of video surveillance systems will encompass activities such as peace treaty verification, border monitoring, surveillance of facilities in denied areas, hazard detection in industrial facilities and automated home security.

1.4.2.Domestic Applications in Video Surveillance

In-house safety is a key problem because deaths or injuries for domestic incidents grow each year. This is particularly true for people with limited autonomy, such as visually impaired, elderly or disabled. Currently, most of these people are aided by either a one-to-one assistance with an operator or an active alarming system in which a button is pressed by the person in case of an emergency. Also, the continuous monitoring of the state of the person by an operator is provided by using tele viewing by means of video surveillance or monitoring worn sensors. However, these solutions are still not fully satisfactory, since they are both too expensive or require an explicit action by the user that is not always possible in emergency situations. Furthermore, in order to allow ubiquitous coverage of the persons movements, indoor environments often require a distributed setup of

sensors, increasing costs and/or required level of attention (since, for example, the operator must look at different monitors). To improve efficiency and reduce costs, on the one hand the hardware used must be as cheap as possible and, on the other hand, the system should ideally be fully automated to avoid both the use of human operators and explicit sensors. Again, standard CCTV surveillance systems are not so widespread in domestic applications since people do not like to be continuously controlled by an operator. Privacy issues and the “big brother” syndrome prevent their capillary distribution, even if the technology is now cheaper (cameras, storage, and so on). Conversely, new solutions fully automated and without the need of a continuous monitoring of human operators, are not invasive and can be acceptable in a home system. Automating the detection of significant events by means of cameras requires the use of computer vision techniques able to extract objects from the scene, characterize them and their behavior, and detect the occurrence of significant events. Peoples safety at home can be monitored by computer vision systems that, using a single static camera for each room, detect human presence, track peoples motion, interpret behavior, assess dangerous situations completely automatically and allow efficient on-demand remote connection. Since the features required by these systems are always evolving, general purpose techniques are preferable to ad-hoc solutions.

1.4.3.People Tracking in Video Surveillance

Review the majority of relevant work directly in the correspondent chapter. In this section, we will focus on famous generic video surveillance systems proposed in the literature only. Some noteworthy prototypes of CV-based people tracking systems have been developed in the last decade, especially in the U.S.A., and funded by DARPA (Defence Advanced Research Projects Agency) programs. One of the pioneering systems of people tracking is Pfinder (“Person Finder”) [20], developed at MIT Media Labs, that employs the Maximum A Posteriori (MAP) probability models to detect human body in 2D image planes, especially in indoor scene. The famous W4 (“What, Where, When, Who”) system developed at University of Maryland, is able to detect multiple people in the outdoors and to analyze body silhouette for inferring people’s activity [21]. VSAM (Video Surveillance And Monitoring), developed at Carnegie Mellon

University, was a big project of cooperative multi-camera surveillance applied in the University campus [22]. Similar research has been carried out in private US research Labs: at IBM, the group of People Vision Project [23] proposed new solutions for appearance-based tracking, also in cluttered indoor environments; at the Siemens labs, in the Imaging and Visualization Department [24] the first formulation of tracking based on mean-shift techniques was defined, in order to follow also body parts in crowded environments. In Europe, since 1996, the group of Prof. Blake at Oxford university proposed Condensation (Conditional Density Propagation) [25] approach to track moving objects also from moving cameras.

Many European projects were funded for video surveillance which includes Advisor and Avitrack. At the ImageLab Laboratory in Italy the Sakbot (Statistical And Knowledge-Based Object Tracker) system [17] has been developed to detect and track people and vehicles using an approach which is robust to occlusions and shadows. It has been used in projects in collaboration with University of California at San Diego [26] for security and with European companies in the area of intelligent transportation systems [27]. Nowadays, many consolidated techniques have been tested for tracking single people from single fixed cameras, and possibly extracting body information, if the camera is placed in an adequate position to have enough resolution for the body shape. Therefore, many researches worldwide are now focusing on distributed cameras and multi-modal acquisition, such as fixed and moving pan-tilt-zoom (PTZ) cameras. Hu et al. [28] report a good survey of multi-camera surveillance systems. Mubarak Shah at University of South Florida [29] proposed an approach for learning geometrical information for consistent labeling or spatially coherent labeling [30], i.e. to maintain the identification of a person and its trajectory when he/she is moving from the field of view of a camera to the one of another camera, by means of homographic geometrical reconstruction. An improved approach has been defined also by the NPD partner. This approach exploits both homography on the ground plane and epipolar geometry, by using the automatically-extracted feet and the head position, respectively [31]. This allows the tracking and disambiguation of groups of people in areas covered by multiple cameras. The use of active (PTZ) cameras to acquire high-resolution images of portions of

the scene or to follow (and “keep in the scene”) interesting people has been proposed recently in the literature [32]. On this topic, the NPD partner has been on the frontier by being first to propose a system based on a single PTZ camera [33].

1.5. Overhead Camera and Side-view Camera

Solutions using fixed cameras that use standard image processing techniques can be separated into two types. In the first, an overhead camera that contains “virtual gaits” that counts the number of people crossing a pre-determined area is used. Clearly, segmentation of a group of people into individuals is necessary for this purpose [34]. The second type attempts to count pedestrians using people detection and crowd segmentation algorithms. In the overhead camera scenario, many difficulties that arise with traditional side-view surveillance systems are rarely present. For example, overhead views of crowds are more easily segmented, since there is likely space between each person, whereas the same scenario from a side-view angle could be incorrectly segmented as one continuous object. When strictly counting people, some surveillance cameras are placed at bottlenecked entrance points where, at most one person at any given time, is crossing some pre-determined boundary (such as a security checkpoint or an access gate at a subway terminal). A potential drawback is that overhead views are prone to tracking errors across several cameras (unless two cameras are operating in stereo), since human descriptors for overhead views are only reliable for a small number of pedestrians [35], using multiple cameras may further complicate crowd counting. In the cases where over-head surveillance views are not available, side-view cameras must be used to count people, and the multiple problems associated with this view (such as crowd segmentation and occlusion) come into play. In the case of crowd segmentation, some solutions that have been proposed include shape indexing, face detection, skin color, and motion [36],[38].

Most of these methods rely heavily on image quality and frame rate for accurate results. Shape indexing and skin colors are considered robust to poor video quality, while motion and face detection are most dependent on video quality. Occlusion is another problem, since all or part of a person may be hidden from view. Some

techniques try to mitigate this issue by detecting only heads [37] or omega-shaped regions formed by heads and shoulders [38].

1.6. Proposed new approach using Side-view Camera

This thesis proposes a method to automatically count passersby by recording images using virtual, vertical measurement lines. The process of recognizing a passerby is performed using an image sequence obtained from a USB camera placed in a side-view position. While different types of cameras work from three different viewpoints (overhead, front, and side views), the earlier proposed methods were not applicable to the widely installed side-view cameras selected for this work.

This new approach uses a side view camera that faced and solved new challenges:

1. two passersby walking in close proximity to each other, at the same time, and in the same direction as shown in Figure 1(a);
2. two passersby moving simultaneously in opposite directions as shown in Fig. 1(b);
3. a passerby moving in a line followed by another, or more, in quick succession as shown in Fig. 1(c).

(a) (b) (c)

Fig. 1 : Examples of passersby image with a side view camera

In this study, space-time image is representing the time as a pixel distance which is used to support the algorithm to achieve the accurate counting. The human regions treated using the passerby segmentation process. The system is fixing automatically to select the

best matching which determine passerby direction and speed. In the experiment, the camera is installed on the left side of the room near the entrance. The experimental results verify the effectiveness of the presented method. In other words, every person was detected by the proposed method and the passerby record image is correctly generated. Moreover, an additional and significant result was that the number of people passing the camera was successfully determined and counted.

1.7. Organization of the Thesis

This thesis is contents five chapters, in this chapter, an introduction of the people counting process and introduction of computer vision and video surveillance. Chapter two is an overview of the related work in counting, detecting and tracking people also contains summary of some earlier approaches. Chapter three is the heart of this thesis which contains the proposed approach of counting people using space-time images divided into subsections and discusses each in details. Chapter four shows the experiments in details and express the results based on the experiments in tables and images. Chapter five is the conclusion and some future work for the development on people counting.

CHAPTER 2

RELATED WORK

Recently, real-time people flow estimation can be very useful information for several applications like security or people management such as pedestrian traffic management or tourists flow estimation. The use of video cameras to track and count peoples increased considerably in the past few years due to the advancement of image processing algorithms and computers' technology. This chapter outlines the main approaches considered in the solution of people tracking and counting. The focus is put on related work in the individual fields of background modeling, people detection, space-time image and tracking as well as combined approaches. During the literature research, it was noted that extensive research were conducted in the fields of people detection and tracking both from static and from mobile platforms. In either set-up static or mobile, the choice and combination of approaches strongly depend on the scenario.

Beside approaches combining vision with auxiliary sensors [39], [40] there are various approaches to people detection and tracking using monocular vision. The static camera set-up, with respect to the scene, allows foreground segmentation by using an inferred background model. Monocular approaches using background modeling typically have problems with sudden illumination changes, ghost effects, partial occlusions, and people remaining in the same position for a long time. There are even several approaches that apply shadow suppression [41], [42] in order to overcome these ghost effects, however struggle with occlusions and precise localization of people in the scene.

In order to better deal with illumination changes, occlusions and allowing more robust and precise localization of people in the scene, people detection and tracking system are increasingly based on disparity maps. Including disparity/ depth information makes it easier to segment an image into objects (distinguishing people from their

shadows) and produces more accurate localization information for tracking people. Hence, the use of stereo cameras for people detection and tracking has been exploited [43], [44], [45], [46] in order to present a relatively inexpensive, compact solution, that is easy to calibrate and set up.

2.1. People Detection and People Tracking

A number of surveillance scenarios require the detection and tracking of people. Although person detection and counting systems are commercially available today, there is need for further research to address the challenges of real world scenarios. Video surveillance systems seek to automatically identify events of interest in a variety of situations. Example applications include intrusion detection, activity monitoring, and pedestrian counting. The capability of extracting moving objects from a video sequence is a fundamental and crucial problem of these vision systems.

2.1.1. People Detection

The key challenges in people detection can be summarized by the following points:

- The appearance of people yields high variability in pose, texture, and size range. In addition, people can carry different objects and wear different clothes.
- People must be identified in public transportation scenarios, which includes a wide range of illumination, fast changing lighting conditions (entering a tunnel, sudden direct sunlight), that decrease the quality of sensed information (poor contrast, shadows).
- Interclass occlusions (people standing close to each other or partially occluding each other) and appearance in the context of cluttered background.
- People must be detected in highly dynamic scenes, since motion is generally very high shortly before approaching and during stops.
- Appearance at different viewing angles resulting in perspective distortion.

Some of the above key challenges can be overcome by foreground segmentation. Foreground segmentation helps to neglect image regions corresponding to background and hence only feeding a list of ROIs (Region of Interest) to the detection module that are likely to contain people. A large scale survey on pedestrian detection carried out by D. Geronimo et al. [47] yields a comprehensive comparison and points out key differences between wide ranges of people detection approaches.

Investigating several people detection approaches can be classier in two groups. The first group includes general image based people detection methods that classify the ROIs as person or nonperson with the aim to minimize the number of false positive and false negative detections. The second group includes detection methods that do not classify foreground ROIs as person or non-person; however, yield valid people detection results under the assumption that foreground objects are people. As proposed by D. Geronimo et al. [47] the image-based people detection methods for person or non-person classification can be broadly divided into silhouette matching and appearance.

Silhouette matching. B. Wu and R. Nevatia [48] propose a novel approach to detect partially occluded people by matching small chamfer segments and combining the results with a probabilistic voting scheme. A disparity template matching method was proposed by D. Beymer and K. Konolige [44] that downscales foreground disparities according to their dominant disparity value and matches small binary person templates.

Appearance. These methods define a space of image features (also known as descriptors), and a classifier is trained by using ROIs, known to contain examples (people) and counterexamples (non-people). Another established human classification scheme is Histograms of Oriented Gradients (HOGs) by N. Dalal and B. Triggs [49]. HOGs utilize features similar to SIFT [50] features, where gradients of an image region are computed and are further divided into orientation bins. The authors propose to use a linear Support Vector Machine (SVM) learning method.

The second groups, including non-classification detection methods, yield valid people detection results under the assumption that foreground objects are people. T. Darrell et al. [46] propose plan-view analysis, which consists of projecting foreground points onto a ground plane and of grouping the projected points according to model assumptions. S. Bahadori et al. [43] apply this plan-view analysis in their approach to refine foreground segmentation and compute observations for tracking.

2.1.2. People Tracking

As pointed out above, people tracking using stereo cameras has been exploited [43], [44], [45], [46] in many applications. The purpose of tracking is to avoid false detections over time and making useful inferences about people's behavior (e.g. walking direction). In order to track people over subsequent input frames in the presence of temporary occlusions, it is necessary to have a model of the tracked people. Several models for people tracking have been developed [51], [52], including color histograms and color templates, in addition to 3D size restrictions. Most of the above applications apply a Kalman filter [53] framework using constant velocity models. For tracking multiple objects in 3D (using cues on silhouette, texture, and stereo), Giebel et al. [54] use particle filtering. Particle filters are widely used in tracking [55], [56]. Both types of filters, however, rely on a first order Markov assumption, and hence, carry the danger of drifting away from the correct target. Research on detection in crowded scenes has led to coupled detection-tracking approaches, which share information between both stages instead of treating the information sequentially [57].

2.2. Surveillance Applications

In this context of video surveillance, most of the emphasis is devoted to techniques capable of execution in real-time on standard computing platforms and with low cost off-the-shelf cameras. Additionally, in indoor surveillance of people’s behavior the techniques must cope with problems of robustness and reliability: for instance, in videos acquired with a fixed camera, the visual appearance of a person is often cluttered and overlapped with home

furniture, other people, and so on. We will introduce some of the video surveillance application in the next subsections.

2.2.1.Human Behavior Analysis

Emerging technologies can offer a very interesting contribution in improving the quality of life of people staying at home or working indoors. Most of these techniques and the related systems are converging in the new discipline of Ambient Intelligence that includes ubiquitous computer systems, intelligent sensor fusion, remote control, tele-healthcare, video surveillance and many other pervasive infrastructure components. One important goal of these systems is human behavior analysis, especially for safety purposes: non invasive techniques, such as those based on processing videos acquired with distributed camera, enable us to extract knowledge about the presence and the behavior of people in a given environment. Recent research in computer vision on people surveillance jointly with research in efficient remote multimedia access makes feasible a complex framework where people in the home can be monitored in their daily activities in a fully automatic way, therefore in total agreement with privacy policies.

There are many reference surveys in the field of human motion capture (HMC) and Human behavior analysis (HBA), for instance, the ones by Cedras and Shah [58], Gavrila [58], Aggarwal and Cai [59], and Moeslund and Granum [60], or more recently, Wang et al. [61]. The basic aim of these models and algorithms is to extract suitable features of the motion and the visual appearance of people (e.g., shape, edges, or texture), in order to classify and recognize their behavior. Many works employ a precise reconstruction of the body model in order to detect the motion of each part of the body. This is normally done for virtual reality and computer graphics application with people moving in structured environments [62].

2.2.2.Classify Trajectory Based Human Behavior

In surveillance, people are normally not collaborative, they are moving in cluttered scenes interacting each other’s and with objects, people carrying packs or sitting on a bench, and the acquisition is

usually done with large-FOV cameras (resulting in images acquired with low resolution). In such cases, useful features that can be analyzed are the people's trajectory and the changes in motion status and direction. For this reason, there is a growing research activity in trajectory analysis. For instance, in [63] classification of vehicle trajectory is done in order to extract abnormal trajectories. In [64] Abstract Hidden Markov Models are exploited to recognize special trajectories and monitor special behaviors in indoor areas. In [65] trajectories of people walking in outdoor are represented by graphs, and trajectory comparison is done by means of graph matching. Some works as [66] deal with single interaction between pairs of trajectories and typically refer to very simple interactions (such as the "follow", or "approach-talkcontinue together") or divergence to typical paths. All these techniques are employed to classify a given trajectory as being significant (abnormal) or normal and, based on that, describe the people's behavior. Richer information can be extracted from persons' appearance, posture and gait.

In recognizing behavior based on a person's shape there are two main approaches. The static one is concerned with spatial data, one frame at a time, and compares pre-stored information (such as templates) with the current image. The goal of static recognition is mainly to recognize various postures, e.g., pointing [67],[68], standing and sitting [69]. The second type of approaches is dynamic recognition where here temporal characteristics of moving target are used to represent its behavior. Typically, simple activities such as walking are used as the test scenarios. Both low and high level information is used. Low-level recognition is based on spatio-temporal data without much processing, for instance, spatio-temporal templates [68],[70] and motion templates [71]. High level recognition are based on pose estimated data and include silhouette matching [72], HMMs [73],[74] and neural networks [75]. In the work of [73] the idea of representing motion data by "movements" (similar to phonemes in speech recognition) is suggested. This enables to compose a complex activity ("word") out of a simple series of movements ("phonemes"). An HMM is used to classify three different categories: running, walking and skipping. This type of high level symbolic representation is also used in [68] who automatically build a "behavior alphabet" (a behavior is similar to a movement) and model

each behavior using an HMM. The alphabet is used to classify different types of actions in a simple virtual reality game and to distinguish between the playing style of different subjects. Another successful application of this symbolic approach is in recognizing signed-language [74].

2.2.3. Tracking Human Body in 3D Model

Recently an increasing number of computer vision projects deal with detection and tracking of human posture as well. An exhaustive review of proposals addressing this field was written by Moeslund and Granum in [60], where about 130 papers are summarized and classified according with several taxonomies. The posture classification systems proposed in the past can be differentiated by the more or less extensive use of a 2D or 3D model of the human body [60]. In accordance with this, we can classify most of them into two basic approaches to the problem. From one side, some systems (like Pfinder [20] or W4 [76]) use a direct approach and base the analysis on a detailed human body model: an effective example is the Cardboard Model [77]. In many of these cases, an incremental predict-update method is used, retrieving information from every body part. Many systems use complex 3D models, and require special and expensive equipment, such as 3D trinocular systems [78], 3D laser scanners [79], thermal cameras [80], or multiple video cameras to extract 3D voxels [81]. Due to the need for real time performance and low cost systems, we discarded complex and/or 3D expensive solutions. In addition, these are often too constrained to the human body model, resulting in unreliable behaviors in the case of occlusions and perspective distortion, that are very common in cluttered, relatively small, environments like a room.

A second way consists in an indirect approach that, whenever the monitoring of single body parts is not necessary, exploits less, but more robust, information about the body. Most of them extract a minimal set of low level features exploited in more or less sophisticated classifiers. One frequent example is the use of neural networks, as in [82],[83]. However, the use of NN presents several drawbacks due to scale dependency and unreliability in the case of occlusions. Another interesting example of this class is the analysis of

AC-DCT coefficients in the MPEG compressed domain [84]: this has proven to be also insensitive to illumination changes, but the reported examples only classify different standing postures (with different pointing gestures), while we are interested in classifying very different postures, such as standing up and laying on the floor. Eventually, in [85], a Universal EigenSpace approach is proposed: this presents insensitivity to clothing, but it assumes that most of the possible postures (with most of the possible occlusions) have been learned, and this is far from being realizable.

