

# **A Background Subtraction Algorithm for a Pan-Tilt Camera**

by

Ying Chen

A thesis submitted to the Faculty of Graduate Studies and Research  
in partial fulfillment of the requirements for the degree of

**Master of Science**

Department of COMPUTING SCIENCE  
University of Alberta

©Ying Chen, 2014

# Abstract

This thesis develops a background subtraction algorithm for detecting moving objects with a pan-tilt camera. Traditional background subtraction methods assume that the camera is static, limiting their applications on a moving platform. To relax this assumption, the camera motion is compensated by registering the current image with respect to the background model image, and then conventional background subtraction approaches can be applied. Precise image registration is non-trivial and pixels can be misaligned. Appearance-based background subtraction approaches are sensitive to pixel misalignment which may generate false positives, i.e., background pixels labelled as foreground. Motion information can be used to separate these noises from truly moving objects. However, motion estimation can be inaccurate, and thus may also contribute to false positives. We observe that pixel misalignment and motion inaccuracy tend not to co-occur for a given pixel. This leads to a decision rule that a pixel can be foreground only if it neither follows the appearance nor the motion background model. This strategy can largely reduce false alarms though it may be conservative in terms of false negatives, i.e., foreground pixels labeled as background. Consequently, we propose to classify a pixel by evaluating its appearance and motion models marginally. We can minimize the detrimental effect of spurious false negatives caused by this conservative classification rule by imposing spatial constraints on pixel label in a Markov-Random-Field (MRF) framework - solved via the graph-cut algorithm. To be more specific, a two-layer graph model is proposed - one layer for evaluating appearance model and the other the motion model - where two types of spatial constraints are imposed, one of which considers the correspondence of pixel labels in two layers (label consistency constraint) whereas the other considers the spatial information between neighbouring pixels (spatial coherence constraint). The final label of a given pixel can be determined by anding its corresponding labels (background/foreground) in two layers according to the strategy. We show through extensive experiments (both outdoors and indoors) that our solution is superior to the competing background subtraction algorithms designed for dealing with a pan-tilt camera.

*You've gotta dance like there's nobody watching,  
Love like you'll never be hurt,  
Sing like there's nobody listening,  
And live like it's heaven on earth.*

– William W. Purkey

# Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Hong Zhang, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate his vast knowledge and skill in many areas, and his continuous assistance which help me through the learning process and research.

I would like to thank the other members of my committee, Herb Yang and Nilanjan Ray for taking time out from their busy schedules to read and comment my thesis.

I would also like to thank my family for the support they provided me through my entire life and my husband for his understanding and encouragement. I would like to thank all my collages and friends who helped and supported me in the last three years.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problems in Existing Background Subtraction Methods from a Pan-Tilt Camera	2
1.2	Thesis Objective and Contribution	2
1.3	Organization	4
<b>2</b>	<b>Related Work</b>	<b>5</b>
2.1	Introduction	5
2.2	Background Subtraction from a Static Camera	6
2.2.1	Parametric Background Modelling	7
2.2.2	Non-parametric Background Modelling	9
2.3	Background Subtraction with a Pan-Tilt Camera	13
2.3.1	Image Registration	14
2.3.2	Background Subtraction Approaches with a Pan-Tilt Camera	15
2.4	Summary	19
<b>3</b>	<b>Methodology</b>	<b>21</b>
3.1	Introduction	21
3.2	Image Registration	22
3.3	Background Modelling	25
3.3.1	Appearance Model	25
3.3.2	Motion Modelling	28
3.4	Graph Construction and Graph Cuts	29
3.4.1	Graph Cuts Algorithm	29
3.4.2	A Two-layer Graph Model	30
3.5	Summary	32
<b>4</b>	<b>Experimental Results and Discussion</b>	<b>34</b>
4.1	Datasets	35
4.1.1	Outdoors Sequences	35
4.1.2	Indoors Sequences	35
4.2	Evaluation Metrics	37
4.3	Approaches for Comparison	38
4.4	Qualitative Results	39
4.4.1	Ex.1 - Appearance Only vs Appearance and Motion	40
4.4.2	Ex.2 - Importance of Spatial Constraints: with vs without	41
4.4.3	Ex.3 - Modelling Appearance and Motion: Jointly vs Marginally	41
4.5	Quantitative Results	43
4.6	Summary	43
<b>5</b>	<b>Conclusion and Future Work</b>	<b>52</b>
	<b>Bibliography</b>	<b>54</b>

# List of Tables

3.1	Weights definition of the edges in the two-layer graph model $\mathcal{G}$ . . . . .	33
4.1	Summarization of our approach and three other approaches for comparison. "Marginally modelled" and "Jointly modelled" refer to as the two ways of incorporating appearance cue and motion cue in background modelling. .	40
4.2	Qualitative comparison on outdoor sequence 1. . . . .	44
4.3	Qualitative comparison on outdoor sequence 2. . . . .	45
4.4	Qualitative comparison on outdoor sequence 3. . . . .	46
4.5	Qualitative comparison on indoor sequence 1. . . . .	47
4.6	Qualitative comparison on indoor sequence 2. . . . .	48
4.7	Qualitative comparison on indoor sequence 3. . . . .	49
4.8	Quantitative evaluation on outdoor sequence 1. . . . .	49
4.9	Quantitative evaluation on outdoor sequence 2. . . . .	49
4.10	Quantitative evaluation on outdoor sequence 3. . . . .	50
4.11	Quantitative evaluation on indoor sequence 1. . . . .	50
4.12	Quantitative evaluation on indoor sequence 2. . . . .	50
4.13	Quantitative evaluation on indoor sequence 3. . . . .	50