Another large class of approaches are based on human silhouette analysis. The work of Fujiyoshi et. Al. [86] uses a synthetic representation (Star Skeleton) composed by outmost boundary points. A similar approach is proposed in [87] where a skeleton is extracted from the blob by means of morphological operations and then processed using a HMM framework. This approach is very promising and has the unique characteristic of also classifying the motion type, but it is very sensitive to segmentation errors and in particular to occlusions. Moreover, no scaling algorithm to remove perspective distortion is proposed making this approach unfeasible for our target application.

Another approach based on silhouette analysis is reported in [88],[89] where a 2D complex model of the human body is matched with the current silhouette by genetic algorithms. In addition to the problems of segmentation errors and occlusions, this approach also suffers from dependency of the model on the view. In [76], Haritaoglu et al. add to W4 framework some techniques for human body analysis using only information about the silhouette and its boundary. They first use hierarchical classification in main and secondary postures, processing vertical and horizontal projection histograms from the body's silhouette. Then, they locate body parts on the silhouette boundary's corners.

2.2.4.Detecting and Counting People in Surveillance Applications

In [90], foreground segmentation is supposed to be influenced only by noise or light. It brings up a particular background model to

fix the problem. Yuk et al. [91] develop the head contour detection process to detect the object which has head contours and painted the trajectories. Elgammal et al. [92] present a probabilistic framework for tracking regions based on their appearance. Buzan et al. [93] propose a system that tracks moving objects in a video dataset so as to extract a representation of the objects' 3D trajectories. In [94], Porikli et al. propose the representation of trajectory using hidden Markov model. Liu et al. [95] consider that the person detection and counting systems are commercially available today, and the focus of their work is the segmentation of groups of people into individuals. In [96], Zhao et al. present a real time system to count human beings passing through a doorway. They use the characteristic of LAB color space and local features of moving targets to track the human accurately. Rittscher et al. [97] propose an algorithm based on partitioning a given set of image features using a likelihood function that is parameterized on the shape and location of potential individuals in the scene.

Terada et al. creates a system that can determine people direction movement and so count people as they cross a virtual line [98]. The advantages of this method is it avoids the problem of occlusion when groups of people pass through the camera's field of view. To determine the direction of people, a space-time image is used. Like Terada et al, Beymer and Konolige also use stereo-vision in people tracking [99]. Their system uses continuous tracking and detection to handle people occlusion. Template based tracking is able to drop detection of people as they become occluded, eliminating false positives in tracking. Using multiple cameras improve the resolution of occlusion problem. But the problem is the need to have a good calibration of two cameras (when 3D reconstruction is used).

Hashimoto et al. resolve the problem of people counting using a specialized imaging system designed by themselves (using IR sensitive ceramics, mechanical chopping parts and IR-transparent lenses) [100]. The system uses background subtraction to create "thermal" images (place more or less importance depending of the regions of the image; Region of Interest) that are analyzed in a second time. They developed an array based system that could count persons at a rate of 95%. So their system is extremely accurate but with certain conditions. In order to work in good conditions, the system requires a

distance of at least 10 cm between passing people to distinguish them and thus to count them as two separate persons. The system also shows some problem in counting with large movements from arms and legs. So this system will be not so appropriate in commercial centre because of the high density traffic when people entering or exiting. In fact, most of the time, person come in supermarkets with their family so make a close group of people which is the most difficult problem to resolve for counting people system.

Another method of separation of people from a background image is used by Schofield et al. [101]. The entire background segmentation algorithm is done by simulating a neural networks5 and uses a dynamically adjusted spacing algorithm in order to solve occlusions. But because of the reduce speed of neural network, the algorithm only deal with counting people in a specific image. This paper is just an approach of how resolve people counting by using neural networks. Tracking people is not considered.

As simple and effective approach, Sexton et al. use simplified segmentation algorithm [102]. They test their system in a Parisian railway station and get error rate ranging 1% to 20%. Their system uses a background subtraction to isolate people from the background. The background image (reference frame) is constantly updated to improve the segmentation and reduce the effect of lighting or environment modification. The tracking algorithm is simply done by matching the resulting blobs, given by the background subtraction process, with the closest centroids6. Means that the tracking operation is operated frame to frame and the label of the blob resulting with the current frame is the same that the blob resulting with the previous frames which has the closest centroid. In order to avoid the majority of occlusions, an overhead video camera is used.

2.3. Methods Based Background Modeling

Different methods for background modeling and updating have been proposed. Depending on the approach to background modeling and model updating, we can divide into two classification directions. The first classification direction includes approaches on how to model the background using depth information. Statistical models have been

widely used, either in the form of single Gaussians [103], [104] or mixture of Gaussians [105], [106]. The choice of the model mostly depends on the kind of scenario in which the application runs. Single Gaussian models are not adequate in environments with actuating background or scenarios that do not allow background inference of the empty scene. In addition to the purely depth/ disparity-based background modeling methods, many combined approaches are proposed including intensity, edge, or motion information [107], [43]. These combined approaches have been proven to be effective.

Secondly, we look at the approaches to correctly update the background model. In our scenario of having a purely static background inside trains, adaptively of the background model does not need to be considered. By purely static background, we understand that all furniture and interior objects are mounted in a fixed position and are not displaced over time. However it is noteworthy that in general background modeling approaches it is necessary to implement a method that is adaptive to dynamic changes in the background model. A basic adaptive model can be achieved by maintaining the status of the background (mean and variance at each pixel for single Gaussian models) and updating this status with the current observation [104]. Kalman filtering is used [108], [109] to achieve adaptively by tracking samples over time for each pixel. In any case, a trade-off between false positives due to foreground objects integrated in the background model and the reactivity of the model to adapt to background changes must always be considered.

2.4. Counting using crowd Segmentation Technique

Liu et al. present in [110] a tracking based approach to count people. They use a combination of foreground detection, crowd segmentation and tracking. The persons are counted when they cross a “virtual” gate. The advantage of this system is, that it can really detect where people are going as opposed to just deciding how many persons are present. The algorithm presented is not only applicable to human tracking, but can be extended to other classes as well. The downside of the approach is that it heavily depends on foreground detection and does not detect humans by a sophisticated model, but by foreground blobs of human proportions. There are some problems caused by

foreground segmentation. First of all, shadows are often also detected as foreground. Second, a person can be split in several not connected blobs. Multiple persons in close proximity form one big foreground cluster. The system tries to solve this problem by using a crowd segmentation technique. This works, when the cluster is composed by only few persons. But when strong occlusions are present or the crowd gets large it fails, since it uses cues based on edges of the foreground map.

The two most important steps of the system are the tracking and the model based segmentation, described in the following sections. The tracker follows the persons and uses the segmentation algorithm in case the persons clutter. Virtual gates are defined in the picture. These virtual gates can for example be lines, but also more complex geometric forms. When a person crosses such a virtual gate, it is counted.

The Tracker The tracker uses an adaptive appearance based approach. Here a color model and a probability for every pixel in the model to belong to the foreground are adaptively trained. This information is used to refine the information provided by the foreground segmentation to make a measurement of the current location of the person. If a measurement is in close proximity to the prediction, it is set as the new location of the person. To overcome the negative effects clutter has on the tracking, large foreground regions and regions with a close proximity of persons are forwarded to the segmentation process.

The Model Based Segmentation Features are extracted based on the foreground segmentation. These features are used to detect cliques which represent a person. Now a large set of hypothetical persons is available. A combination is searched that assigns each feature to at most one person. To find the combination with the biggest likelihood, the EM algorithm is employed.

2.5. Real Time People Tracking System for Security

Real-time people flow information is very useful source for security application as well as people management such as pedestrian

traffic management, tourist flows estimation. To track and count moving people is considered important for the office security or the marketing research. Many of such measurements are still carried out on manual works of persons. Therefore it is necessary to develop the automatic method of counting the passing people.

Several attempts have been made to track pedestrians. Segen and Pingali [111] introduced a system in which the pedestrian silhouette is extracted and tracked. The system runs in real-time, however, the algorithm is too heavy to track many people simultaneously and cannot deal well with temporary occlusion. Masoud and Papanikolo poulos [112] developed a real-time system in which pedestrians were modeled as rectangular patches with a certain dynamic behavior. The system had robustness under partial or full occlusions of pedestrians by estimating pedestrian parameters. Rossi and Bozzoli [113] avoided the occlusion problem by mounting the camera vertically in their system in order to track and count passing people in a corridor, but assumed that people enter the scene along only two directions (top and bottom side of the image). Terada [4] proposed a counting method which segmented the human region and road region by using the three dimensional data obtained from a stereo camera. However, this system also assumed only simple movement of pedestrians.

Segen and Pingali concentrate on image processing after segmentation [114]. A standard background algorithm is used to determine the different regions of interest. Then, in each of those areas, the algorithm identifies and tracks features between frames. All the paths of each feature is stored and represent the motion of person during all the process. Then, by using those paths, the algorithm can easily determine how many people crossed a virtual line and the direction of this crossing. This system does not deal with occlusion problems and can be reduce in performance if there is a lot of persons in the field of the video camera. In fact, the paths' data will be big which will complicate the calculation of intersection between the line and all the paths. Haritaoglu and Flickner adopt an another method to resolve the problem of real time tracking of people [115]. In order to segment silhouettes from the background, they choose to use a background subtraction based with color and intensity of pixel values.

Those information's will help to classify all the pixels in the image. Three classifications are used : foreground, background and shadow. Then all the pixels classified as foreground make different regions. These entire foreground groups are then segmented into individual people by using 2 different motion constraints as temporal and global. In order to track these individuals, the algorithm uses an appearance model based on colour and edge densities.

Matsuyama, Wada, Habe and Tanahashi [116] proposed a real-time people counting system with a single camera for security inside the building. The camera is hung from the ceiling of the gate so that the image data of the passing people are not fully overlapped. The implemented system recognizes people movement along various directions. To track people even when their images are partially overlapped, the proposed system estimates and tracks a bounding box enclosing each person in the tracking region. The approximated convex hull of each individual in the tracking area is obtained to provide more accurate tracking information.

2.6. Multi-target Tracking Systems

When designing a tracking based counting system, you are primary designing a multitarget tracker. The tracker has to be able to track varying number of targets. Yet it suffers from problems arising due to person interaction and occlusion. Person interaction and occlusions are big problems for tracking algorithms, because they imply a strong dependency of the individual targets. These dependencies are too costly to calculate exactly, thus an efficiently calculable solution has to be found. In this section, some systems are presented and their capabilities are described. Two systems of special interest for this diploma thesis are then described in more detail.

In case of multiple, non-labeled measurements, the problem of data association is important. When the targets are close both in space and appearance, it is not an easy task to associate the right measurement to the right tracker. Reids Multiple Hypothesis Tracker (MHT) [117] and the joint probabilistic data association filter (JPDAF) [118],[119] handle this problem. While the MHT can deal with a changing number of targets, the number of targets in the

JPDAF remains fixed. These methods are not usable for the purposes of this thesis, because their runtime increases exponential with the number of targets. The probability hypothesis density (PHD) filter first developed by Mahler in [120] retains the joint nature of the multiple target tracking and models the appearance and disappearance of targets directly in the filter as opposed to a superordinate process. The resulting filter is the multi-target equivalent to a constant-gain Kalman filter. The resulting equations are still not efficiently computable. For that reason, implementations approximate the PHD, for example using particle filters [121]. Although the PHD models the multitarget problem principled and efficient, the problem is coarsely approximated by utilizing the constant-gain Kalman filter. In [122], a probabilistic exclusion principle is presented, which prevents persons from occupying the same space. Additionally, the technique of partitioned sampling is introduced in this work to handle the occlusion problem. Partition sampling decomposes the joint structure of the occlusion problem. If it is for example known that target A occludes target B, first the configuration for target A can be calculated and later used to infer the configuration for target B. The problem with this approach is that it assumes that the spatial distribution of two targets is known. But in practice, one could have two hypotheses for target A, one in the front and one in the back of the picture, while target B stands in between. Now the decomposition as suggested by partitioned sampling is not possible anymore.

In [123] Yu et al. developed a filter using Pairwise Markov Random Fields (PMRFs) to avoid coalescence of different targets. Coalescence means that two trackers lock on the same target. This phenomenon occurs, when two similar looking targets stand very near to each other. To avoid coalescence, PMRFs model pairwise interactions between two targets. This interaction can for example prevent two persons from occupying the same space. That way, coalescence can be prevented.

2.7. Detection and Tracking of Multiple Humans

In [124] Wu et al. present an interesting approach for tracking multiple humans. They use a part-based detector to identify humans. If possible, the new detections are simply matched to similar old

detections. If this fails, a mean shift based tracker [125] is employed to find the person. This system and its predecessor in [126] are the only systems known to the author evaluated on the CAVIAR dataset. The advantage of this system is that it can handle partially occluded humans (both inter-person and inter-scene) by utilizing part-based detectors. Their edgelet feature provides a robust description of human. Another advantage is that by utilizing a state of the art human detector, they can discriminate humans and non-humans. The biggest drawback of the presented approach is the inability to cope with fully occluded persons. The system divides into two parts: first a part-based detection is performed. Then the detection results are used to reliably track humans.

Part-based Detection The part-based detector developed in this work uses edgelets as features. An edgelet is a local shape feature, that should detect the silhouette of a human. Examples are lines and circles. These edgelets are weak classifiers, which are used in a boosting method [127] to train a cascade-of-rejectors. That way, several classifiers are trained, as visualized. For every body part several views are trained. These views share the same root node in the cascade-of-rejectors, making the computation more efficient.

The full body part and the head-shoulder detector are used to find hypotheses for humans in the image. Then the torso and legs detectors are used to scan in the region of these hypotheses. Now combined responses are formed, which “fuse” together part detectors belonging to the same human: the hypotheses for the humans are analyzed to see whether there are other part-detectors supporting the hypotheses or not. To account for occlusions, an occupancy map is built from the hypotheses which marks occluded body parts with do not care. Then a Bayesian approach is chosen to find the best fitting mapping given the image observation. That way, false alarms and false negatives can be filtered out and the so-called “combined responses” from the part-based trackers are built.

Tracking Based on Detections In the first stage of tracking, the detection results are matched to the existing trackers. This is done by defining an affinity measure which regards position, size and color appearance. The persons are matched to the detection with the biggest

affinity. Potential tracks are initialized every time a detection has not been matched to a tracker. The potential tracks become a confident trajectory, if they have been matched to detection for a certain amount of time. The amount of time needed is determined by the affinity of the match and the probability for consecutive false measurements. Track deletion is done in a similar way.

At every time-step, persons are tracked by matching the detections to the existing trajectories. When this fails, a mean shift tracker is used to track the parts individually. The probability distribution tracked by the mean shift tracker is composed of the color based appearance model; a Kalman based dynamic model and the detection confidence. To improve the performance of the probability from the appearance model, principal component analysis (PCA) is used to model shape constraints.

2.8. Counting People using Video Cameras

In the past years there has been a bulk of research work in the area of image processing with the objective of obtaining more accurate and reliable people-count estimations. An intuitive solution to the problem of estimating the size of a crowd in an image will be, literally, to obtain a head count. While this would be a tedious, but yet feasible, task for a human it certainly is a difficult problem for an automatic system. That is exactly the problem tackled as in [128], where wavelets are used to extract head-shaped features from the image. Further processing uses a support vector machine to correctly classify the feature as a “head” or “something else” and applies a perspective transform to account for the distance to the camera. A similar idea is used in [129], where a face detection program is used to determine the person count. Unfortunately, as pointed out by its authors, this method is affected by the angle of view at which the faces are exposed to the camera. Additionally, images where a person’s back is only visible will result in a poor estimation as well. Another approach has been suggested in [130], it aims to obtain an estimation of the crowd density, not the exact number of people. It requires a reference image—where no people are present, in order to determine the foreground pixels in a new image. A single layer neural network is fed with the features extracted from the new image (edge

count and densities of the background and crowd objects) and the hybrid global learning algorithm is used to obtain a refined estimation of the crowd density.

The reconstruction of 3D information from stereo image pairs is one of the key problems in computer vision and image analysis. A variety of algorithms have been proposed for this purpose. These algorithms can be divided into four method categories, depending on their strategy of solving the correspondence problem. The first category includes area-based methods [131], [132], [133], that correlate image patches by comparing local similarity measures. These methods assume constant disparities within the correlation image patch and thus, yield incorrect results at depth discontinuities, which lead to blurred object boundaries. However, these methods allow real-time disparity calculation. The second method category includes feature-based methods [134], [135] that make use of characteristic image features based on corners or edges. The third category includes phase-based methods [136], [137], which estimate displacements via the phase in the Fourier domain. The fourth category includes energy-based techniques [138], [139], [140] that aim to minimize variation formulations. These methods then penalize deviations from data and smoothness constraints and seek to infer a global optimum.

Damian and Valery compare different classification algorithms for estimating the number of people in an image obtained from a video surveillance camera. This approach differs from previous works in that we do not attempt to obtain and count specific features from the images (head shaped objects in [128] or faces in [128]). We just exploit the correlation between the percentage of foreground pixels and the number of people in an image [141].

Tesei et al. use image segmentation and memory to track people and handle occlusions [142]. In order to highlight regions of interests (blobs3), the system uses background subtraction. It consists to subtract a reference frame (background image previously compute) from the current frame and then threshold it (this algorithm will be more detailed in the analysis section). Using features such as blob area, height and width, bounding box area, perimeter, mean gray level,

the blobs are tracked from frame to frame. By memorizing all this features over time, the algorithm cans resolve the problem of merging and separating of blobs that occurs from occlusion. In fact, when blobs merge during occlusion a new blob is created with other features but the idea of this algorithm is that in this new blob, it stores the blobs' features which form it. So when the blobs separate themselves, the algorithm cans assigned their original labels. This system doesn't resolve all the problems but it's a good idea and does not request a lot of computing operations.

Shio and Sklanksy try to improve the background segmentation algorithm (detect people occlusion) by simulating the human vision more particularly the effect of perceptual grouping4 [143]. First, the algorithm calculates an estimation of the motion from consecutive frames (frames differencing is more detailed in the analysis section) and use this data to help the background subtraction algorithm (segment people from the background) and try to determine the boundary between closer persons (when occlusions occurs, make boundaries to separate people by using frame differencing information). This segmentation uses a probabilistic object model which has some information like width, height, direction of motion and a merging/splitting step like this seen before. It was found that using an object model is a good improvement for the segmentation and a possible way to resolve the occlusions problem. But using perceptual grouping is totally ineffective in some situations like, for example, a group of people moving in the same direction at speed almost equals.

2.9. Crowd Modeling for Surveillance

A people tracking system cans be used for others applications not necessarily just for counting people. For example, these systems can be extended for security application. In fact, a real-time people tracking system provides enough information in order to make a good video surveillance. Detect strange behaviors of people (like violent gesture, fight or running people) and store these information’s on a database [144]. This type of system cans be very interesting for storekeeper or supermarket. Another application is for marketing. It can analyze the behaviors of clients and make conclusion. For

example, measure the impact of an advertisement or modification in the arrangement. Determine the period and place of good and bad influence. Moreover, we also hope to find out what behavior-oriented state the crowd is in, i.e. states representing the crowd. This work has potential for applications in different areas of our everyday life. First of all, it can be applied to intelligent surveillance systems in security agencies. In pieces such as banks, airports, public squares and casinos, the crowds should be monitored for detection of abnormal behaviors. Secondly, this research will enhance the intelligent level of transportation systems in areas such as evacuation, crowd guidance, etc. The understanding of crowd distributions and motions will make crowd control easier and more skillful. The schedule of public transportation vehicles such as trains and buses can be optimized to provide larger load. Thirdly, for the tourism agencies, crowd modeling and monitoring technology can be employed and further improved to optimize the controlling and guiding tourists flow for safety, comfort, and protection of resources.

Gary Conrad and Richard Johnsonbaugh simplify the entire people counting process by using an overhead camera (it permits to greatly reduce the problem of occlusions) [145]. To avoid the problem of light modification, they use consecutive frames differencing instead of using background subtraction. To limit computation, their algorithm reduces the working space in a small window of the full scene perpendicular to the flow traffic. At any given time, their algorithm is able to determine the number of people in the window and the direction of travel by using the centre of mass in each little images of the window. With a quick and simple algorithm, they obtained very good results and achieved a 95,6% accuracy rate over 7491 people.

2.10. Counting People without People Models or Tracking in Crowd Monitoring

Chan et al. [147] presents a non tracking based approach to count large numbers of people and to distinguish in which directions they are walking. To achieve this, first the crowd is separated into groups with different motions using the mixture of dynamic textures model [148]. Then for each region a set of simple features is extracted

which are afterwards classified by a Gaussian Process [149]. One advantage of this system is, that it is privacy preserving because no separation of single persons takes place in the classification process. It can deal with a large number of pedestrians, but the downside again is, that no counting with respect to "who went where" can be performed. However, this may be solved by simply tracking the different regions. Tracking regions instead of persons has the advantage that a large number of persons can be summarized in one region, resulting in a lower computational effort. The problem might be that regions can merge and split. Another problem of the presented approach is that both the mixture of dynamic textures model and the Gaussian Process have to be learned.

First, the picture is split into regions of different direction using the mixture of dynamic textures model. This model is learned with the Expectation Maximization (EM) [150] algorithm and is described in detail in [148]. Before the feature extraction can take place, the effects of perspective must be considered. If not, closer objects have more influence because they are bigger. To reduce this effect, a simple approach is chosen to generate an approximate perspective map. The pixels are now weighted according to the distance information provided by this map. Afterwards 28 features are extracted, which can be classified as segment features (e.g., the total number of pixels in the segment), internal edge features (e.g., the total number of edge pixels in the segment) and texture features (e.g., the homogeneity of the texture at different angles).

People have different appearances depending on their walking direction. Thus for each walking direction a separate Gaussian Process has to be trained to map feature output to crowd size. The kernel function for the Gaussian Process is modeled by a linear and a RBF kernel. It was chosen because normally the features should linearly correspond to crowd size, but some nonlinearities arise due to various reasons like occlusion, segmentation errors and spacing within a segment [147].

2.11. Summary of Earlier Approaches for People Counting and Tracking

Accurate people detection can increase management efficiency in public transportation by marking areas with high congestion or signaling areas that need more attention. Estimation of crowds in underground transit systems can be used to give passengers a good estimate of the waiting time in a queue. Multiple solutions to automate the crowd-counting process have been proposed, including solutions from a moving platform (such as a camera on a bus) [151] that analyze the optic flow generated from the moving objects as well as the moving platform. Researchers have identified crowd counting to be often highly-sensitive to training data [152], in which cases algorithms or crowd density classifiers [153] will greatly benefit from having a realistic and robust training dataset. New techniques for creating human crowd scenes are continuously being developed, especially due to the growing demand from the motion picture industry [154]. Simulated crowds have been widely studied in many application domains, including emergency response [155] and large-scale panic situation modeling [156], [157]; perhaps simulated crowds [158] or flow models could also potentially offer visual surveillance researchers a new way to efficiently generate training data. An analysis of some earlier approaches for people counting and tracking can be seen in table 1 and table 2. [O: Outdoor, R: Real time, C: Crowed]

Table 1 analysis of some earlier approaches for people counting and tracking

Author	Yr	Behaviors	dataset	O	R	C	Ref .
Lengvenis	13	passenger counting in public transport	214 passengers , used a video image where 3 different people were boarding the bus one by one	Y	Y	N	[159]
Terada	09	counting passersby, generating passerby record images	an average of 33 persons passed by during the daytime, on line detect using internet camera	N	Y	N	[160]
Xi	09	counting	dataset of more than	Y	N	Y	[15]

		people, face detection	160 potential people trajectories				
Li	08	crowd counting	classifier training 1755 positive samples of 32x32px, and 906 for testing. counting testing 12 minutes of video	Y	N	Y	[38]
Dong	07	people counting, crowd density	2 videos, the overall detection rate was found to be 94.25%.	Y	Y	Y	[36]
Harasse	07	counting people from a video stream in a noisy environment	camera on the bus,	Y	Y	N	[161]
Ke	07	picking up object, waiving, pushing elevator button	20 minutes of video, 160x120px	Y	Y	Y	[162]
Rabaud	06	crowd density	900 320x240px images, and 1000 640x480px images	Y	N	Y	[163]
Rahmalan	06	crowd counting	150 200x200px training and 75 testing images	Y	N	Y	[152]
Wu	06	people counting, crowd density	70 320x240px images	Y	N	Y	[153]
Liu	05	virtual gate crowd counting, proximity to tracks	1 10 minute video	N	N	Y	[34]
Reisman	04	crowd detection	320x240px video from mobile platform	Y	Y	Y	[151]

Table 2 Efficiency, feature and results of some earlier approaches for people counting and tracking

Ref.	Feature	Results
[151]	Optic flow	No empirical analysis, the system can reliably detect crowd at distances of up to 70 meters
[152]	Grey Level Dependency Matrix (GLDM), Minkowsky Fractal Dimensions (MFD), Translation Invariant Orthonormal Chebyshev Moments (TIOCM)	TIOCM (novel) is compared with MFD and GLDM (see right). Accuracy for TIOCM reported as approx. 86% (based on chart), compared to approx. 35% for MFD, and approx. 80% for GLDM. Results based on morning and afternoon conditions. One operating point is used, and no false alarm rates given.
[153]	Statistical methods (Grey Level Dependency Matrix, GLDM)	Total error is less than 12%. No FP rate is reported
[34]	Motion, Blob's color, position, shape, and trajectory	Only visual sample results, no empirical analysis
[161]	skin color model	A 85% counting success rate is achieved compared to the real count, False positives were caused by some arms being counted
[36]	Silhouettes of connected blobs, Fourier descriptors,	Confusion matrix and Receiver Operating Characteristic Curve given. Overall accuracy reported as 94.25%
[38]	Histogram of oriented gradients	Shown by Receiver Operating Characteristic Curve analysis
[163]	Feature tracking based on KLT, connectivity graphs	Average error ranges from 6.3% to 22%. No FP rates reported.
[162]	Spatiotemporal shape contours, optical flow	Shown by Precision and Recall graph, one for each event detected

[15]	people counting approach based on face detection, Kalman filter with kernel based tracking algorithm	Approach displays an accuracy rate up to 93%.
[160]	Space-time image , five the center of gravity, the position of the person's head, the brightness, the size, and shape of the person characteristics	The automatically recorded images coincided with manual selection for about 94% of the total , dropped to 82% in just one case when sunlight entered from an eastern window. Using Internet camera
[159]	Method Based On Intensity Maximum Detection [ABIMD]	ABIMD method allowed achieving an increased 90% accuracy but this method only worked properly when a single person was present

CHAPTER 3

COUNTING PEOPLE

3.1. Introduction

Security and security-related topics have recently received much attention, particularly national, personal, Internet, information, and banking security. Abraham Maslow's Hierarchy of Needs Theory provides a possible explanation for the abiding interest in security matters [164]. According to Maslow, after the most basic physiological needs are satisfied (the need for air, food, and water, for example), human attention shifts to concerns of safety and security: protection from the natural elements, order, law, and stability.

Within the multi-faceted field of security, many scholars have investigated video surveillance technology. These technologies have a wide range of applications in both restricted and open areas. Surveillance systems can provide extra protection for families, homes, and businesses [165] [166]. Advances in object tracking have made it possible to obtain the spatiotemporal motion trajectories of moving objects from videos and use that data to analyze complex events. Moving-object tracking includes accurately counting people, an important task performed by automatic surveillance systems.

According to the Japanese National Police Agency (NPA), many types of violent crimes have increased and become serious problems for numerous institutions and commercial areas [3]. Therefore, observing people's movements for protection and security have become important for these commercial areas and companies, and counting people supports this goal [160]. Moreover, accurately counting people improves business results by accurately analyzing its performance. As management consultant Peter Drucker said, "If you can't measure it, you can't manage it." [2]. However, many such measurements are still performed manually [4]. Therefore, it is necessary to develop an automatic method to count passing people.

3.2. People Counting Systems Based on Image Processing

People counting systems have numerous implications in business world. It could be used to: identify the effects of a new competitor, adjust the rental rate according to the actual density of human traffic, evaluate the success of the activities of marketing and renovations, and confirm the success of the organization's promotions and other marketing activities or to get a snapshot of the activity for each area or each shop. Organizations, interested in counting, could be museums, shops, airports and shopping centers.

There are two main methods for counting people: a manual method or an automatic one. Manual counting is performed by an auditor sitting in the entrance to count people walking in front of him, or her. Such a human counters can achieve a good level of accuracy when few people enter and exit the organization at the same time. However, they lose their reliability over time and when the ins and outs are many. We could then face all kinds of human errors: the numbers are incorrect, entering people are counted as leaving ones (and vice versa) by mistake; Auditors forget to count their colleagues, etc. It is nearly impossible to accomplish this task to perfection through manual auditors. Besides, costs associated with manual counting are very difficult to control as it requires a lot of labor and time if we want to keep an acceptable level of data quality. Additionally, manual counts should be entered manually into a system if it is needed to be processed and analyzed.

On the other hand, most of these deficiencies could be handled through automatic people counting systems. In such systems, counting is performed through many approaches among which is a real-time image processing approach, where a video camera is used to capture video sequences of crossing people and export them to a software package for being processed and interpreted. Such an image-processing- based system could, not only handle the pre-mentioned deficiencies of a manual counting system, but also provide the many advantages shown below:

- It is not affected by the number of people entering or leaving at the same time or the number of consecutive hours spent in service

- Timeliness. Image processing greatly accelerates the work traditionally done for photo interpretation because it is automated and therefore reproducible.
- Repeated and tedious work. This work (geometric correction, calibration, database management,..) are often difficult, if not impossible, to be achieved manually.
- Possibility of repeating the analysis and interpretation. Human interpretation (i.e. photo interpretation) is often tedious, time consuming, costly, and very difficult to reproduce.
- Makes Image interpretation an operator independent process. While the interpretation of the same image by several human photo interpreters often produces very different results, because humans do not have the same sensitivity to colors, context of scanned objects, etc.
- Quantification. A major advantage of image processing is to provide quantified information:
  - Dimensions (area, perimeter, etc..) of all or part of an image (i.e., number of pixels), to form a digital or manual delineation.
  - Chemical concentrations through the spectral characteristics of objects in the image.
  - Physical quantities (roughness, etc.) and optical properties such as reflectance.
- Capability of easily manipulating images and change their veracity
- People are counted as they cross a virtual digital line and thus avoiding the occlusion problem that happens when the camera's field of view is passed by groups of people.
- Counting of several persons while crossing a virtual line in the same or opposite direction.
- Higher speed acquisition allows for dynamic processes to be observed in real time, or stored for later playback and analysis. When combined with a high image resolution, it can generate vast quantities of raw data.
- Performs relatively well in situations where traditional people counting systems fail: such as crowds moving out or in simultaneously.
- Allowing for more functionality at low additional costs,

therefore making them more cost effective.

- Providing a directional counting which gives an extra dimension to the collected data as it enables us to know how long are visitors spend in the organization.
- Such systems are usually scalable so that it could be easily updated according to the progress in image processing techniques
- Allows for monitoring people on a regular and non-interrupted basis (i.e., day-round and year-round) under all weather and lighting conditions.
- The ability to store and post-process images with PC-based software packages enables for re-examining the previously obtained data, compare them with more recently produced images, and optimize them.
- Management of several buildings at a time through a centralized database where an easy access to statistics of every building is available in different time intervals (daily, monthly or yearly).
- The possibility of expanding the area over which data are collected by processing data from more than one camera.
- Invisibility to customers and visitors

3.3. The Proposed Algorithm Overview

The first step in proposed algorithm is acquiring the frame from the camera. The next step in the people counting algorithm is very important called preprocessing step which contents background subtraction, removing noise and establish the four measurement lines. After that movement detection is the key-point for the algorithm to start generate the space-time images. Then treat the human pixel area via segmentation process. The next step is performing the template matching to determine direction and speed. Finally, count according to the direction and number of passing people. The schematic flow chart for the proposed algorithm is shown as Fig.2. More details for the algorithm steps are discussed on the following sub-sections.

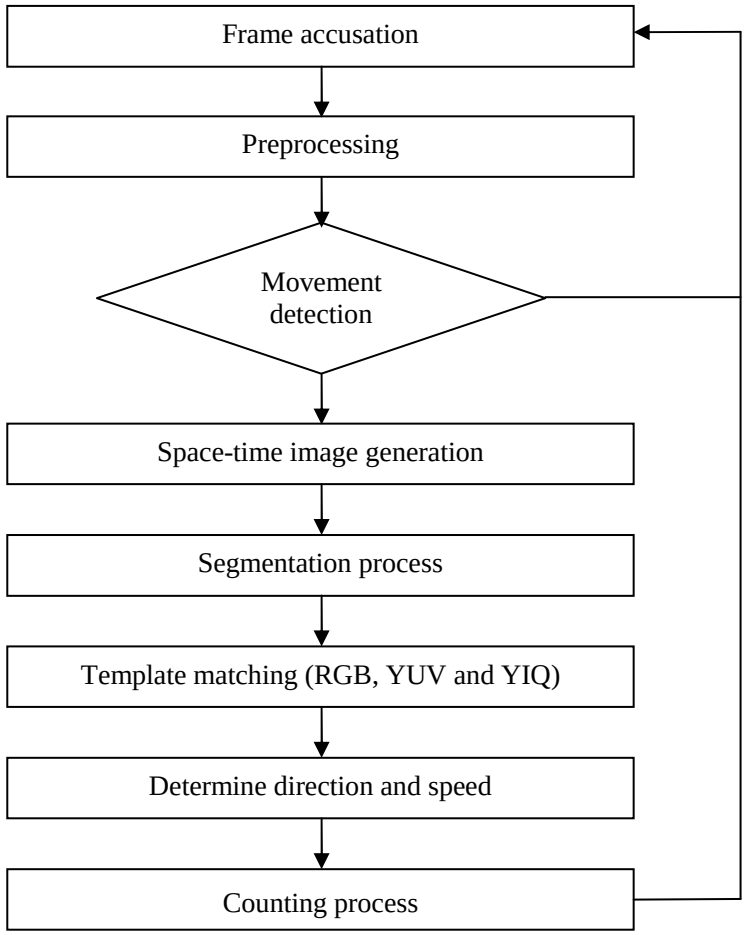


Fig.2: The flow of the proposed algorithm scheme.

3.4. Frame Acquisition

Frames are captured continuously by a camera installed at a surveillance site. The surveillance camera is connected to a personal computer to acquire the image data. Image data of passersby, from the time they enter the frame until they exit the frame, are extracted from the acquired image series. Fig.3 shows an example of image sequences captured by the camera. The 320 × 240 pixel images are captured by the camera as bitmaps. The image is obtained by a USB camera at an average of 17 frames per second. The images in Fig.3 were acquired via a camera installed on the left side of the room near the entrance. In Fig.3 (a), a person enters the frame zone and movement is detected by the algorithm. In Fig.3 (b), one passerby is crossing in front of the camera. In Fig.3 (c), two passersby are walking in close proximity to each other, at the same time, and in the same direction. In Fig.3 (d), two passersby are moving simultaneously in opposite directions.

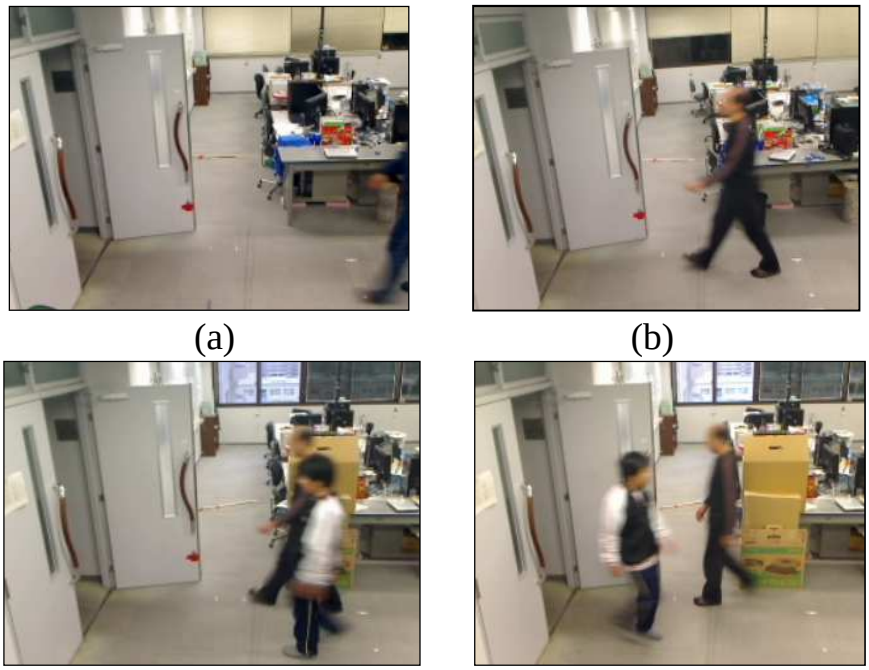


Fig.3: Frames acquired using the surveillance camera: (a) and (b) show a single person passing; (c) and (d) show different examples of two persons passing.

3.5. Image Preprocessing

Using regular surveillance cameras to count visitors introduces some consequential problems. The video feed acquired from the cameras show an exact image of the environment and circumstances. For example, infrared cameras do not give rise to the same problem since anything but heat sources (such as human beings) is filtered out. Because of this all images must be processed before being analyzed. Processing the image involves removing all excessive data from the image such as background and noise. After processing the image one will have an image which clearly shows only the foreground of the image.

Possible detections of passersby are extracted from the frame using a preprocessing stage to generate space-time images. The algorithm employs the following steps to detect movement:

- The first step is to construct a static image to be used as a static background image. This image can be acquired by capturing a frame without any motion. Moreover, the image is

then used as a background reference in order to obtain subsequent images, via pixel subtraction. After subtracting the background reference, the remaining pixels represent possible detection of motion in the frame; this subtraction process is continuous for each frame.