# List of Figures

1.1	Problems in existing background subtraction methods from a pan-tilt camera. (a) From left to right: the current frame; image difference between the warped image and the current frame; the ground truth; the detection result. (b) From left to right: the current frame; the motion magnitude of the current frame: the darker the intensity, the smaller the motion magnitude; the ground truth; the detection result. . . . .	3
1.2	Appearance difference and motion difference to the background model. . . . .	4
2.1	Illustration of traditional background subtraction approaches and their limitation on the moving platform.(a) An example of a traditional background subtraction approach; (b) An example when background subtraction fails (the camera pans to the left). From left to right of both (a) and (b): the current frame, the estimated background, the detection. . . . .	6
2.2	Illustration of kernel density estimation. (a): The comparison of the construction of histogram and kernel density estimators using six data points (red dots on the $x$ -axis). The horizontal axis is divided into bins covering the range of the data. A box of height of $1/12$ is placed whenever a data point falls inside this interval. For the kernel density estimate, a normal kernel (a red dash curve) is placed on each of the data point. The final kernel density estimate is the blue curve. (b): KDE with different bandwidths of a random sample with 1000 points from a standard normal distribution. Green: the true density (standard normal). Red: KDE with $H = 0.2$ ; Blue: KDE with $H = 1.37$ ; Magenta: KDE with $H = 3.16$ . . . . .	10
2.3	Illustration of support vector machine. $H_1$ does not separate the classes. $H_2$ does but the margin is small. $H_3$ is an appropriate hyperplane with the maximum margin. . . . .	11
2.4	A standard graph designed for background subtraction . . . . .	12
2.5	A simple example: illustrate two types of outliers that background subtraction approaches may fail to deal with: a) the outliers from pixel misalignment produced by imprecise image registration; b) the outliers from inaccurate motion estimate if motion is incorporated. . . . .	18
3.1	The main steps of our approach . . . . .	22
3.2	Feature-based global image registration by applying Kanade-Lucas-Tomasi Tracker for tracking detected features (features matching) and RANSAC to reduce the features that are outliers. (a) Left: the target frame. Right: the reference frame which is shifted to the right. (b) Matched feature points. Green rectangle: the location of the target frame in the reference frame. . . . .	24
3.3	Background/Foreground Classification with ViBe . . . . .	26
3.4	Background Model Update Scheme of ViBe . . . . .	27
3.5	The two-layer graph model of our proposed approach. . . . .	31
4.1	The datasets used for evaluation . . . . .	36
4.2	Illustration of the motion field of Frame 155 <sup>th</sup> in outdoor sequence 1. . . . .	42

# Chapter 1

## Introduction

Moving object detection is one of the active and challenging problems for high-level tasks of computer vision applications, such as object tracking, video surveillance and indoor/outdoor navigation. A good object detection approach requires low false alarm rate and miss-detection rate that are of great importance for the higher level tasks.

There has been considerable work on moving objects detection which can be divided into three categories: feature-based object detection, motion segmentation and background subtraction. The idea of feature-based object detection is to model the moving objects with representative features. Its performance heavily depends on the features' quality, which might be sensitive to illumination changes and scale changes, etc. The number of detected features also plays an important role. A moving object with too few features, e.g., an object with only few pixels, might not be detected with feature-based approaches. Motion segmentation methods detect moving objects by grouping those pixels with similar motion. Yet, these approaches may be sensitive to noises and have difficulty in detecting exact object boundaries. Different from the first two categories, background subtraction models the background and thus, moving objects, which deviate from the background model can be detected. Compared to the first two categories, background subtraction is stable and robust which can identify any moving object independent of its size or shape.

The rationale of background subtraction is to identify the moving objects from a video frame that differ significantly from a background model [37]. The success of background subtraction algorithms has led to their ubiquitous use in surveillance system with stationary cameras. However, the common requirement of traditional background subtraction approaches that the camera remains static severely limits its usage to moving platforms, such as mobile and PTZ cameras.

In this thesis, we propose a novel background subtraction method which is able to detect moving objects both in a static camera or a pan-tilt camera. In section 1.1, we discuss the problem in detail and Section 1.2 describes our goal and contributions. An outline of the thesis is given in Section 1.3.



## 1.1 Problems in Existing Background Subtraction Methods from a Pan-Tilt Camera

Traditionally, the extension of background subtraction to a pan-tilt camera can be achieved by compensating the camera motion. Here pixels are aligned with their corresponding background models which are built and maintained from the previous frames, thus conventional background subtraction can be performed. Yet precise image registration is non-trivial and a pixel may turn out to be a false positive, i.e., a false foreground detection, due to pixel misalignment. False positives usually occur on edges since the background models on the boundaries are very different, e.g., a misaligned pixel (marked as red surrounded by a red circle) in Figure 1.1 (a), is falsely labeled as foreground due to the large difference to its unmatched background model.

This problem can be alleviated by exploiting pixel wise motion information, since we assume that a background pixel does not move after the camera motion is compensated. However, as we mentioned in the previous section, motion estimate itself can be inaccurate. For example, in Figure 1.1 (b), the motion of a pixel in the grass area (surrounded by a red circle) is incorrectly large due to the lack of texture leading to a false detection. These inaccuracies can result in either false positives (when motion estimate is incorrectly large) or false negatives (when motion estimate is incorrectly small).