- The second step of pre-processing is removing the noise from the frame by applying a specific morphological filter, a labelling based lookup table, which is used to remove any irrelevant small areas. Therefore, the noise caused by lighting and the color of clothing is reduced.

The images in fig. 4 shown the pre-processing steps, fig. 4(c) image depicts a passerby walking to exit the door. In order to extract the foreground of the person in the image, the background image in fig. 4(b) must be prepared. When the background is known can be used the background subtraction algorithms to remove the background from the image. The algorithms can also be tweaked to exclude some sorts of foreground objects, such as shadows, from the image. The second step is removing noise on top of the images.. The image in fig. 4(c) shows the foreground objects, often referred to as blobs. In this image there is only one blob showing, the person which was walking in front of the camera. In the passersby counting the foreground image, or blob image, shows any objects which move through the frame. In order to count one has to analyses the blobs and track each individual through the frame. It is therefore vital that the background subtraction algorithms return as accurate blob images a possible, so that all steps of the visitor counting are performed correctly.

(a) (b) (c)

Fig. 4: the resulted image after the pre-processing step that contains background subtraction and removing noise.(a)is the original captured image (b)is the background image (c) is the resulted image

3.6. Measurement Line Characteristics

To represent and establish the measurement lines, four vertical lines are set in the image, each line is two pixels in width.(Whenever the line is wide, the size of the passerby appears to be wide, inside the space-time image, and the magnitude of the human-pixel area is represented with a larger amount of pixels.). Two of the four lines, “middle lines,” are in the middle of the image. The two remaining lines, “outer lines,” are located to the left and to the right of the middle lines. The measurement lines used in this study are shown as in Fig.5. The position of the lines is precisely selected, in order to clearly determine the movement directions. A separate background is prepared, from the static background image, for each of the four measurement lines. The data contained inside the four two-pixel-wide measurement lines will be used to generate space-time images.

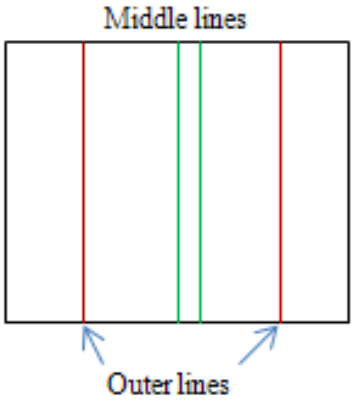


Fig.5: The measurement lines used to generate the space-time image

3.7. Motion Detection

In video surveillance, motion detection refers to the capability of the surveillance system to detect motion and capture the events. Motion detection is usually a software-based monitoring algorithm which, when it detects motions will signal the surveillance camera to begin capturing the event; also called activity detection. An advanced motion detection surveillance system can analyze the type of motion to see if it warrants an alarm. After subtracting each frame from the background to extract the foreground pixels, a connect component operator applied to the result image for clustering to extract the moving objects and also removed the small area (noise).

3.8. Generating Space-Time Images

As discussed, in the previous subsection, human regions are extracted from the captured images using background subtraction, and noise suppression via a labeling filter. Virtual measurement lines are superimposed on the original frame in order to obtain a measurement-line image. Space-time image generating from a video sequences recorded by video camera. The recorded sequences have limited spatial and temporal resolution. Their limited resolutions are due to the space-time imaging process, which can be thought of as a process of blurring followed by sampling in time and in space.

By repeating the process of recording measurement-line image, and arranging measurement line images together with the x-axis (time) and the y-axis (space), a space-time image is produced. An example of how space-time images are generated is shown in Fig.6. After subtracting the static background from each frame, if a motion is detected the measurement lines of the current image are captured; the corresponding static background measurement lines, from the preprocessing stage, are likewise continuously subtracted. The resulting difference of the subtraction process is continuously recorded on the space-time image.

A space-time image contains data for all passersby. When a passerby moves left or right the resulting image is obtained. Since the measurement line are vertical, movement of passersby are seen moving through the measurement lines in a horizontal direction. This causes the shape of the passersby to also appear in a vertical position in the space-time image. When labeling is applied to the space-time image for shape extraction purposes, human objects can be identified as shown in fig.7. The space-time image is representing the shape of the human region. Fig.7 representing sample of result of the space-time Fig.7 (a) the color space-time image after generating Fig.7 (b) the resulted image after applied labeling to extract human objects

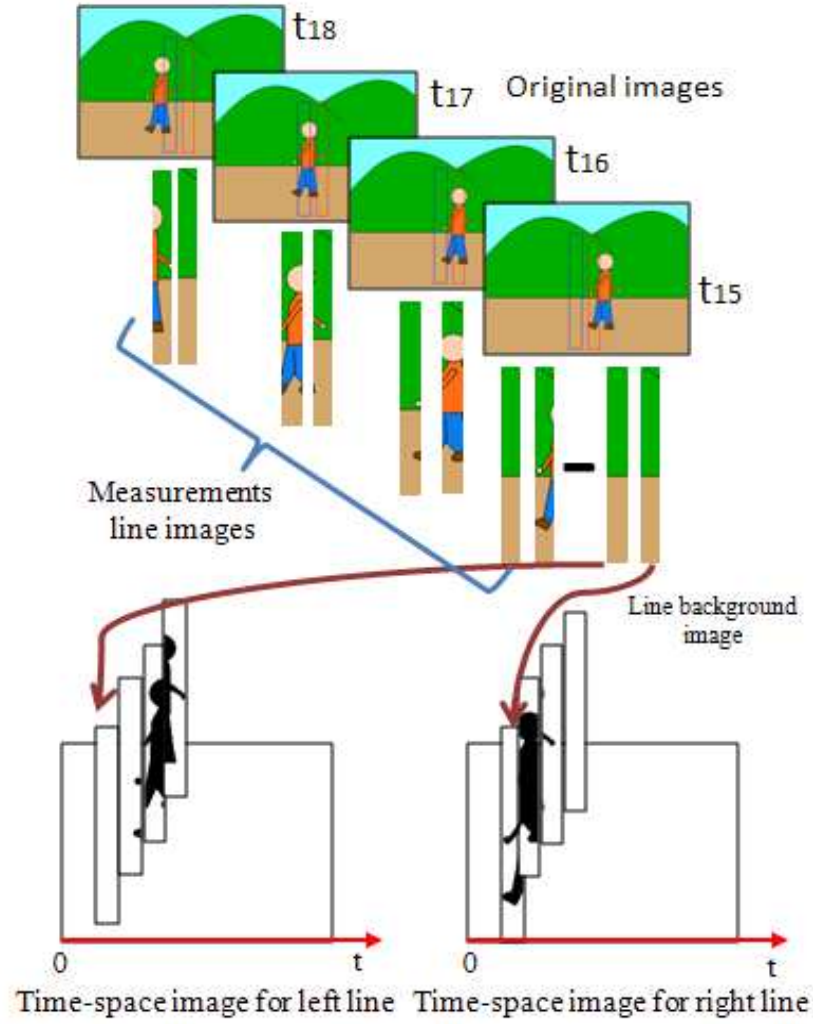
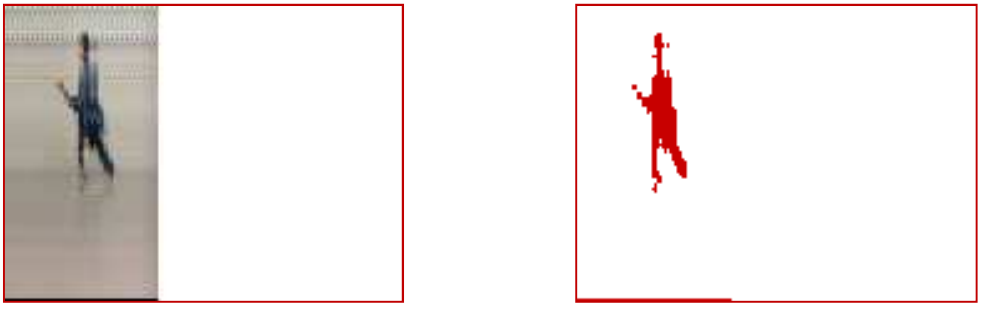


Fig.6: Space-time images generation according to the time.



(a) (b)

Fig.7: sample of result of the space-time (a) the space-time image after generating (b) the result image when labeling is applied to extract human objects

After generating the space-time image the system treats the passerby as one single component, without specifying individual body

parts which influences the counting results. Therefore, a segmentation process is applied to the passerby in order to assemble the disconnected components into a recognizable human shape by using a process of labeling.

3.9. Segmentation

Image segmentation is a fundamental process in many image, video, and computer vision applications. It is often used to partition an image into separate regions, which ideally correspond to different real-world objects. It is a critical step towards content analysis and image understanding. Extensive research has been done in creating many different approaches and algorithms for image segmentation, but it is still difficult to assess whether one algorithm produces more accurate segmentations than another, whether it be for a particular image or set of images, or more generally, for a whole class of images.

3.9.1. Threshold based segmentation

Thresholding is probably the most frequently used technique to segment an image [167][168]. The thresholding operation is a grey value remapping operation g defined by:

$$g(v) = \begin{cases} 0 & if \quad v < t \\ 1 & if \quad v \geq t \end{cases}$$

Where v represents a grey value, and t is the threshold value. Thresholding maps a grey-valued image to a binary image. After the thresholding operation, the image has been segmented into two segments, identified by the pixel values 0 and 1 respectively [169],[170]. If we have an image which contains bright objects on a dark background, thresholding can be used to segment the image fig.8. Since in many types of images the grey values of objects are very different from the background value, thresholding is often a well-suited method to segment an image into objects and background. If the objects are not overlapping, then we can create a separate segment from each object by running a labeling algorithm on the thresholded binary image, thus assigning a unique pixel value to each object.

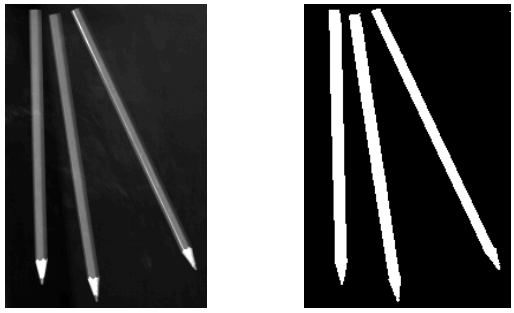


Fig.8: Example of segmentation by thresholding. On the left, an original image with bright objects on a dark background. Thresholding using an appropriate threshold segments the image into objects and background

Many methods exist to select a suitable threshold value for a segmentation task. Perhaps the most common method is to set the threshold value interactively [171][172]; the user manipulating the value and reviewing the thresholding result until a satisfying segmentation has been obtained. The histogram is often a valuable tool in establishing a suitable threshold value.

3.9.2. Segmentation of the Passerby

One of the difficulties for the segmentation algorithm is the background noise that sometime produces different quantities of connected components for the same passerby [174]. Because, inside the space-time image, the shape of each passerby appears almost identical, it is necessary to sometimes assemble the appropriate shape of each passerby via segmentation. Additionally, this problem influences the template matching process: accurately matching the passersby. Template matching is discussed in more detail. This problem also affects the magnitude and size of the human pixels area.

To solve the problem mentioned in the previous paragraph, the method calculates and counts the connected components that represent the same passerby with a different labeling object. This assigns all the connected components the same labeling number. In this case, the passerby is represented as one component. The position of the passerby is vertical. Thus the system search, for other connected components of the passerby, which is also vertical and is located between the left and right boundaries of the pre-segmented connected components. Fig.9 shows the passerby shape before and after segmentation process with the color space-time image.

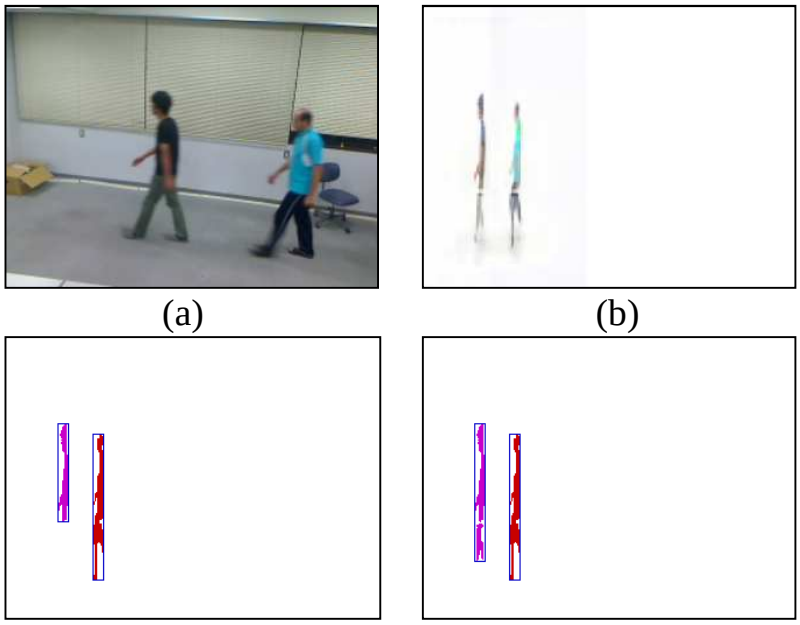


Fig.9: The passerby shape before and after segmentation process: (a) the original image (b) the original color space-time image. (c) before segmentation process. (d) after segmentation process.

3.10. Color Spaces

A color space is a method by which we can specify, create and visualize color. As humans, we may define a color by its attributes of brightness, hue and colorfulness [177],[187]. A computer may describe a color using the amounts of red, green and blue phosphor emission required to match a color. A printing press may produce a specific color in terms of the reflectance and absorbance of cyan, magenta, yellow and black inks on the printing paper[179],[180]. A color is thus usually specified using three co-ordinates, or parameters. These parameters describe the position of the color within the color space being used. They do not tell us what the color is, that depends on what color space is being used.

Different color spaces are better for different applications, for example some equipment has limiting factors that dictate the size and type of color space that can be used. Some color spaces are perceptually linear; a 10 unit change in stimulus will produce the same change in perception wherever it is applied. Many color spaces, particularly in computer graphics, are not linear in this way [182]. Some color spaces are intuitive to use, it is easy for the user to

navigate within them and creating desired colors is relatively easy. Other spaces are confusing for the user with parameters with abstract relationships to the perceived color. Finally, some color spaces are tied to a specific piece of equipment while others are equally valid on whatever device they are used.

3.10.1. RGB Color Space

The RGB color model is an additive color model in which red, green and blue light is added together in various ways to reproduce a broad array of colors. The main purpose of the RGB color model is for the sensing, representation, and display of images in electronic systems, such as televisions and computers [181],[183]. A color is represented by 3 components, corresponding to the relative proportions of the display primaries required to produce it fig.10 and table 3. Although easy to implement it is nonlinear with visual perception. However, RGB is not very efficient when dealing with difficult color images. All three RGB components need to be of equal bandwidth to generate any color within the RGB color cube. Also processing an image in the RGB color space is usually not the most efficient method.

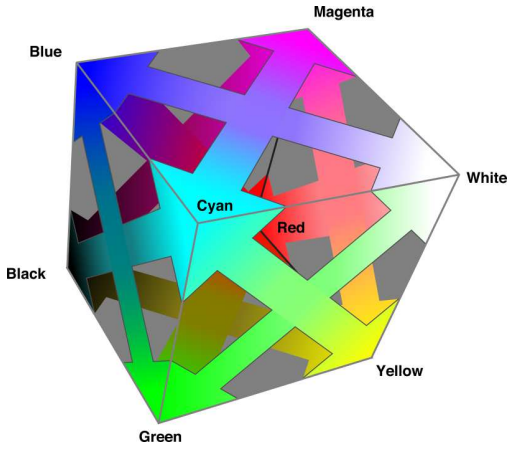


Fig.10 The RGB color space cube

Table 3 The RGB color space cube bars

	Range	white	Black	Red	Green	Blue	Yellow	Magenta	Cyan
R	0 to 255	255	0	255	0	0	255	255	0
G	0 to 255	255	0	0	255	0	255	0	255
B	0 to 255	255	0	0	0	255	0	255	255

3.10.2. YUV Color Space

YUV is a color space typically used as part of a color image pipeline. It encodes a color image or video taking human perception into account, allowing reduced bandwidth for chrominance components, thereby typically enabling transmission errors or compression artifacts to be more efficiently masked by the human perception than using a RGB-representation [184],[180]. Other color spaces have similar properties, and the main reason to implement or investigate properties of YUV would be for interfacing with analog or digital television or photographic equipment that conforms to certain YUV standards by the following equation (1):

$$\left.\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ U &= -0.147 \times R - 0.289 \times G + 0.436 \times B + 128 \\ V &= 0.615 \times R - 0.515 \times G - 0.100 \times B + 128 \end{aligned}\right\} \quad (1)$$

3.10.3. YIQ Color Space

The Y component represents the luma information, and is the only component used by black-and-white television receivers. I and Q represent the chrominance information[186],[182]. In YUV, the U and V components can be thought of as X and Y coordinates within the color space. For digital RGB values with range of 0 to 255, Y has a range of 0 to 255, I has a range of -152 to +152 and Q has a range of -134 to +134. I and Q can be thought of as a second pair of axes on the same graph, rotated 33°; therefore IQ and UV represent different coordinate systems on the same plane. The YIQ system is intended to take advantage of human color-response characteristics [180]. The RGB to YIQ conversion is defined as formula (2):

$$\left.\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ I &= 0.596 \times R - 0.275 \times G - 0.321 \times B + 128 \\ Q &= 0.212 \times R - 0.532 \times G + 0.311 \times B + 128 \end{aligned}\right\} \quad (2)$$

Or we can use formula (3) or formula (4):

$$\begin{bmatrix} I \\ Q \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \cos(33) & \sin(33) \\ -\sin(33) & \cos(33) \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} \quad (3)$$

$$\left.\begin{aligned} R &= Y + 0.956 \times I + 0.621 \times Q \\ G &= Y - 0.272 \times I - 0.647 \times Q \\ B &= Y - 1.107 \times I + 1.704 \times Q \end{aligned}\right\} \quad (4)$$

3.11. Template Matching

Template matching is conceptually a simple process. We need to match a template to an image, where the template is a sub image that contains the shape we are trying to find. Accordingly, we centre the template on an image point and count up how many points in the template matched those in the image [187]. The procedure is repeated for the entire image, and the point that led to the best match, the maximum count, is deemed to be the point where the shape (given by the template) lies within the image.

3.11.1. Template Matching Technique

Template matching is a technique in digital image processing for finding small parts of an image which match a template image. It can be used in manufacturing as a part of quality control, a way to navigate a mobile robot, or as a way to detect edges in images. Template matching can be subdivided between two approaches: feature-based and template-based matching [188]. The feature-based approach uses the features of the search and template image, such as edges or corners, as the primary match-measuring metrics to find the best matching location of the template in the source image[189]. The template-based, or global, approach, uses the entire template, with generally a sum-comparing metric (using SAD, SSD, cross-correlation, etc.) that determines the best location by testing all or a sample of the viable test locations within the search image that the template image may match up to.