Motion and appearance information are important to detect moving objects from background. In this thesis, how to incorporate the motion and appearance information in background subtraction to improve the performance of detection in a pan-tilt camera is our main concern.

## 1.2 Thesis Objective and Contribution

In this thesis, we propose a novel and robust background subtraction method to detect moving objects from a pan-tilt camera.

Existing extension of background subtraction to a pan-tilt camera is to compensate the background motion caused by camera movements, i.e., the consecutive frames are aligned in the same coordinate system so that pixels can be compared to its corresponding background models. Image registration can be global or local. In global approaches, a homography is computed and used to align the background model image with the current image before applying background subtraction. Since it is hard to align objects with different depth variations properly with a global homography transformation, these approaches usually focus on improving the robustness of the conventional background subtraction approaches. One way to reduce the false positives caused by pixel misalignment is to consider the spatial information in the local region, assuming that pixels can be aligned properly from its

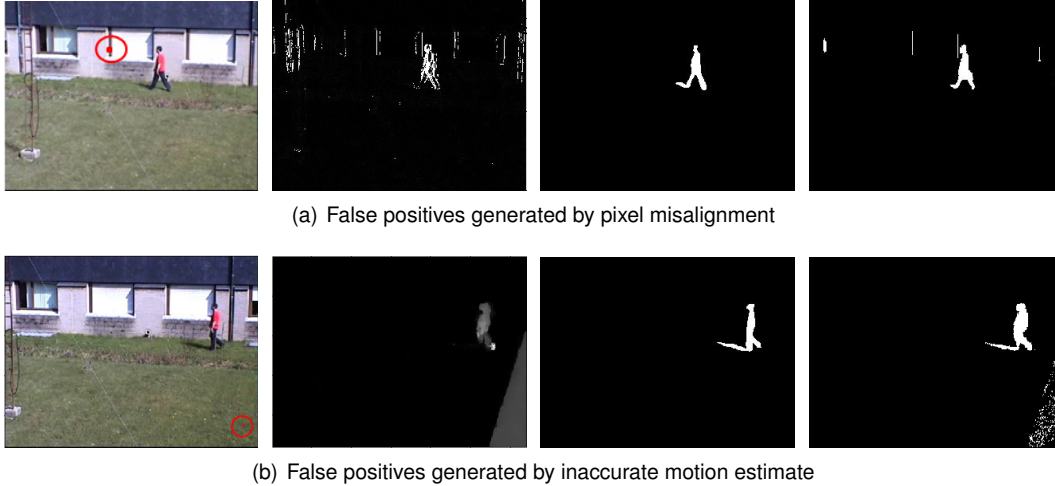


Figure 1.1: Problems in existing background subtraction methods from a pan-tilt camera. (a) From left to right: the current frame; image difference between the warped image and the current frame; the ground truth; the detection result. (b) From left to right: the current frame; the motion magnitude of the current frame: the darker the intensity, the smaller the motion magnitude; the ground truth; the detection result.

neighbourhoods. Even though, these approaches can not deal with large misalignment especially when the camera moves fast. Local image registration, on the other hand, focuses on improving the accuracy of pixel alignment by registering each pixel locally. In local approaches, an Expectation-Maximization (EM) framework is usually employed to iteratively switch between pixel wise motion estimation and background subtraction to detect moving objects. One drawback of these local registration approaches is that it is susceptible to reach in the local minima.

Most of the conventional background subtraction approaches are based on appearance only. These approaches are vulnerable to illumination changes, etc. Besides, these approaches are sensitive to outliers in image registration. Motion can be incorporated with appearance to provide higher discriminative power. However, accurate motion estimate is impractical, and these inaccuracies can result in either false positives, i.e., background pixels can be classified as foreground when their apparent motion estimate is incorrectly large, or false negatives, i.e., foreground pixels can be classified as background when the motion estimate is incorrectly small.

As both motion and appearance are important for detecting moving objects, yet their inaccuracies may contribute to false positives, our goal is to design a background subtraction algorithm using both appearance and motion cues that can suppress significantly false positive detections while introducing minimum false negatives.

Our algorithm exploits the observation that pixel misalignment and inaccurate motion estimation tend to occur not simultaneously or concurrently at a pixel (Figure 1.2), i.e., the

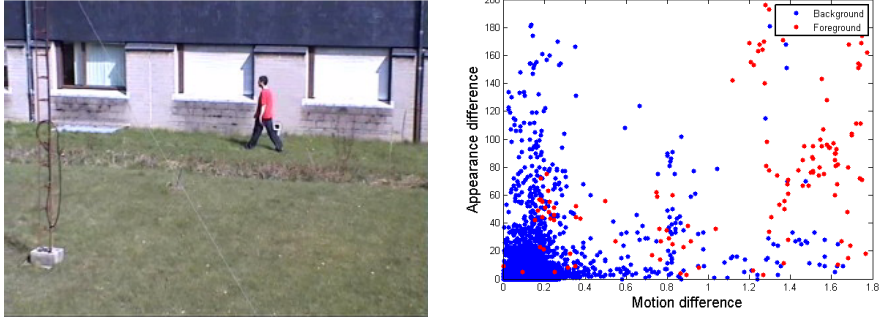


Figure 1.2: Appearance difference and motion difference to the background model.

majority of the outlier background pixels exhibit the property that one of the two features (appearance or motion) is correct while the other is wrong. In other words, a foreground pixel can be determined if and only if it is different from background from the perspective of both motion and appearance. This strategy is effective to reduce false positives, but rather "conservative", i.e., it will generate false negatives if the difference to background model from either cue is not significant enough. Considering spatial information around the neighbourhoods and the spatial constraint on pixel labels are useful to reduce false negatives while maintaining low false alarm rate. Particularly, we propose a robust and real-time background subtraction approach which considers the marginal statistical models of appearance and motion separately and also the spatial constraints in a Markov Random Field framework (MRF) - solved via the graph-cut algorithm, to reduce the false positives at the minimum expense of recall. The extensive experiments show that our approach is superior to other competing background subtraction algorithms designed for a pan-tilt camera.

### 1.3 Organization

The rest of the thesis is organized as follows. In Chapter 2, a review of background subtraction for both static and pan/tilt camera is given. Descriptions of background modelling and object detection of our approach are discussed in detail in Chapter 3. Chapter 4 provides both the qualitative and quantitative results of the experiments on different datasets and the comparison with different background subtraction methods. Our conclusions are given in Chapter 5.

# Chapter 2

## Related Work

In this chapter we present the background and related work of the thesis, which includes background subtraction from a static camera and a pan-tilt camera. Then we illustrate the motivation of our approach which designs a novel and robust background subtraction method to detect objects from a pan-tilt camera.

### 2.1 Introduction

Background subtraction is a group of algorithms used to detect moving objects in video sequence from a static camera. It is one of the basic low-level operations in many computer vision applications, including human-computer interaction [35], video surveillance [11, 50], robotics [20, 43] and traffic monitoring [9, 25]. Intuitively, moving objects differentiate from background in terms of both appearance and motion. Assuming that the background is static, background subtraction "subtracts" the observed image from the estimated background, i.e., objects that are static, and areas with large differences are referred to as the foreground, i.e., the moving objects (Figure 2.1(a)).

Traditional background subtraction approaches assume that the camera is static, otherwise the new field-of-view becomes unmatched with the prebuilt background model (Figure 2.1(b)). This assumption severely limits its usage in the moving platforms, such as cell phones, PTZ cameras, etc. Efforts have been made to extend background subtraction to moving cameras by compensating the camera motion. The consecutive frames, including the background model image, are registered to the same coordinate system, thus conventional background subtraction approaches can be applied to detect the moving objects. Precise image registration is not trivial and pixels can be misaligned. False alarms usually occur on the misaligned pixels which are compared to the unmatched background models, deteriorating the performance of background subtraction. Post-processing can be applied to filter out the noises, yet it does not work effectively when the size of the noises are comparable to that of moving objects.

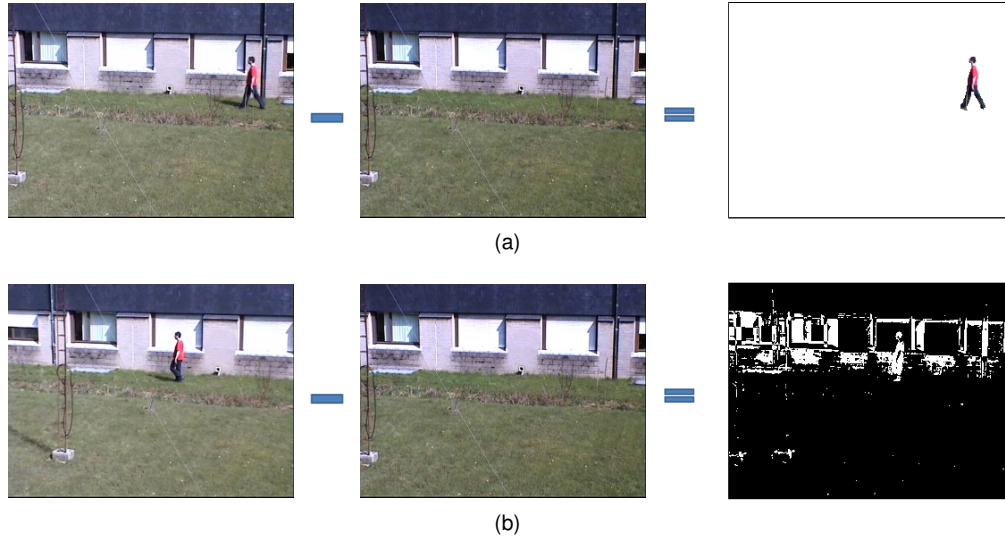


Figure 2.1: Illustration of traditional background subtraction approaches and their limitation on the moving platform.(a) An example of a traditional background subtraction approach; (b) An example when background subtraction fails (the camera pans to the left). From left to right of both (a) and (b): the current frame, the estimated background, the detection.

This problem can be alleviated by incorporating motion information via optical flow estimation, assuming that background pixels are static after camera motion is compensated whereas the foreground pixels move. Meanwhile, motion can also provide higher discriminative power than the conventional appearance-based background subtraction approaches which are sensitive to illumination changes. However, pixel wise motion estimate can be inaccurate, and together with the inaccuracies from image registration, may contribute to the false detections.