3.11.2. Feature-based Approach

If the template image has strong features, a feature-based approach may be considered; the approach may prove further useful if the match in the search image might be transformed in some fashion. Since this approach does not consider the entirety of the template image, it can be more computationally efficient when working with source images of larger resolution, as the alternative approach,

template-based, may require searching potentially large amounts of points in order to determine the best matching location[190].

3.11.3. Template-based Approach

For templates without strong features, or for when the bulk of the template image constitutes the matching image, a template-based approach may be effective. As aforementioned, since template-based template matching may potentially require sampling of a large number of points, it is possible to reduce the number of sampling points by reducing the resolution of the search and template images by the same factor and performing the operation on the resultant downsized images (multi resolution, or pyramid, image processing), providing a search window of data points within the search image so that the template does not have to search every viable data point, or a combination of both[189].

3.11.4. Template-based Matching and Convolution

A basic method of template matching uses a convolution mask (template), tailored to a specific feature of the search image, which we want to detect. This technique can be easily performed on grey images or edge images. The convolution output will be highest at places where the image structure matches the mask structure, where large image values get multiplied by large mask values [190],[192].

This method is normally implemented by first picking out a part of the search image to use as a template: We will call the search image S(x, y), where (x, y) represent the coordinates of each pixel in the search image. We will call the template T(x t, y t), where (xt, yt) represent the coordinates of each pixel in the template. We then simply move the center (or the origin) of the template T(x t, y t) over each (x, y) point in the search image and calculate the sum of products between the coefficients in S(x, y) and T(xt, yt) over the whole area spanned by the template[193]. As all possible positions of the template with respect to the search image are considered, the position with the highest score is the best position. This method is sometimes referred to as 'Linear Spatial Filtering' and the template is called a filter mask.

For example, one way to handle translation problems on images, using template matching is to compare the intensities of the pixels, using the SAD (Sum of absolute differences) measure [190]. A pixel in the search image with coordinates (xs, ys) has intensity Is(xs, ys) and a pixel in the template with coordinates (xt, yt) has intensity It(xt,yt). Thus the absolute difference in the pixel intensities is defined as Diff(xs, ys, x t, y t) = | Is(xs, ys) – It(x t, y t) |.

$$SAD(x,y) = \sum_{i=0}^{T_{rows}} \sum_{j=0}^{T_{cols}} Diff(x+i, y+j, i, j)$$

The mathematical representation of the idea about looping through the pixels in the search image as we translate the origin of the template at every pixel and take the SAD measure is the following:

$$\sum_{x=0}^{S_{rows}} \sum_{y=0}^{S_{cols}} SAD(x,y)$$

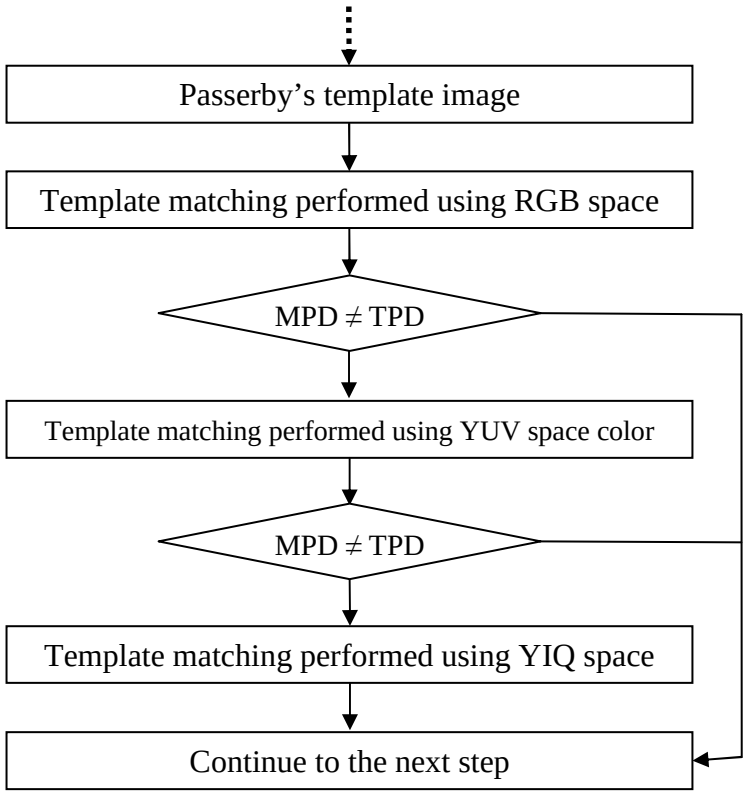
S_{rows} and S_{cols} denote the rows and the columns of the search image and T_{rows} and T_{cols} denote the rows and the columns of the template image, respectively [190],[193]. In this method the lowest SAD score gives the estimate for the best position of template within the search image. The method is simple to implement and understand, but it is one of the slowest methods.

3.11.5. Overview of the Problem

In general, the binary image is used to perform template matching, because it is fast. However, accurately matching more than one passerby appearing in the same space-time image is difficult because of the problem of mismatching. For example, when using the binary image, the shapes of the passersby are nearly identical. This contributes to the problem of mismatching. In this case, when dealing with more than one passerby, using color space such as RGB, YUV, and YIQ is better for accurate matching. Details about these three space colors are discussed.

3.11.6. Optimal Match

[Note: This section introduces two acronyms: MPD for measured pixel-distance and TPD for template-matching-resulting pixel-distance.] A match can be achieved by treating the area of a passerby as a template image taken from the left middle-measurement line of the space-time image and then performing template matching. When two shapes are detected in the space-time image (for an example, see Fig. 9), the pixel-distance between the two shapes is measured (MPD). The template matching process uses different space colors (RGB or YUV), as shown in Fig. 11, and the result covers the whole, or part of, the passerby's shape. The label of the shape can then be identified by reviewing the labeling lookup table and detecting old labels inside the matching result area. After determining the labels of the two shapes, the pixel-distance between the two resulting shapes in the same space-time image is likewise measured (TPD). Comparing the MPD and TPD then yields the optimal match.



MD: Measured pixel-distance between the two shapes.
TD: Measured pixel-distance between the two template-matching result shapes.

Fig. 11. Optimal match flowchart

3.12. Detection of the Direction of the Passersby

To determine the direction of passersby, two space-time images, one for each middle measurement line, are used. The distance between the two middle measurement lines needs to be considered. A passerby completely crosses one middle measurement line before crossing the second. The distance between the middle measurement lines is sufficient to detect the direction of at least one normal walking step. Therefore, measuring the difference between the passerby's position in the two space-time images can determine the direction of motion.

After template matching, a match can be achieved. The passerby's exact position can be determined by applying the labeling concepts in the reference table to the resulting match. Comparing the left position of the passerby in the both space-time images shows the person's direction, as illustrated in Fig. 12. Using the passerby's exact position to detect the direction produces more accurate results than using the passerby's matching position.

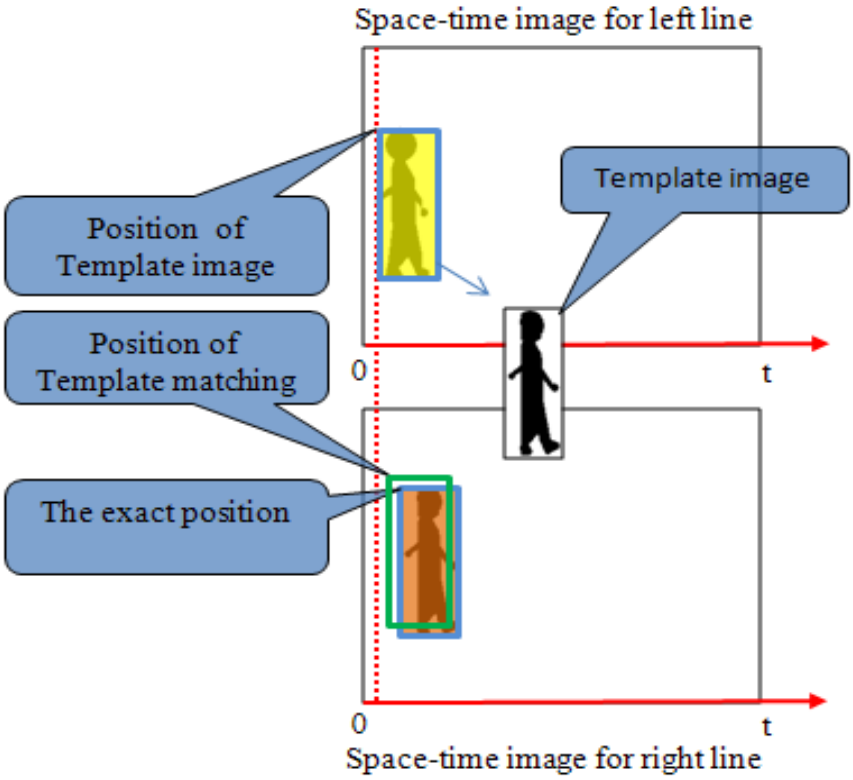


Fig. 12. direction detection using Space-time image.

3.13. Head Position

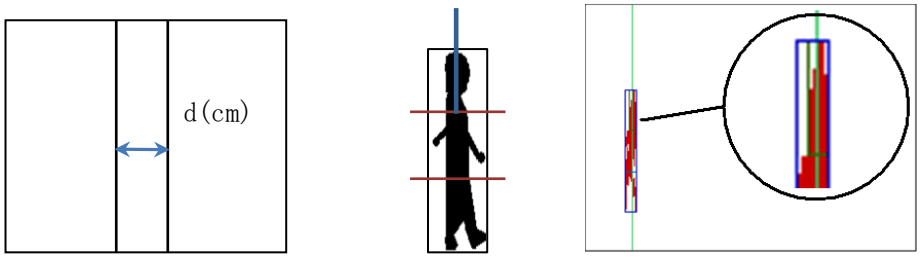
Since the measurement lines are vertical, passersby move horizontally through the measurement lines, causing their shape to appear vertically in the space-time image, as discussed. Dividing the passerby's shape into three equal parts and taking the uppermost part as the passerby detects the position of the head, as shown in Fig. 13(b) and (c). Therefore, the position of the passerby's head is detected twice, once by each middle measurement lines in the space-time image.

3.14. Time Determination

The difference between the passerby's positions in the two space-time images is used to determine the direction of the passerby. Since the x-axis represents time, this difference can be used to calculate the elapsed time. The calculated time, though, is not the precise elapsed time, which is needed; without the correct elapsed time, the speed calculation will be inaccurate. In this case, the exact elapsed time can be calculated by using the vertical center of the body position. The center of the body position can be calculated using the head position. The difference between the centers of the passerby's body position in the two space-time images is calculated in pixels. Since each measurement line is 2 pixels wide, the number of frames can be counted. As a result, multiplying the frame rate by the number of frames yields the precise elapsed time.

3.15. Measurement of the Speed of the Passerby

To determine speed, the two middle measurement lines space-time images are used. The distance between them is measured manually in centimeters, as shown in Fig. 13(a). The calculation of the elapsed time is discussed later. After calculating the precise elapsed time, dividing the distance by the elapsed time yields the passerby's speed.



(a) (b) (c)

Fig. 13. Speed calculation with (a) the distance in d(cm); (b) and (c) represent the passerby's head.

3.16. Human-pixel Area

The human-pixel area is the number of pixels that represent the magnitude of the passerby's shape in the space-time image. Four important factors influence the size and magnitude of the passerby's shape (human-pixel area): the measurement line width, passerby segmentation, frame rate, and speed of the passerby. As discussed passerby segmentation and the width of the measurement lines are influence the human-pixel area. Using different examples, the following discussion focuses on the influence of the frame rate and passerby's speed.

- Frame rate: As illustrated in the following cases, the influence of the frame rate on the human-pixel area can be observed. When using a slow frame rate to acquire the frame, the passerby appears to be thin inside the space-time image, and a small amount of pixels represents the magnitude of the human-pixel area. On the other hand, when using a high frame rate to acquire the frame, the passerby appears to be wide, and a large amount of pixels represents the magnitude of the human-pixel area.
- Speed: When the frame rate is constant, the influence of the passerby's speed on the human-pixel area can be clearly observed clearly. When a passerby's speed is high (fast), the passerby appears thin inside the space-time image, and the magnitude of the human-pixel area is represented with a smaller amount of pixels than used to represent a passerby walking at a normal speed. On the other hand, when a passerby's speed is slow, the passerby appears to be wide inside the space-time image, and the magnitude of the human-pixel area is represented with a larger

amount of pixels than used to represent a passerby walking at a normal speed.

3.17. Pixel-speed Ratio

The human-pixel area is counted, and then the passerby's speed determined. Multiplying the human-pixel area by the speed of the passerby generates the pixel-speed ratio (R), as shown in equation 3:

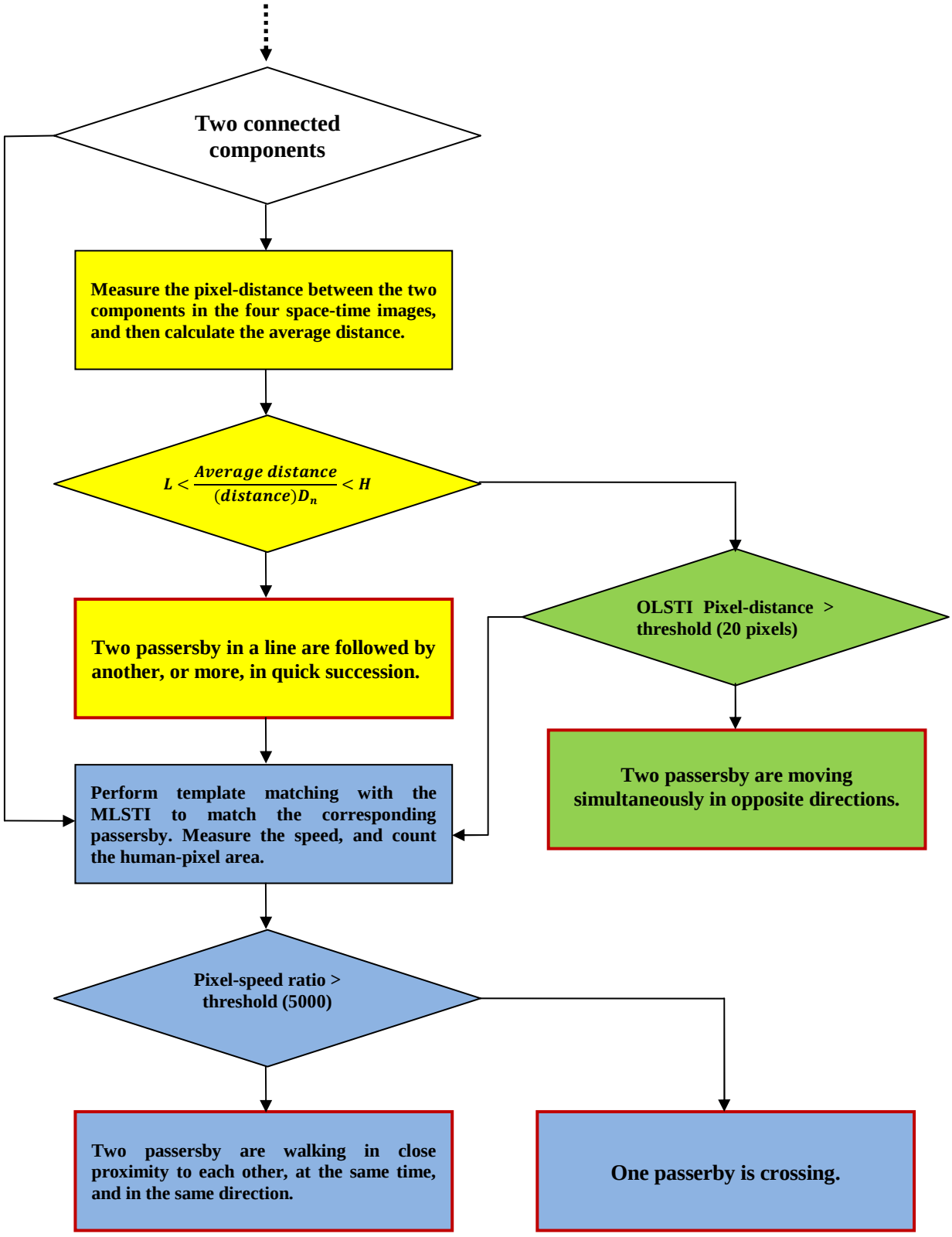
$$R = human - pixel\ area \times passerby\ speed \quad (3)$$

3.18. Counting Process

This section discusses in detail the counting process for different situations: (1) one or two passersby moving in the same direction, (2) two passersby moving in opposite directions, (3) and one passerby followed by another. Fig. 14 provides an overview of the counting process algorithm. As noted earlier, the direction, speed, and human-pixels area are determined first, and then each passerby is counted based on the direction of movement. The following sections explain the counting process for each situation.

3.18..1. Counting Passersby Walking in the Same Direction

When two passersby walk in close proximity, at the same time, and in the same direction, their combined shape appears to be wide in the space-time image, and the magnitude of human-pixel area is represented with a large amount of pixels. This situation is similar to that of one passerby passing at a slow speed, as discussed. Therefore, using only the passerby's shape, it is difficult to determine whether there are one or two passersby. Therefore, a ratio to distinguish between single and multiple shapes is needed.



MLSTI: Middle-measurement line space-time image.

OLSTI: Outer-measurement line space-time image.

Fig. 14. Counting processing algorithm.

3.18..2. Counting Using the Pixel-speed Ratio

The pixel-speed ratio can determine whether the shape is of one or two passersby. If the average value of the ratio in the two middle-measurement line space-time images is more than the value of the threshold chosen after many experiments (set at 5,000 in this work), the system detects two passersby walking in the same direction; otherwise, the system detects one passerby fig.15 . In other words,

$$if\ R \begin{cases} < \text{threshold} & one\ passerby \\ > \text{threshold} & two\ passerby \end{cases}$$

Fig15. Two passersby walk in close proximity to each other, at the same time, and in the same direction.

3.18..3. Counting Passersby Walking in Opposite Directions

Instead of the two middle measurement lines, the two outer measurement lines in the two space-time images are used to count two passersby walking in opposite directions. A passerby crosses one of the two outer lines before crossing the second one. The time needed to reach the second outer line is represented with distance in the space-time image. If two connected components are detected in one or both of the space-time images, the distance between the two passersby is measured in pixels. If the pixel-distance is greater than the established threshold for the elapsed time to complete crossing the two outer lines

(set at 20 pixels in this work), the system recognizes and counts two passersby walking in opposite directions fig.16.

Fig.16 two passersby move simultaneously in opposite directions.

3.18..4. Counting One Passerby Followed by Others.

The counting function is modified to count passersby following each other in a line in quick succession. When two connected component shapes are detected, the pixel-distance (D) between the shapes in the space-time images is measured. The measurement process is repeated and applied to all four space-time images. The average value of the four distances is then calculated using the following equation (4):

$$Average\ distance = \frac{D_1+D_2+D_3+D_4}{4} \qquad (4)$$

After calculating the average value, the relationship between the average value and the distance is defined based on equation 5. The main purpose of this equation is to determine the proportional (commensurate) of the four distances. The low (L) and high (H) values are threshold values. After experimentation, the most effective L and H values for the purposes of counting are chosen. Using equation 5, if the four values of the dividing results have achieved a relationship, the system detects the shapes of two passersby followed by one another.

$$L < \frac{Average\ distance}{(distance)D_n} < H \quad (5)$$

Finally, template matching is performed to match the corresponding passersby. Therefore, achieving optimal matching requires using the measured pixel-distance between the passersby in the space-time images. In addition, the pixel-speed ratio is calculated to determine whether the shape is of one or two passersby. This processing is done independently for each passerby fig.17.