In this thesis, we propose a novel background subtraction method that evaluates the marginal probabilities of motion and appearance and considers the spatial constraints existing in pixel labels of motion and appearance, and also in neighbouring pixels, to suppress most of the false positives at the minimum expense of miss detections. Section 2.2 gives a literature review of background subtraction from a static camera, while in Section 2.3 introduces the extension of background subtraction to a moving camera.

## 2.2 Background Subtraction from a Static Camera

The simplest way to model the background is to obtain a background image without any moving object. This approach will only work when all background pixels are static. Considering robustness and adaptability to complex background, repetitive motion from clutters and long-term/short-term scene changes, numerous background models have been proposed which can be divided into two categories: a) parametric background modelling [27,

44, 57, 48, 25, 49] and b) non-parametric background modelling [7, 55, 53, 14, 36, 29, 24]. In this section, we present an overview of each category in background modelling from a static camera.

## 2.2.1 Parametric Background Modelling

If the background is nearly static and the variation of a pixel's value is mainly due to camera noise, it is natural to model each pixel as Gaussian distribution. Deriving from this assumption, parametric methods describe each pixel as a parametric distribution or a mixture of distributions, e.g., Kalman filter [25], Wiener filter [49], running Gaussian Average (single Gaussian) [52], and Gaussian Mixture Model (GMM) [44, 57].

### Kalman Filter

A Kalman filter-based adaptive background model [25, 41] allows the background estimate to evolve as the illumination changes gradually. For each frame, segmentation is performed by thresholding the difference  $D_t$  between the current frame and the background model  $B_{t-1}$ . The background is updated at each frame by involving the previous estimated background model  $B_{t-1}$  and the difference  $D_t$  using the following update equation:

$$B_t = B_{t-1} + (\alpha_1(1 - M_t) + \alpha_2 M_t) D_t \quad (2.1)$$

where  $M_t$  is a binary mask of the current frame. The gains  $\alpha_1$  and  $\alpha_2$  are based on the estimates of the rate of change of background.

### Running Gaussian Average (Single Gaussian)

A comparable approach, running Gaussian average scheme [52] is proposed where the background for each pixel  $p$  at time  $t$  is modelled as a Gaussian distribution  $\mathcal{N}(\mu_{p,t}, \sigma_{p,t})$  with mean  $\mu_{p,t}$  and variance  $\sigma_{p,t}$ . Thus pixel  $p$  is classified as a foreground pixel if the inequality between its observation  $I_{p,t}$  at time  $t$  and its background model  $\mathcal{N}(\mu_{p,t}, \sigma_{p,t})$ :

$$|I_{p,t} - \mu_{p,t}| > k\sigma_{p,t} \quad (2.2)$$

holds; otherwise,  $p$  will be classified as background; For the background pixel  $p$ , the background model is updated as:

$$\mu_{p,t} = \alpha I_{p,t} + (1 - \alpha)\mu_{p,t-1} \quad (2.3)$$

where  $\alpha$  is an empirical weight as a trade-off between stability and quick update. Kalman filtering and running Gaussian Average have the advantages of low computation and low memory requirement. Their learning rates enable the background model to adapt to gradual lighting changes. However, they can't handle dynamic background with more than one Gaussian distribution.

## Gaussian Mixture Model (GMM)

To account for complex backgrounds containing more than one Gaussian distribution, [44] models each pixel as a mixture of  $K$  Gaussians corresponding to either background or foreground. The probability of the occurrence of a current pixel is:

$$P(\mathbf{I}_{p,t}) = \sum_{i=1}^K \omega_{i,p,t} * \eta(\mathbf{I}_{p,t}; \mu_{i,p,t}, \Sigma_{i,p,t}) \quad (2.4)$$

where  $\eta(\mu_{i,p,t}, \Sigma_{i,p,t})$  is the  $i^{th}$  background Gaussian model and  $\omega_{i,p,t}$  its weight. Pixel values that do not fit the background distributions are considered as foreground until there is sufficient and consistent evidence to initiate a new Gaussian to support them. The background Gaussians can be determined in terms of its persistence and the variance which can be measured by  $\omega/\sigma$ . This value increases both as a distribution gains more confidence and more persistent. After ordering the Gaussians by  $\omega/\sigma$ , the first  $B$  distributions are chosen as the background model, where

$$B_{p,t} = \arg \min_b \left( \sum_{i=1}^b \omega_{i,p,t} > T \right) \quad (2.5)$$

where  $T$  is a measure of the minimum portion of the data that should belong to background. Thus  $I_{p,t}$  is labeled as background if it is within 2.5 standard deviation of a background Gaussian model.

GMM has gained vast popularity [30, 38, 28, 31]. Yet [14] points out that it fails to achieve sensitive detection in the case where the background has very high frequency variations such as waving water or shaking tree leaves, i.e., background having fast variations cannot be accurately modelled with just a few Gaussians. Another important point is its ability to adapt to sudden change in the background which depends on the learning rate. Low learning rate is suitable for long-term change but it has a poor adaptivity to sudden change. High learning rate can adapt to changes quickly, but slowly moving objects can be easily incorporated into background.

## Summary

Parametric background modelling fits a probability density function that can be modelled parametrically (e.g. Gaussian distribution) on the previous background observations to adapt to the changes of background. The pixel-wise background model refers to only a few parameters which guarantees low time complexity and memory load. However, the natural scene in the real world is usually complex that cannot be modelled by just some specific distributions. Another limitation is that most of the parametric background modelling methods only consider each pixel independently which rely heavily on post-processing to filter out the

noises. This problem becomes serious in background subtraction from a moving camera where the size of a false positive caused by pixel misalignment is comparable to a moving object. It would be a dilemma to trade off between miss-detections and false-detections. For parametric approaches, a training process is required to learn the parameters which may not satisfy the need of quick initialization and quick object detection in the new field of view when the camera moves. Furthermore, most of the parametric background modelling approaches are appearance-based only, which are sensitive to appearance changes due to camera distortion, weather, etc.

## 2.2.2 Non-parametric Background Modelling

In most of the time, scenes are complex that cannot be described by parametric methods. To improve the robustness of background subtraction towards complex background, non-parametric background modelling approaches estimate the density function directly from the data without any underlying distribution assumptions. Histogram over time [55], kernel density estimation [13], background clustering including support vector machines [29, 1] and codebook [53], graph cut [10, 22, 46], and ViBe [3] are some of the non-parametric background modelling methods.

### Histogram over Time

An approximation of the background probability density function can be given by the histogram of the most recent values classified as background values [55]. However, as the number of samples is necessarily limited, it only has probability density function (pdf) on those background values existed in the observing window, but miss the "tails" of the true pdf [37]. Another drawback is the lack of convergence to the right density function if the dataset is small. Besides, it is not suitable for higher dimensional features.

### Kernel Density Estimation (KDE)

Unlike histograms, kernel density estimation [13] guarantees a smooth, continuous and differentiable density estimate to any density shape with no assumption on the underlying distribution (Fig 2.2 (a)). Different from Gaussian Mixture Model, kernel density function does not need the number of Gaussian distributions. With kernel density estimators learned from the previous samples, KDE can handle shaking videos caused by camera jittering or dynamic background such as waving trees, fountains/falls [15, 47, 12]. Besides, KDE can be performed in a higher-dimensional space integrating appearance and motion features for the purpose of modelling the background density [13]. In general, the probability density function can be given as a weighted sum of kernels in terms of each sample  $x_{i,p}$  for pixel



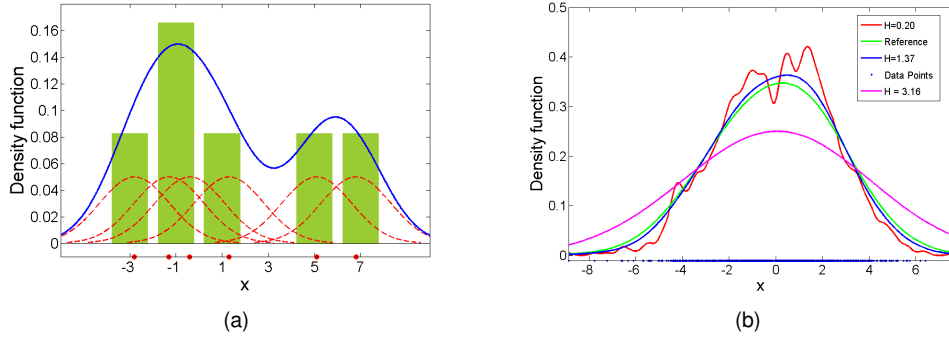


Figure 2.2: Illustration of kernel density estimation. (a): The comparison of the construction of histogram and kernel density estimators using six data points (red dots on the  $x$ -axis). The horizontal axis is divided into bins covering the range of the data. A box of height of  $1/12$  is placed whenever a data point falls inside this interval. For the kernel density estimate, a normal kernel (a red dash curve) is placed on each of the data point. The final kernel density estimate is the blue curve. (b): KDE with different bandwidths of a random sample with 1000 points from a standard normal distribution. Green: the true density (standard normal). Red: KDE with  $H = 0.2$ ; Blue: KDE with  $H = 1.37$ ; Magenta: KDE with  $H = 3.16$ .

$p, i = 1..n$ :

$$P(\mathbf{I}_{\mathbf{p},t}) = \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{I}_{p,t} - \mathbf{x}_{i,p}) \quad (2.6)$$

where  $K_H(\mathbf{x}) = \|\mathbf{H}\|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$ . The bandwidth  $\mathbf{H}$  which is the smoothness parameter, specifies the "width" of the kernel around each samples point  $\mathbf{x}_{i,p}$ . An important issue which exhibits a strong influence on the resulting estimate is the kernel bandwidth selection. Because of the limitation of the number of samples and the real-time computation requirement, the choice of appropriate bandwidth is essential which is shown in Fig 2.2 (b). Small values of  $\mathbf{H}$  make the estimate look "wiggly" and show spurious features (red curve), whereas to big values of  $\mathbf{H}$  lead to an over-smooth estimate that is too biased and may not reveal structural features [51] (magenta curve). Another problem for kernel density estimation is the memory load to keep those samples and the computational cost. What's more, the appearance of the moving objects in the training phase can bias the probability density estimation and thus may cause miss detection.

### Background modelling via Clustering

Object detection can also be regarded as clustering problem, and a number of background subtraction algorithms based on clustering have been proposed during the past 20 years [29, 8, 7, 24, 53]. Here we introduce SVM and codebook which are used widely in background subtraction.