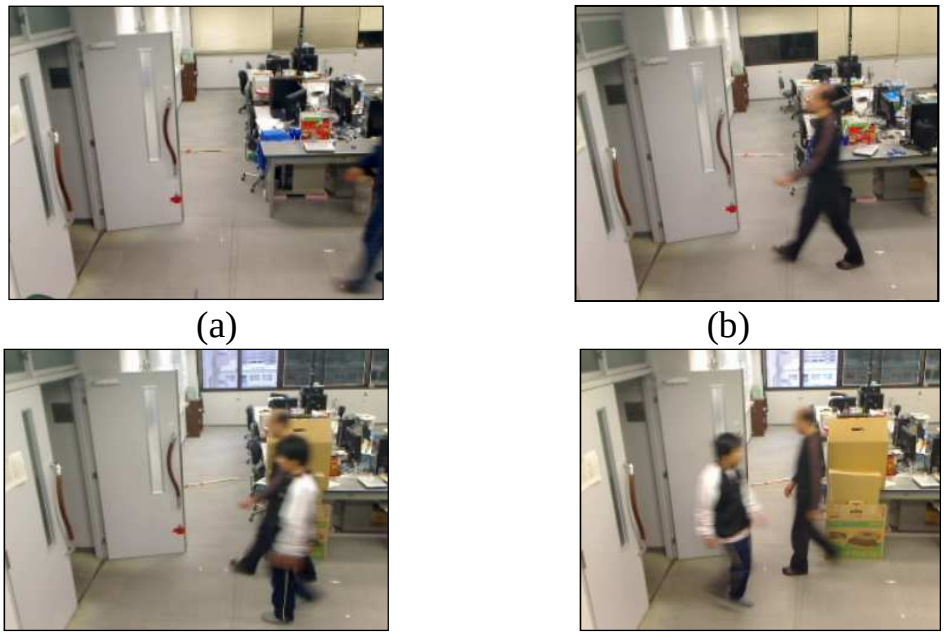
Fig.17 Two passersby walk in close proximity to each other, in the same direction, followed by another two passersby

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1. Experimental Data

The method was tested using a series of video sequences of various scenarios. Fig. 18 shows images captured by a USB camera installed on the left side of a room near the entrance which captured 320 × 240 pixel images at an average of 17 frames per second. In Fig. 18(a), a person enters the frame zone, and the algorithm detects motion. In Fig. 18 (b), one passerby crosses in front of the camera. In Fig. 18(c), two passersby walk in close proximity to each other, at the same time, and in the same direction. In Fig. 18(d), two passersby move simultaneously in opposite directions. The proposed method was tested with 50 cases in each situation. In addition to more than 40 short captured video about 5 minutes and also tested with 9 long captured video.



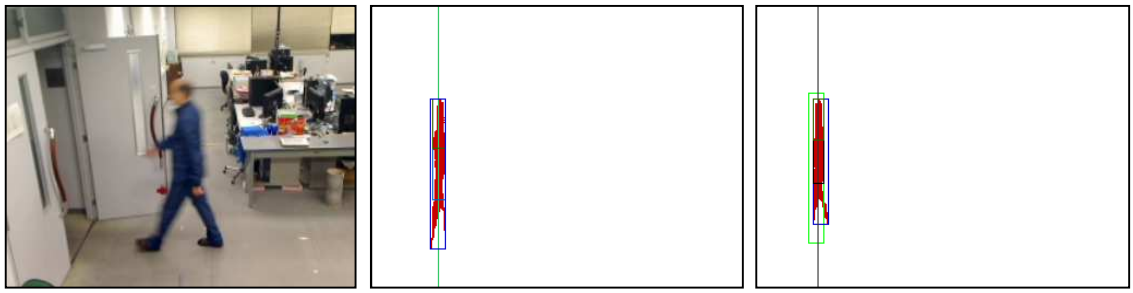
(a) (b)

(c) (d)

Fig. 18. Frames acquired using the surveillance camera: (a) and (b) show a single person passing, (c) and (d) show different examples of two people passing.

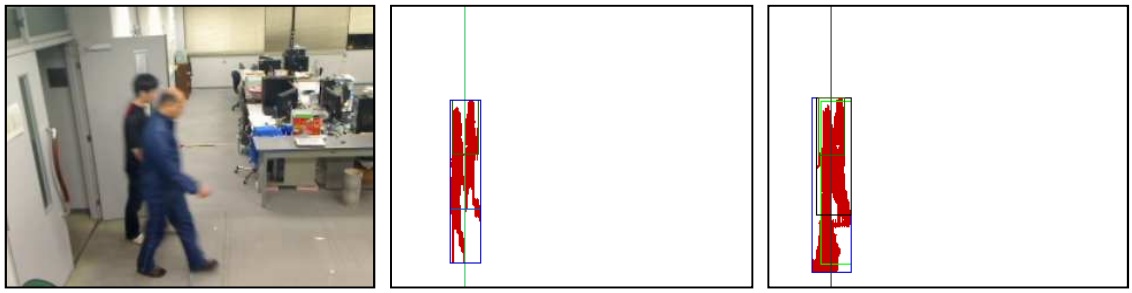
4.2. Experimental Observations

The method successfully counted a single passerby walking in any direction, incoming or outgoing, based on the pixel-speed ratio using the middle-measurement lines space-time images, as shown in Fig. 19. When two passersby walked together at the same time in the same direction, the method counted the two passersby based on the pixel-speed ratio using the middle-measurement line space-time images, as shown in Fig. 20.



(a) (b) (c)

Fig. 19. Single passerby walking in the direction of the exit: (a) the original images, (b) and (c) the left and right middle-measurement lines space-time images.



(a) (b) (c)

Fig. 20. Two passersby walking together in the same direction: (a) the original images, (b) and (c) the left and right middle-measurement lines space-time images.

When two passersby walked in opposite directions, the method precisely counted the two passersby based on the measured pixel-distance between them in the outer-measurement line space-time image, as shown in Fig. 21. Finally, multiple passersby in a line following one another in quick succession were counted based on the measured pixel-distance between two passersby in the four space-time images (equations 4 and 5) and the pixel-speed ratio, as shown in Fig.

22. in which two passersby follow one another, the method measured the pixel-distance between the two shapes found in all four space-time images.

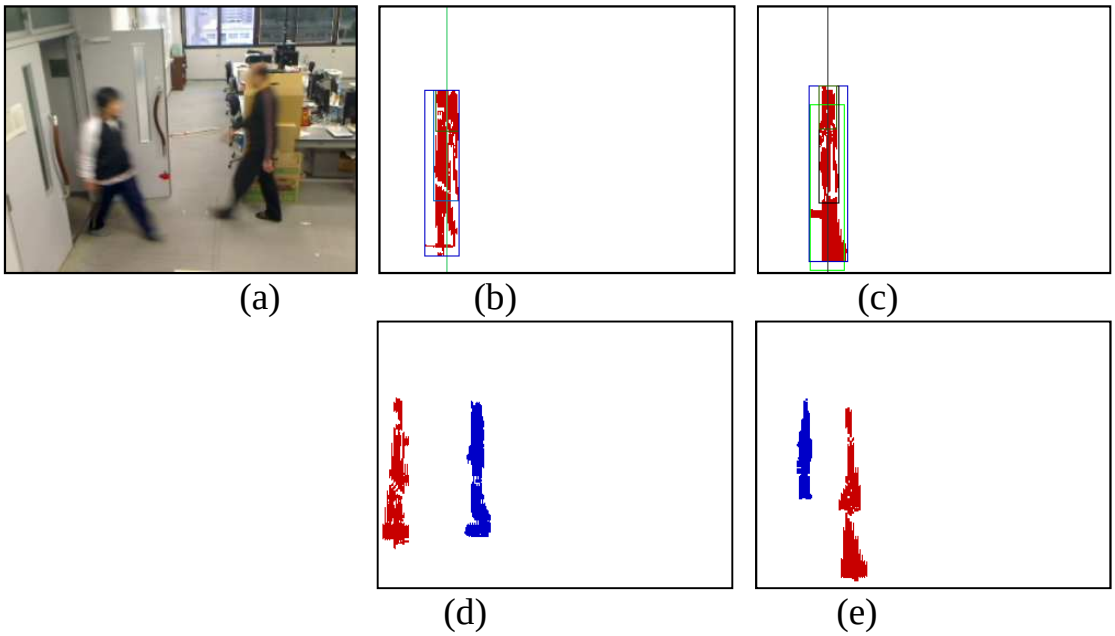


Fig. 21. Two passersby walking in opposite directions: (a) the original images, (b) and (c) the left and right middle-measurement lines space-time images, and (d) and (e) the outer measurement-lines space-time images.

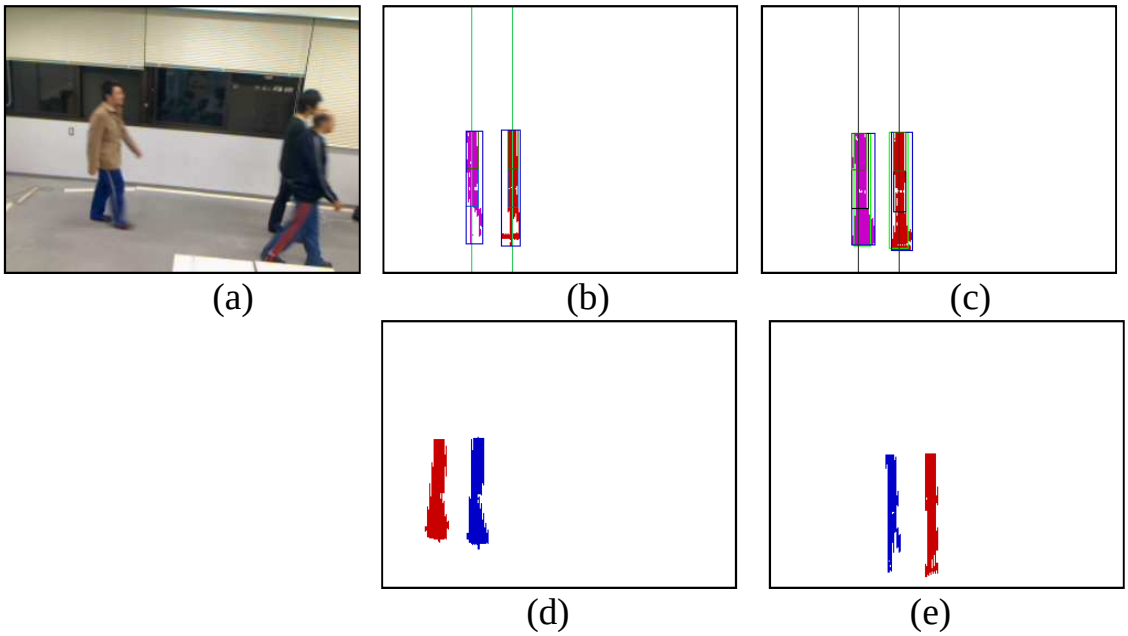


Fig. 22. Two passersby walking together in opposite directions: (a) and (c) time-space image without any processing, (b) and (d) the space-time images.

4.3. Experimental Results

This section discusses our experimental results which demonstrate the successful, accurate matching, accurate direction detection and automatic counting of passersby in various cases and directions in different video sequences.

4.3.1. Matching Accuracy

Accurately matching more than one passerby in the same space-time image is difficult because of the problem of mismatching. When dealing with more than one passerby, the system uses space colors such as RGB, YUV, and YIQ to achieve accurate matching. When the method used only RGB space colors in template matching, the error rate of the results was approximately 15% to 25%. Using RGB and YUV space colors reduced the error rate to approximately 7% to 12 %. Finally, when using three space colors (RGB, YUV, and YIQ), the error rate was virtually unnoticeable (approximately 3%). Fig. 23 illustrates the determination of optimal matching.

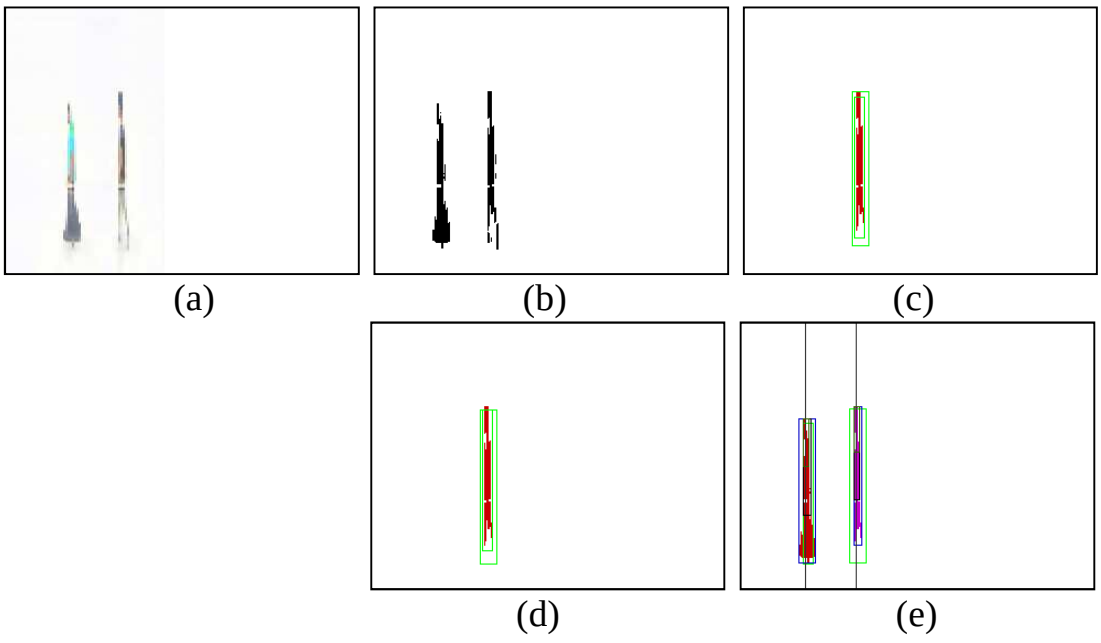


Fig. 23. Template matching that determines the optimal match: (a) color space-time image, (b) binary image, (c) and (d) mismatching with RGB and YUV space colors, (e) correct matching with YIQ space colors.

4.3.2. Passersby Direction Accuracy

Determine the direction of passersby is based on using two space-time images one for each middle measurement line. A passerby crosses one middle measurement line before crossing the other line. By measuring the difference between the passerby's position in the two space-time images we achieve 100% of determine the direction of the passersby. Moreover the passerby's exact position can be determined by applying the labeling concepts (lookup table) in the reference table to the resulting match. Using the passerby's exact position to detect the direction produces more accurate results than using the passerby's matching position.

4.3.3. Passerby Counting Accuracy

The method automatically counted the passersby in various cases and directions from different video sequences. Table 4 shows the counting accuracy for multiple experiments with different situations: one or two passersby moving in the same direction, two passersby moving in opposite directions, and one passerby followed by another when the number of passersby is one, two, three or four and the speed is measured only sometimes.

TABLE 4. Experimental results of the counting algorithm in various situations.

Status		Number of passersby	Undetected	Detected direction (%)	Speed	Speed-pixel ratio	Counting accuracy (%)
One passerby		One	0	100	Measured	Used	100
Two in close proximity		Two	0	100	Measured	Used	90
Opposite direction		Two	0	100	Not measured	Not used	100
Passersby in a line followed by another, or more	One followed by one	Two	0	100	Measured for each	Used	100
	One followed by two in close proximity	Three	0	100	Measured for each	Used	95
	Two in close proximity followed by two in close proximity	Four	0	100	Measured for each	Used	90

Testing the system to count single passerby in any direction the system gave a good result for counting single passerby. Based on using the pixel-speed ratio for the middle-measurement lines space-time images the system counted the two passersby walked together at the same time in the same direction. Also using the pixel-distance in the space-time images the system can count the passersby in opposite direction. The results for 50 cases in each situation, presented in Table 1, confirm that the new method effectively and efficiently counts passersby. Table 5 displays a sample of the system's accuracy at counting passersby in a fixed time (5 minutes). From the sample results the long experiment, the manual count was Exit: 185 and Enter: 209, and the method determined Exit: 180 and Enter: 205. Significantly, the number of passersby was determined and counted successfully, with a high accuracy of approximately 97%.

Table 5. Accuracy of the system's counts of people passing the camera in a fixed time.

Time (min)	Manual count		System count		Accuracy (%)
	Exit	Enter	Exit	Enter	
5	18	21	18	20	99
5	15	15	15	15	100
5	11	18	12	18	98
5	16	17	16	17	100

4.4. Discussion

This work used five characteristics to detect the position of a person's head: the center of gravity, human-pixel area, speed of passerby, and distance between people. These five characteristics enable accurate counting of passersby. The proposed method does not involve optical flow or other algorithms at this level. Instead, human images are extracted and tracked using background subtraction and time-space images. Our method used the widely installed side-view camera, for which the earlier approaches were not applicable. On the other hand, the overhead camera is useful for only one proposes counting people and cannot be used for any other functions compared to the side-view camera. Our method does not require the same

conditions as the earlier methods, such as a distance of at least 10 cm to distinguish passersby and thus count them as two separate people. Our method can overcome this challenge and, in this case, count two passersby using the five previously mentioned characteristics. In addition, the earlier methods sometimes failed to count people with large arm and leg movements. This proposed method, however, can count not only one passerby but also two passersby walking in close proximity at the same time in the same direction or opposite directions and passersby moving in a line in quick succession. As mentioned earlier, using different space colors to perform template matching and automatically select the optimal matching accurately counts passersby with an error rate of approximately 3%, lower than earlier proposed methods.

CHAPTER 5

CONCLUSION

5.1. Conclusion

This thesis proposes a new approach to automatically count passersby using four virtual, vertical measurement lines, which are precisely positioned in the frame. Four space-time images are generated, one for each measurement line, based on a sequence of images obtained from a USB camera. A PC was connected to the camera, which had a rate of 17 frames per second. The USB camera placed in a side-view position, while different types of cameras work from three different viewpoints (overhead, front, and side views). The earlier proposed methods were not applicable to the widely installed side-view cameras selected for this work. This new approach uses a side-view camera that solves three new challenges: (1) two passersby walking in close proximity to each other, at the same time, and in the same direction; (2) two passersby moving simultaneously in opposite directions; and (3) a passerby moving in a line followed by another, or more, in quick succession.

In this approach the measurement lines, four vertical lines, are represent and establish in the frame, each one 2 pixels wide. The wide of the measurement lines is chosen precisely, and the four lines are positioned also precisely to generate the space-time images. Two lines called the middle lines are in the middle of the image and the two remaining lines called the outer lines are to the left and right of the middle lines. Four space-time images are generating one for each measurement lines. The space-time images represent human regions, which are treated with labeling to remove any noise.