- **Support Vector Machines (SVMs)**. Generally speaking, SVMs [29] separate the

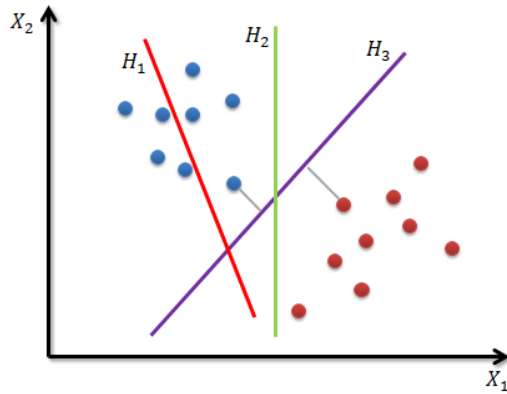


Figure 2.3: Illustration of support vector machine.  $H_1$  does not separate the classes.  $H_2$  does but the margin is small.  $H_3$  is an appropriate hyperplane with the maximum margin.

data points with a hyperplane or set of hyperplanes in a high- or infinite-dimensional space constructed by maximizing its functional margin to any class, given that a larger margin indicates lower generalization error of the classifier (Fig 2.3). Training two-class SVM needs data points for both background and foreground. However, collecting samples for foreground can be rather expensive, or just impossible as foreground has a wide range of variance. One-class SVM [8] can handle background/foreground segmentation with the absence of foreground training data. One important issue of SVM is the selection of the kernel bandwidth, but little guidance for choosing an appropriate bandwidth is available.

- Codebook.** Apart from SVM, foreground can also be extracted by applying k-means clustering [7], sequential clustering [53], and codebook [24] by quantizing the background into clusters/codewords. For example, in [24], each pixel is attached with a adaptive and compressed background model encoded with a set of codewords learning from the previous frames. Each codeword defines a vector of features describing some statistics, such as color value, minimum and maximum pixel brightness, and codeword access frequency [24]. A thresholding operation is performed based on the distance of the occurrence from the nearest codeword. modelling the background with codebook is capable of coping with dynamic background such as shaking leaves, for each object that has been seen frequently in the pixel will be encoded with a codeword. It is also robust in dealing with illumination changes. Unlike kernel density estimation, it allows moving foreground objects in the scene during the initial training period. Furthermore, it can model a background from a long training sequence. The limitation is that a codebook needs a long training sequence which may not be reuseful in other sequences.

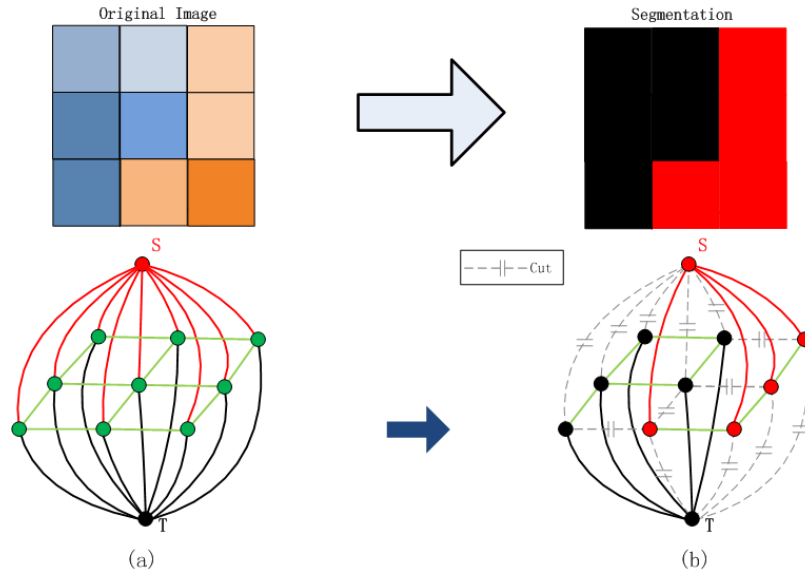


Figure 2.4: A standard graph designed for background subtraction

### Graph Cut

Background/Foreground segmentation can be formulated in terms of energy minimization integrating regional bias and spatial dependence. Such energy minimization problems can be reduced to instances of the maximum flow problem in a graph which can be solved by graph cut model. Specifically, Fig 2.4 shows a typical graph derived from an image. Each pixel corresponds to a node, and all pixel nodes are connected to two special terminals, one is the source  $s$  representing foreground, and one is the sink  $t$  for background. The weights of the links between pixel vertices and  $s/t$  derive directly from the similarity between the current frame and the foreground/background model where the background can be built with the existing background modelling methods. The neighbour links are assigned weights in terms of the difference of the neighbouring pixels. Once a graph is constructed, a global optimal (minimum cost) cut is achieved by energy minimization framework separating the source from the sink. The graph-based approach can overcome the effects of noise by aggregating the neighbourhood information. Other priors can also be imposed when constructing a graph, such as motion prior [10], shape prior [19], etc.

### Visual Background Extractor (ViBe)

Most of these methods construct background models based on temporal information where the first  $n$  frames are used to train the parameters in the group of parametric methods or collect the background samples in the group of non-parametric methods. Moving objects are not allowed to appear in the training phase otherwise it will bias the background model.

Instead of training the background models with the temporal sequences, ViBe model proposed in [3] claims that the probability density function learned from temporal sequences can be replaced by spatial distribution, and thus the background model can be built within one frame and no training phase is needed. The underlying assumption is that those neighbouring pixels should share similar appearance and variations. In this approach, a pixel-wise background model is constructed by sampling  $N$  data points randomly from its connected neighbourhood, and the moving objects is detected if the appearance is close to sufficiently enough samples in its background model. ViBe is a parameter-free method with low computational load and the background model can be constructed instantly. Though it is a simple technique, it is robust to noise and able to deal with dynamic background.