In the segmentation process, the system calculates and counts the connected components that represent the same passerby, with different labels, and then assigns all the connected components same labeling number. The system vertically searches for the other connected components of the passerby and horizontally searches for

assigns all the connected components with the same labeling number. So, that the passerby is represented as one component.

The proposed approach is clearly determining the direction of movement, based on using two space-time images from the middle measurement lines. By measuring the difference between the passerby's positions in the two space-time images we achieve 100% accuracy of determine the direction of the passersby. Moreover using the passerby's exact position to detect the direction produces more accurate results than using the passerby's matching position.

In the proposed method, correctly matching more than one passerby in the same space-time image by overcomes the problem of mismatching. Overcoming mismatching requires assembling the correct shape of each passerby through the segmentation process so that the shape appears as one component. Additionally, different color spaces, such as RGB, YUV, and YIQ, are used to perform template matching. When the method used only RGB color space in template matching, the error rate of the results was approximately 15% to 25%. In the other hand, using two, RGB and YUV, color spaces reduced the error rate to approximately 7% to 12 %. Finally, when using three color spaces, RGB, YUV, and YIQ, the error rate was virtually unnoticeable (approximately 3%). Based on the pixel-distance between the two shapes of the passersby in the space-time images the system automatically selects the optimal matching.

In the experiments, a side-view camera was fixed on the left side of the room near the entrance. A PC was connected to the camera, which had a rate of 17 frames per second. The experiment results confirm that the new method effectively and efficiently counts passersby. The method was tested in multiple situations: (1) one passerby walking in any direction; (2) two passersby walking in close proximity, at the same time, and in the same direction; (3) two passersby moving simultaneously in opposite directions; and (4) a passerby followed by another, or more, in a line in quick succession.

This work used five characteristics, the position of a person's head, the center of gravity, human-pixel area, speed of passerby, and the distance between the passersby. These five characteristics enable

accurate counting of passersby. The proposed method does not involve optical flow or other algorithms at this level. Instead, human images are extracted and tracked using background subtraction and time-space images. Our method used the widely installed side-view camera, for which the earlier approaches were not applicable. On the other hand, the overhead camera is useful for only one proposes counting people and cannot be used for any other functions compared to the side-view camera.

As an additional, significant result, the number of passerby was determined and counted successfully. This accurate counting used the pixel-distance between passersby and the relationship between the speed of the passersby and the human-pixel area in the space-time images (pixel-speed ratio). The number of passersby was determined and counted successfully, with a high accuracy of approximately 97%.

The proposed method does not require the same conditions as the earlier methods, such as a distance of at least 10 cm to distinguish passersby and thus count them as two separate people. Our method can overcome this challenge and, in this case, count two passersby using the five previously mentioned characteristics. In addition, the earlier methods sometimes failed to count people with large arm and leg movements. This proposed method, however, can count not only one passerby but also two passersby walking in close proximity at the same time in the same direction or opposite directions and passersby moving in a line in quick succession. As mentioned earlier, using different space colors to perform template matching and automatically select the optimal matching accurately counts passersby with an error rate of approximately 3%, lower than earlier proposed methods.

5.2. Future Work

Counting people and observing their movements are essential processes for security, organizational administration, and the improvement of corporate business results. Counting people is a challenging problem in the field of image processing and computer vision that has recently received much attention in the last few years. The people counter method is able to accurately count in many situations that are difficult for other existing counters method.

However, there are still many aspects of the matching problem that have yet to be explored. A matching error need to be improved to get more accurate matching. Future work could also focus on making the background subtraction process more sensitive to environmental changes and automating updating of the background. Additional improvements could include enabling the method to count groups of people.

REFERENCES

[1] K. Terada, D. Yoshida, S. Oe, and J. Yamaguchi, “A method of counting the passing people by using the stereo images”, International conference on image processing, pp. 338-342,1999.

[2] Byrne, John A.; Gerdes, Lindsey (November 28, 2005). "The Man Who Invented Management". BusinessWeek. Retrieved November 2, 2009.

[3] NPA White Paper: http://www.npa.gov.jp/hakusho.

[4] K. Terada, D. Yoshida, S. Oe and J. Yamaguchi.” A counting method of the number of passing people using a stereo camera”, IEEE Proc. of Industrial Electronics Conf., Vol. 3, pp.1318-1323, 1999.

[5] B. Son, S. Shin, J. Kim, and Y. Her “Implementation of the Real-Time People Counting System using Wireless Sensor Networks”, International Journal of Multimedia and Ubiquitous Engineering, Vol. 2, No. 3, pp. 63-79 July, 2007

[6] A. Vicente, I. Muñoz, P. Molina, and J. Galilea “embedded vision modules for tracking and counting people”, IEEE transactions on instrumentation and measurement, vol. 58, no. 9, pp. 3004-3011, September 2009

[7] N. K. Kanhere, S. J. Pundlik, and S. T. Birchfield. “Vehicle segmentation and tracking from a low-angle off-axis camera”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1152–1157, June 2005.

[8] S. Avidan. “Support vector tracking”. In IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 26, pages 1064–1072, Aug. 2004.

[9] G. Pingali, Y. Jean, and I. Carlbom. “Real-time tracking for enhanced tennis broadcasts”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 0, pages 260–260, June 1998.

[10] A. Leykin. “Visual Human Tracking and Group Activity Analysis: A Video Mining System for Retail Marketing”. PhD thesis, Dept. of Computer Science, Indiana University, 2007.

[11] A. Anjulan and N. Canagarajah. “Object based video retrieval with local region tracking”. Signal Processing: Image Communication, 22 pp.607–621, Aug. 2007.

[12] N. Li, J. Song, R. Zhou, and J. Gu. “A people-counting system based on bp neural network”. In Proc. 4th International Conference on Fuzzy Systems and Knowledge Discovery, Washington, DC, IEEE Computer Society USA, pp 283–287, 2007.
[13] S. H. Laurent, L. Bonnaud, and M. Desvignes. “People counting in transport vehicles”. World Academy of Science, Engineering and Technology(4), 2005.
[14] H. Septian, J. Tao, and Y. Tan. “People counting by video segmentation and tracking”. In Proc. 9th International Conference on Control, Automation, Robotics and Vision, pages 1–4, 5–8 December 2006.
[15] X. Zhao, E. Dellandrea, and L. Chen. “A People Counting System based on Face Detection and Tracking in a Video”. In 6th IEEE International Conference on Advanced Video and Signal Based Surveillance, September 2009.
[16] V. Rabaud and S. Belongie. “Counting crowded moving objects”. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pp. 705–711, 17–22 June 2006.
[17] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, “Detecting Moving Objects, Ghosts and Shadows in Video Streams,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.
[18] B. Boghossian and J. Black, “The challenges of robust video surveillance systems,” in Proceedings of the IEE International Symposium on Imaging for Crime Detection and Prevention, ICDP, pp. 33– 38,2005.
[19] B. E. Flinchbaugh and T. J.Olson, “Autonomous video surveillance,” Proceedings of the SPIE, vol. 2962, pp. 144–151, 1997.
[20] C. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, “Pfinder: real-time tracking of the human body,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780–785, July 1997.
[21] [5] I. Haritaoglu, D. Harwood, and L.S. Davis, “W4: real-time surveillance of people and their activities,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[22] R. Collins, A. Lipton, and T. Kanade, “A System for Video Surveillance and Monitoring: VSAM Final Report,” Tech. Rep., CMU-RI-TR-Robotics Institute, Carnegie Mellon University, May, 2000.
[23] J. Connell, A. W. Senior, A. Hampapur, Y.-L. Tian, L. Brown, and S. Pankanti, “Detection and Tracking in the IBM People Vision System,” in Proceedings of Int’l Conference on Multimedia and Expo, vol. 2, pp. 1403–1406,2004.
[24] D. Comaniciu and P. Meer, “Mean Shift: A Robust Approach Toward Feature Space Analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603–619, 2002.
[25] M. Isard and A. Blake, “CONDENSATION - Conditional Density Propagation for Visual Tracking,” International Journal of Computer Vision, vol. 29, no. 1, pp. 5–28, 1998.
[26] A. Prati, I. Mikic, M.M. Trivedi, and R. Cucchiara, “Detecting Moving Shadows: Algorithms and Evaluation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 918–923, July 2003.
[27] R. Cucchiara, R. Melli, A. Prati, and L. De Cock, “Predictive and Probabilistic Tracking to Detect Stopped Vehicles,” in Proceedings of Workshop on Applications of Computer Vision (WACV), vol. 1, pp. 388–393,2005.
[28] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” IEEE Trans. on Systems, Man, and Cybernetics - Part C, vol. 34, no. 3, pp. 334–352, Aug. 2004.
[29] S. Khan and M. Shah, “Consistent labeling of tracked objects in multiple cameras with overlapping fields of view,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1355–1360, Oct. 2003.
[30] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, “Principal Axisbased Correspondence Between Multiple Cameras for People Tracking,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 663–671, 2006.

[31] A. Calderara, A. Prati, R. Vezzani, and R. Cucchiara, “Consistent Labeling for Multi-camera Object Tracking,” in Proc. of IEEE Int’l Conference on Image Analysis and Processing, 2005.
[32] A. D. Bagdanov, A. Del Bimbo, and F. Pernici, “Acquisition of Highresolution Images Through Online Saccade Sequence Planning,” in Proc. Of ACM Workshop on Video Surveillance and Sensor Networks (VSSN), 2005.
[33] R. Cucchiara, A. Prati, and R. Vezzani, “Advanced Video Surveillance with Pan Tilt Zoom Cameras,” in Proc. of ACM Workshop on Video Surveillance and Sensor Networks (VSSN), 2006.
[34] X. Liu, P.H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, “Detecting and counting people in surveillance applications,” IEEE Conference Advanced Video and Signal Based Surveillance, pp. 306-311, 2005.
[35] I. Cohen, A. Garg, and T.S. Huang, “Vision-based overhead view person recognition,” IEEE Int. Conference on Pattern Recognition, vol. 1, pp. 1119-1124, 2000.
[36] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami, “Fast crowd segmentation using shape indexing,” IEEE Int. Conference on Computer Vision, pp. 1-8, 2007.
[37] S. Lin, J. Chen, and H. Chao, “Estimation of number of people in crowded scenes using perspective transformation,” IEEE Trans. Systems, Man, and Cybernetics Part A, vol. 31, no. 6, pp. 645-654, 2001.
[38] M. Li, Z. Zhang, K. Huang, and T. Tan, “Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection,” Int. Conference Pattern Recognition, pp. 1-4, 2008.
[39] A. Ess, B. Leibe, K. Schindler, and L. van Gool, “Robust multiperson tracking from a mobile platform," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 31, pp. 1831-1846, October 2009.
[40] L. Spinello and R. Siegwart, “Human detection using multimodal and multidimensional features," in Proc. of the IEEE International Con-ference on Robotics and Automation (ICRA), pp. 3264-3269, 2008.

[41] A. M. Elgammal, D. Harwood, and L. Davis, “Non-parametric model for background subtraction.," In Proc. of the 6th European Conference on Computer Vision (ECCV), pp. 751-767, 2000.
[42] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, “Improving shadow suppression in moving object detection with HSV color information.," Proc. of IEEE Conference on Intelligent Transportation Systems (ITS), pp. 334-339, 2001.
[43] S. Bahadori, L. Iocchi, G. Leone, D. Nardi, and L. Scozzafava, “Realtime people localization and tracking through _xed stereo vision," Ap- plied Intelligence, vol. 26, no. 2, pp. 83-97, 2007.
[44] D. Beymer and K. Konolige, “Real-time tracking of multiple people using continuous detection," IEEE Frame Rate Workshop, 1999.
[45] I. Haritaoglu, D. Harwood, and L. Davis, “W4s: A real-time system detecting and tracking people in 2 1/2d.," In Proc. of the 5th European Conference on Computer Vision (ECCV), pp. 877-892, 1998.
[46] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, “Plan-view trajectory estimation with dense stereo background models.," In Proc. of the 8th International Conference on Computer Vision (ICCV'01), pp. 628-635, 2001.
[47] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, “Survey of pedestrian detection for advanced driver assistance systems," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 32, pp. 1239-1258, July 2010.
[48] B. Wu and R. Nevatia, “Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," In Proc. of 10th IEEE International Conference on Computer Vision (ICCV), vol. 1, pp. 90-97, 2005.
[49] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection," in Proc. of the 18th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005.
[50] D. G. Lowe, “Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision (IJCV), vol. 60, pp. 91-110, Nov. 2004.
[51] J. Li, C. S. Chua, and Y. K. Ho, “Color based multiple people tracking.," In Proc. of 7th International Conference on Control, Automation, Robotics and Vision (ICRACV), 2002.

[52] K. Roh, S. Kang, and S. W. Lee, “Multiple people tracking using an appearance model based on temporal color.," In Proc. of 15th International Conference on Pattern Recognition (ICPR'00), 2000.
[53] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[54] J. Giebel, D. Gavrila, and C. Schnor, “A Bayesian framework for multicue 3D object tracking.," In Proc. of European Conference Computer Vision (ECCV), pp. 241-252, 2004.
[55] M. Isard and A. Blake, “CONDENSATION - Conditional Density Propagation for Visual Tracking," International Journal of Computer Vision (IJCV), pp. 5-28, 1998.
[56] O. Lanz, \Approximate Bayesian multibody tracking," IEEE Transac- tions on Pattern Analysis and Machine Intelligence (PAMI), vol. 28, pp. 1436-1449, 2006.
[57] M. Andriluka, S. Roth, and B. Schiele, “People-tracking-by-detection and people-detection-by-tracking," in Proc. of the 21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2008.
[58] D. M. Gavrila, “The visual analysis of human movement: a survey,” Computer Vision and Image Understanding, vol. 73, no. 1, pp. 82–98, 1999.
[59] J. K. Aggarwal and Q. Cai, “Human motion analysis: a review,” Comput. Vis. Image Underst., vol. 73, no. 3, pp. 428–440, 1999.
[60] T.B. Moeslund and E. Granum, “A Survey of Computer Vision-Based Human Motion Capture,” Computer Vision and Image Understanding, vol. 81, no. 3, pp. 231–268, Mar. 2001.
[61] LiangWang,Weiming Hu, and Tieniu Tan, “Recent developments in human motion analysis.,” Pattern Recognition, vol. 36, no. 3, pp. 585–601, 2003.
[62] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, “Human Body Model Acquisition and Tracking Using Voxel Data,” International Journal of Computer Vision, vol. 53, no. 3, pp. 199–223, July-August 2003.

[63] X. Wang, K. Tieu, and E. Grimson, “Learning Semantic Scene Models by Trajectory Analysis,” in Proceedings of IEEE European Conference on Computer Vision, pp. 110–123,2006.
[64] N.T. Nguyen, H.H. Bui, S. Venkatsh, and G. West, “Recognizing and monitoring high-level behaviors in complex spatial environments,” in Proceedings of IEEE Int’l Conference on Computer Vision and Pattern Recognition, 2003.
[65] D. Makris and T. Ellis, “Path detection in video surveillance,” Image and Vision Computing, vol. 20, no. 12, pp. 895–903, 2002.
[66] D. Makris and T. Ellis., “Spatial and Probabilistic Modelling of Pedestrian Behavior,” Proceeding of British Machine Vision Conference, pp. 557–566, 2002.
[67] M. J. Swain. R. E. Kahn, “Gesture Recognition Using the Perseus Architecture,” Tech. Rep., TR-Dept. Computer Science, Univ. Chicago, 1996.
[68] C. R. Wren, B. P. Clarkson, and A. P. Pentland, “Understanding Purposeful Human Motion,” in The Fourth International Conference on Automatic Face and Gesture Recognition, 2000.
[69] C. I. Attwood, G. D. Sullivan, and K.D. Baker, “Model-based Recognition of Human Posture Using Single Synthetic Images,” in Fifth Alvey Vision Conference, 1989.
[70] O. Chomat and J. L. Crowley, “Recognizing Motion Using Local Appearance,” in In International Symposium on Intelligent Robotic Systems, 1998.
[71] R. Nelson. R. Polana, “Low Level Recognition of Human Motion,” in In Workshop on Motion of Non-rigid and Articulated Objects, 1994.
[72] T. Darell, P. Maes, B. Blumberg, and A. P. Pentland, “A Novel Environment for Situated Vision and Behavior,” in Workshop for Visual Behaviors At CVPR-94, 1994.
[73] C. Bregler, “Learning and Recognizing Human Dynamics in Video Sequences,” in Proceedings of IEEE Int’l Conference on Computer Vision and Pattern Recognition, 1997.
[74] T. Starner and A. Pentland, “Real-time American Sign Language Recognition from Video Using Hidden Markov Models,” in Proceedigns of International Symposium on Computer Vision, 1995.

[75] B. Heisele and C. Wohler, “Motion-based Recognition of Pedestrians,” in Proceedings of Int’l Conference on Pattern Recognition, 1998.
[76] I. Haritaoglu, D. Harwood, and L. S. Davis, “Ghost: A Human Body Part Labeling System Using Silhouettes,” in Proceedings of Int’l Conference on Pattern Recognition, pp. 77–82,1998.
[77] S.X Ju, M.J. Black, and Y. Yacob, “Cardboard People: A Parameterized Model of Articulated Image Motion,” in In 2 International Conf. on Automatic Face and Gesture Recognition, 1996.
[78] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima, “Human body postures from trinocular camera images,” in Proceedings of IEEE Int’l Conference on Automatic Face and Gesture Recognition, pp. 326–331,2000.
[79] N. Werghi and Y. Xiao, “Recognition of human body posture from a cloud of 3D data points using wavelet transform coefficients,” in Proceedings of IEEE Int’l Conference on Automatic Face and Gesture Recognition, pp. 70–75,2002.
[80] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima, “Real-time human posture estimation using monocular thermal images,” in Proceedings of IEEE Int’l Conference on Automatic Face and Gesture Recognition, pp. 492–497, 1998.
[81] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, “Human Body Model Acquisition and Tracking Using Voxel Data,” International Journal of Computer Vision, vol. 53, no. 3, pp. 199–223, July-August 2003.
[82] S. Pinzke and L. Kopp, “Marker-less systems for tracking workin postures - results from two experiments,” Applied Ergonomics, vol. 32, no. 5, pp. 461–471, Oct. 2001.
[83] J. Freer, B. Beggs, H. Fernandez-Canque, F. Chevriet, and A. Goryashko, “Automatic recognition of suspicious activity for camera based security systems,” in Proc. of European Convention on Security and Detection, pp. 54–58, 1995.
[84] B. Ozer and W. Wolf, “Human Detection in Compressed Domain,” in Proceedings of IEEE Int’l Conference on Image Processing, pp. 77– 82, 1998.

[85] M.M. Rahman, K. Nakamura, and S. Ishikawa, “Recognizing human behavior using universal eigenspace,” in Proceedings of Int’l Conference on Pattern Recognition, vol. 1, pp. 295–298, 2002.
[86] H. Fujiyoshi and A.J. Lipton, “RealTime Human Motion Analysis by Image Skeletonization,” in Fourth IEEE Workshop on Applications of Computer Vision, 1998.
[87] I.-Cheng Chang and Chung-Lin Huang, “The model-based human body motion analysis system,” Image and Vision Computing, vol. 18, no. 14, pp. 1067–1083, Nov. 2000.
[88] Yi Li, Sondge Ma, and Hanging Lu, “Human posture recognition using multi-scale morphological method and Kalman motion estimation,” in Proceedings of Int’l Conference on Pattern Recognition, vol. 1, pp. 175– 177, 1998.
[89] Changbo Hu, Qingfeng Yu, Yi Li, and Sondge Ma, “Extraction of parametric human model for posture recognition using genetic algorithm,” in Proceedings of IEEE Int’l Conference on Automatic Face and Gesture Recognition, pp. 518–523, 2000.
[90] L. F. Teixeira, and Luis Corte-Real “Cascaded change detection for foreground segmentation” IEEE Workshop on Motion and Video Computing. WMVC '07, 2007.
[91] J. S. C Yuk , K.Y. K. Wong , Ronald H. Y. Chung , F. Y. L. Chin, and K. P. Chow, “Real-Time Multiple Head Shape Detection and Tracking System with Decentralized Trackers” 6th Int. Conf. on Intelligent Systems Design and Applications (ISDA'06) .
[92] A. Elgammal, R. Duraiswami, and L. S. Davis “Probabilistic Tracking in Joint Feature-Spatial Spaces” Proceedings of IEEE CVPR. 1682,2003.
[93] D. Buzan, S. Sclaroff, and G. Kollios “Extraction and Clustering of Motion Trajectories in Video” Proceedings of the 17th ICPR 2004.
[94] F. Porikli “Trajectory Distance Metric Using Hidden Markov Model based Representation” IEEE European Conference on Computer Vision, PETS Workshop, 2004.
[95] X. Liu , P. H. Tu , J. Rittscher, A. Perera, and N. Krahnstoever, “Detecting and Counting People in Surveillance Applications” Proc. Advanced Video and Signal Based Surveillance (AVSS)’05. Teatro Sociale, Como, Italy September 15-16, 2005.

[96] C. Zhao and Q. Pan, “Real Time People Tracking and Counting in Visual Surveillance” Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, 2006.
[97] J. Rittscher, P. H. Tu and N. Krahnstoever, “Simultaneous Estimation of Segmentation and Shape” Proceedings of IEEE CVPR, 2005.
[98] K. Terada, D. Yoshida, S. Oe, and J. Yamaguchi, “A method of counting the passing people by using the stereo images”, International conference on image processing, 0-7803-5467-2,1999
[99] D. Beymer and K. Konolige, “Real-time tracking of multiple people using stereo”, Computer Based Learning Unit, University of Leeds, ,1999
[100] Hashimoto, K. Morinaka, K. Yoshiike, N. Kawaguchi, C. Matsueda, S, “People count system using multi-sensing application”, International conference on solid state sensors and actuators, 0-7803-3829-4,1997
[101] A.J. Schofield, T.J. Stonham, and P.A, “A RAM based neural network approach to people counting”, Fifth International Conference on Image Processing and its Applications , 0-85296-642-3,1995
[102] G. Sexton, X. Zhang, D. Redpath, and D. Greaves, “Advances in automated pedestrian counting”, European Convention on Security and Detection , 0-85296-640-7,1995
[103] Eveland, K. Konolige, and R. Bolley, “Background modeling for segmentation of video-rate stereo sequences," IEEE Computer Vision and Pattern Recognition (CVPR), June 1998.
[104] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “P_nder: Real-time tracking of human body.," IEEE Transactions on Pattern Analysis and Machine Intellegence (PAMI), vol. 19(7), pp. 780-785, 1997.
[105] R. Kaestner, N. Engelhard, R. Triebel, and R. Siegwart, “A Bayesian approach to learning 3D representations of dynamic environments," in Proc. of the 12th International Symposium on Experimental Robotics (ISER), 2010.

[106] C. Stau_er and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking.," Proc. Computer Vision and Pattern Recognition 1999 (CVPR '99), June 1999.
[107] J. Lee, H. Jeon, S. Moon, S. Baik, J. Blanc-Talon, W. Philips, D. Popescu, and P. Scheunders, “Background Modeling Using Color, Disparity, and Motion Information, vol. 3708, pp. 611-617. Springer Berlin - Heidelberg, 2005.
[108] K.-P. Karmann, A. von Brandt, and R. Gerl, “Moving object segmentation based on adaptive reference images," Signal Processing V: Thre- ories and Application, 1990.
[109] K.-P. Karmann and A. von Brandt, “Moving object recognition using an adaptive background memory.," Time-Varying Image Processing and Moving Object Recognition, 1990.
[110] X. Liu, P. H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever. “Detecting and counting people in surveillance applications”. In Proc. IEEE Conference on Advanced Video and Signal Based Surveillance, pages 306–311, 15–16 September 2005.
[111] J. Segen, and S. Pingali,” A Camera-based System for Tracking People in Real Time”, IEEE Proc Of Int. Co nf. Pattern Recognition Vol.3, pp.63-67, 1996.
[112] O. Masoud, and N. P. Papanikolo poulos, “novel method for tracking and counting pedestrians in realtime using a single camera”, IEEE Trans. on Vehicular Tech., Vol. 50, No. 5, pp.1267-1278, 2001.
[113] M. Rossi, and A. Bozzoli, “Tracking and Counting Moving People”, IEEE Proc Image Processing Vol. 3, pp.212-216, 1994.
[114] J. Segen and S.G. Pingali, “A camera-based system for tracking people in real time”, Proceedings of the 13th International Conference on Pattern Recognition, 0-8186-7282-X,1996
[115] I. Haritaoglu and M. Flickner, “Detection and tracking of shopping groups in stores”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 0-7695-1272-0,2001
[116] T. Matsuyama, T. Wada, H. Habe and K. Tanahashi, “Background subtraction under varying illumination”. Trans, IEICE; J84-D-II: 2201–2211,2001.
[117] D. B. Reid. “An algorithm for tracking multiple targets”. In Proc. IEEE Conference on Decision and Control including the 17th

Symposium on Adaptive Processes, volume 17, pages 1202–1211, January 1978.
[118] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. “Tracking multiple moving targets with a mobile robot using particle filters and statistical data association”. In IEEE International Conference on Robotics and Automation, pp. 1665–1670, 2001.
[119] J. Vermaak, S. J. Godsill, and P. Perez. “Monte carlo filtering for multitarget tracking and data association”, IEEE Transactions on Aerospace and Electronic Systems, 41, pp.309–332, 2004.
[120] R. P. S. Mahler. “Multitarget bayes filtering via first-order multitarget moments”. IEEE Transactions on Aerospace and Electronic Systems, 39,4 pp.1152–1178, October 2003.
[121] B. N. Vo, S. Singh, and A. Doucet. Sequential monte carlo methods for multitarget filtering with random finite sets. IEEE Transactions on Aerospace and Electronic Systems, 41(4), pp.1224–1245, 2005.
[122] J. MacCormick and A. Blake. “A probabilistic exclusion principle for tracking multiple objects”, In Proc. 7th IEEE International Conference on Computer Vision, volume 1, pages 572–578, 20–27 1999.
[123] T. Yu and Y. Wu. “Collaborative tracking of multiple targets”. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pp. 834–841, 2004.
[124] B. Wu and R. Nevatia. “Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors”. International Journal of Computer Vision, 75(2), pp.247–266, 2007.
[125] D. Comaniciu, V. Ramesh, and P. Meer. “The variable bandwidth mean shift and data-driven scale selection”. In Proc. 8th Intl. Conf. on Computer Vision, pp. 438–445, 2001.
[126] T. Zhao and R. Nevatia. “Tracking multiple humans in crowded environment”, In Proc. IEEE Computer Vision and Pattern Recognition, volume 2, pp. 406–413, 2004.
[127] C. Huang, H. Ai, B. Wu, and S. Lao. “Boosting nested cascade detector for multi-view face detection”, In Proc. 17th International Conference on Pattern Recognition, volume 2, pp. 415–418,Washington, DC, USA,. IEEE Computer Society2004.

[128] S.-F. Lin, J.-Y. Chen and H.-X. Chao, “Estimation of number of people in crowded scenes using perspective transformation”, IEEE Transactions on Systems, Man and Cybernetics, , Part A 31(6), pp.645–654, 2001.
[129] L. Sweeney, and R. Gross, ,” Mining images in publicly-available cameras for homeland security”. AAAI Spring Symposium on AI Technologies for Homeland Security, Palo Alto, 2005
[130] S.-Y. Cho, T.W. S. Chow and C.-T. Leung, “A neural-based crowd estimation by hybrid global learning algorithm”, IEEE Transactions on Systems, Man and Cybernetics, August, Part B 29(4), pp. 535–541,1999.
[131] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International Journal of Computer Vision (IJCV), no. 47(1-3), pp. 7-42, 2002.
[132] K. Konolige and M. Agrawal, “Frame-frame matching for realtime consistent visual mapping," IEEE International Conference on Robotics and Automation (ICRA), pp. 2803-2810, 2007.
[133] H. Hirschmuller, P. R. Innocent, and M. Garibaldi, “Real-time correlation-based stereo vision with reduced border errors.," International Journal of Computer Vision (IJCV), vol. 47(1/2/3), pp. 229- 246, April-June 2002.
[134] W. E. L. Grimson, “Computational experiments with a feature based stereo algorithm," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 7, pp. 17-34, 1985.
[135] A. Geiger, M. Roser, and R. Urtasun, “E_cient large-scale stereo matching," in Asian Conference on Computer Vision (ACCV), (Queenstown, New Zealand), November 2012.
[136] M. Felsberg, “Disparity from monogenic phase”. in L. Van Gool, Editor," Pattern Recognition (PR), vol. 2449 of Lecture Notes in Computer Science, pp. 248-256, 2002.
[137] T. Fr ohlinghaus and J. M. Buhmann, “Regularizing phase-based stereo", In Proc. of 23th International Conference on Pattern Recognition (ICPR), pp. 451-455, August 1996.
[138] N. S. Sudipta, “Graph cut algorithms in vision, graphics and machine learning, an integrative paper," 2004.

[139] L. Alvarez, R. Deriche, J. S_anchez, and J. Weickert, “Dense disparity map estimation respecting image derivatives: a PDE and scale-space based approach.," Journal of Visual Communication and Image Rep- resentation (JVCIR), vol. 13(1/2), pp. 3-21, 2002.
[140] L. Ladick_y, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. S. Torr, “Joint optimization for object class segmentation and dense stereo reconstruction," International Journal of Computer Vision (IJCV), 2011.
[141] A. C. Davies, J.H. Yin and S. A. Velastin, “Crowd monitoring using image processing”, Electronics & Communication Engineering Journal, February vol.7(1), pp.37–47,1995.
[142] A. Tesei, A. Teschioni, C.S. Regazzoni, G. Vernazza, Long-Memory" matching of interacting complex objects from real image sequences”, Conference on Time Varying Image Processing and Moving Objects Recognition, ,1996
[143] A. Shio and J. Sklansky, Segmentation of people in motion, Proceedings of the IEEE workshop on Visual Motion, 0-8186-2153-2,1991
[144] Y., Weizhong, “Crowd Modeling for Surveillance”, Doctoral Thesis, automation and Computer-Aided Engineering, The Chinese University of Hong Kong, May 2008.
[145] Gary Conrad and Richard Johnsonbaugh, “A real-time people counter”, Proceedings of the ACM symposium on Applied computing, 0-89791-647-6 ,1994
[146] M. Rossi and A. Bozzoli, “Tracking and Counting Moving People”, IEEE Proc. of Int. Conf. Image Processing, ,1994
[147] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos. “Privacy preserving crowd monitoring: Counting people without people models or tracking”. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7, 23–28 June 2008.

[148] Member-Chan, Antoni B. and Member-Vasconcelos, Nuno. “Modeling, clustering, and segmenting video with mixtures of dynamic textures”. IEEE Trans. Pattern Anal. Mach. Intell., 30(5) pp.909–926, 2008.

[149] C. E. Rasmussen and C. K. I.Williams. “Gaussian processes for machine learning” (adaptive computation and machine learning). December 2005.
[150] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the em algorithm”. Journal of the Royal Statistical Society. Series B (Methodological), vol. 39(1),pp.1–38, 1977.
[151] P. Reisman, O. Mano, S. Avidan, and A. Shashua, “Crowd detection in video sequences,” Intelligent Vehicles Symposium, pp. 66-71, 2004
[152] H. Rahmalan, M.S. Nixon, and J.N. Carter, “On crowd density estimation for surveillance,” The Institution of Engineering and Technology Conference on Crime and Security, pp. 540–545, 2006.
[153] X. Wu, G. Liang, K. Lee, and Y. Xu, “Crowd density estimation using texture analysis and learning,” IEEE Int. Conference on Robotics and Biometrics, pp.214–219, 2006.
[154] B. Seongmin, I-K. Jeong, and I-H Lee, “Implementation of crowd system in Maya,” Int. Joint Conference on SICE-ICASE, pp. 2713-2716, 2006.
[155] A. Shendarkar, K. Vasudevan, S. Lee, and Y-J. Son, “Crowd simulation for emergency response using BDI agent based on virtual reality,” Proc. of Winter Simulation Conference, pp. 545–553, 2006.
[156] S. Banarjee, C. Grosan, and A. Abraham, “Emotional ant based modeling of crowd dynamics,” Symbolic and Numeric Algorithms for Scientific Computing, pp.8, 2005.
[157] N. Courty and S.R. Musse, “Simulation of large crowds in emergency situations including gaseous phenomena,” Int. Conference on Computer Graphics, pp. 206-212, 2005.
[158] Y-Y. Lin and Y-P. Chen, “Crowd control with swarm intelligence,” Evolutionary Computation, pp. 3321-3328, 2007.

[159] p. lengvenis, r. simutis, v. vaitkus, r. maskeliunas,” application of computer vision systems forpassenger counting in public transport”, elektronika ir elektrotechnika, issn 1392-1215, vol. 19, no. 3, 2013

[160] K. Terada and K. Atsuta, “Automatic Generation of Passerby Record Images Using Internet Camera”, Electronics and Communications in Japan, Vol. 92, No. pp. 553-560 11, 2009.
[161] S. Harasse, L. Bonnaud, M. Desvignes, “People Counting in Transport Vehicles ”, International Journal of Computer, Information Science and Engineering Vol:1 No:4,pp. 651-654, 2007
[162] Y. Ke, R. Sukthankar, and M. Hebert, “Event detection in crowded videos,” IEEE Int. Conference on Computer Vision, pp. 1–8, 2007.
[163] V. Rabaud and S. Belongie, “Counting crowded moving objects,” IEEE Int. Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 705-711, 2006.
[164] R. Cianci, P. A. Gambrel, “Maslow’s hierarchy of needs: Does it apply in a collectivist culture” Journal of Applied Management and Entrepreneurship, vol.8, No.2, pp.143–161.(2003.
[165] A. Cavoukian, “Guidelines for using video surveillance cameras in public places”. Toronto: Information and Privacy Commissioner, Ontario, 2001.
[166] M. Gray, “Urban surveillance and panopticism: Will we recognize the facial recognition society? Surveillance & Society, vol.1(3), pp.314–330, 2003.
[167] K J. Batenburg, and J. Sijbers, "Adaptive thresholding of tomograms by projection distance minimization", Pattern Recognition, vol. 42, no. 10, pp. 2297-2305, April, 2009
[168] K J. Batenburg, and J. Sijbers, "Optimal Threshold Selection for Tomogram Segmentation by Projection Distance Minimization", IEEE Transactions on Medical Imaging, vol. 28, no. 5, pp. 676-686, June, 2009
[169] W. K. Pratt “Digital Image Processing”: PIKS Inside, Third Edition0-471-37407-5 (Hardback); 0-471-22132-5 (Electronic). John Wiley & Sons, Inc.2001
[170] Dwayne Phillips Image Processing in C Second Edition 0-13-104548-2 R & D Publications, 2000
[171] R. Lukac K.N. Plataniotis “Color image processing method and applications” 978-0-8493-9774-5, 2007

[172] S. Al-amri, N.V. Kalyankar and S.D. Khamitkar” Image Segmentation by Using Thershod Techniques” , journal of computing, volume 2, issue 5,pp. 83-86 may 2010
[173] Gonzalez and Woods, "Digital image processing", 2nd Edition, prentice hall, 2002.
[174] Kenneth R. Castelman, "Digital image processing", Tsinghua Univ Press, 2003.
[175] F. Samopa, A. Asano.,"Hybrid Image Thresholding Method using Edge Detection", IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.4, PP.292-299, April 2009.
[176] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, “semantic image segmentation and object labeling”, ieee transactions on circuits and systems for video technology, vol. 17, no. 3, pp. 298-312 march 2007
[177] D.H Brainard, “Calibration of a computer controlled color monitor”, Color Research & Application, 14, 1, pp 23-34 1989.
[178] W.B. Cowan, “An inexpensive scheme for calibration of a colour monitor in terms of CIE standard coordinates”, Computer Graphics, Vol. 17 No. 3, 1983.
[179] R.S. Berns, R.J. Motta, and M.E. Gorzynski, “CRT Colorimetry: Part 1 Theory and Practice, Part 2 Metrology”, Color Research and Application, 18, 1993.
[180] W.N. Sproson, “Colour Science in Television and Display Systems.”, Adam Hilger Ltd, 1983. ISBN 0-85274-413-7
[181] Travis, D, “Effective Color Displays. Theory and Practice.”, Academic Press, ISBN 0-12-697690-2 (This contains C source code for many colour space conversions.) 1991.
[182] “Computer Generated Colour.”, R. Jackson, L. MacDonald, K. Freeman, John Wiley and Sons, 1994.
[183] “Digital Image Processing.”, Rafael C. Gonzalez and Richard E. Woods, Addision Wesley, 1992.
[184] “Fundamentals of Digital Image Processing.”, Anil K. Jain, Prentice-Hall International, 1989.
[185] “Computer graphics : principles and practices.”, James D. Foley, et al. 2nd ed. Addison- Wesley, c1990.

[186] A. ALBIOL, L. TORRES, AND E. J. DELP, “Optimum color spaces for skin detection”. In Proceedings of the International Conference on Image Processing, vol. 1, pp.122–124. 2001.
[187] G. GOMEZ, , AND E.MORALES, “Automatic feature construction and a simple rule induction algorithm for skin detection”. In Proc. of the ICML Workshop on Machine Learning in Computer Vision, 31–38. 2002.
[188] M. Sonka, “Image Processing, Analysis and Machine Vision”, Chapman & Hall, London, 1993.
[189] K. R. Castleman, “Digital Image Processing”, Prentice Hall International, Inc., New Jersey, 1996.
[190] F. van der Heijden, “Image Based Measurement Systems”, John Wiley & Sons, Inc., West Sussex, 1994.
[191] G. A Baxes, “Digital Image Processing”, John Wiley & Sons Inc., Canada, 1994.
[192] A. Low, “Introductory Computer Vision and Image Processing”, McGraw-Hill Book Company, London, 1991.
[193] B. S. Morse, “Segmentation (Matching, Advanced)”, Brigham Young University. 1998.