### **Summary**

Compared to parametric methods, the non-parametric approaches may need more memory to keep the information that is representative to the probability density function, which may also require more computation. Nevertheless, non-parametric background modelling can handle complex and dynamic background where parametric methods fail easily. Besides, it is more straightforward to incorporate different features, e.g., contrast and motion, in background model to improve the robustness. Among the non-parametric methods, graph cut not only take the regional bias but also consider boundary property to reduce noises while ViBe takes the advantage of spatial similarity of neighbouring pixels to speed up the background modelling. These two properties enable the background subtraction algorithm to reduce noises caused by pixel misalignment and adapt the background model to the new field of view. In our approach, we incorporate these two methods and design a robust real-time background subtraction approach to detect moving objects with a moving camera.

## **2.3 Background Subtraction with a Pan-Tilt Camera**

The limitation that traditional background subtraction can only be applied to a static camera severely restricts its applications to moving platforms. This problem arises as the new field-of-view becomes unmatched with the prebuilt background model when the camera moves. Camera motion can be compensated via image registration, which aligns the consecutive frames and the background model image in the same coordinate system, thus the conventional background subtraction approaches can be applied to detect the moving objects. Ideally, pixels are aligned properly and thus background subtraction functions well. However, ideal image registration is non-trivial and pixel misalignment exists when pixels are incorrectly aligned due to geometric distortion, illumination changes and environmen-

tal noises. As a result, those misaligned pixels are compared to unmatched background models and become false positives.

In this section, we will first briefly introduce how pixels can be aligned before applying background subtraction (Section 2.3.1), and then discuss how people deal with pixel misalignment which is a common and important issue when extending background subtraction algorithms to a pan-tilt camera (Section 2.3.2).

### 2.3.1 Image Registration

Image can be registered globally or locally. In global image registration, all pixels are transformed with the same transformation matrix while in local image registration, the transformation is pixel-dependent.

#### Global Image Registration

Global image registration assumes that two images taken from two different point of views can be aligned by a  $3 \times 3$  transformation matrix  $\mathbf{H}$ , which is called homography. To be specific,  $\mathbf{H}$  defines the camera motion - rotation, translation and scale change - thus a 3D vector  $\mathbf{x}'$  can be projected to a new point  $\mathbf{x}$  rigidly by

$$\mathbf{x}' = \mathbf{H}\mathbf{x} \quad (2.7)$$

Generally speaking, this assumption holds in a pinhole camera model where the depth of objects in the scene are not changed, and thus the outliers caused by parallax-translation problem can be reduced. In other words, global image registration can be used in one of the following two cases:

- the optical centre does not move, i.e., pure rotation and possibly change of camera settings, or
- the views scene is planar, i.e., all objects are in the same depth.

The linear transformations are global in nature, thus, they cannot model local geometric differences between images. The accuracy of global registration methods proves insufficient when the underlying assumptions are invalid, e.g., the depth variations of the background objects, which often happens. However, due to its good approximation and fast computation, global image registration is usually used for aligning images before applying background subtraction ([45, 5, 2, 42, 39, 21, 32]).

#### Local Image Registration

Local image registration allows "elastic" or "nonrigid" transformation, which is capable of locally warping the target image to align with the reference image by modelling local geo-

metric differences between images. In local approaches ([4, 26]), an Expectation-Maximum (EM) framework is usually employed to iteratively switch between fine image registration (in pixel level) and background subtraction to detect the moving objects until the process converges. One drawback of the local approaches is that it is susceptible to reach a local minima due to the iterative nature, which gives rise to pixel misalignment. Besides, local image registration may be more sensitive to noises than global image registration, such as local illumination changes.

### **2.3.2 Background Subtraction Approaches with a Pan-Tilt Camera**

Existing background subtraction methods from a static camera can be performed normally after registering the target frame to the reference frame. For example, [45] performed background subtraction through temporal median filter after global image registration, while [5] applies a AR filter to background modelling. As image registration may introduce outliers, the performance of background subtraction approaches may be deteriorated by pixel misalignment compared to the situation in a static camera, thus the conventional background subtraction approaches should be improved to deal with these outliers.

Existing methods to improve the robustness of background subtraction algorithms can be divided into two categories according to the features they use: one is based on appearance cue only, the other is based on both motion and appearance cues. In the following, we will discuss each category in detail.

#### **Background Subtraction with Appearance Only**

Most of the appearance-based background subtraction approaches proposed to deal with pixel misalignment focus on making use of spatial information.

One way to use spatial information is to locate the most appropriate background model from the neighbourhoods. The assumption is that the correct alignment should exist in the surrounding neighbourhoods, so do the appropriate background models. Thus, a background pixel can be determined if it matches with one of the background models from its neighbourhoods. [39] proposed a Spatial Distribution Gaussian model (SDG) based on this assumption whereas [32] also proposed a similar background subtraction method to accommodate small alignment errors.

The main drawbacks of this strategy are from two perspectives. One is that the performance of the background subtraction methods heavily depend on the range of the searching region; Searching within a small region can locate small misalignments but fail to deal with pixels that are misaligned outside this region, whereas searching in a large region will severely slow down the detection process. The other drawback is that the searching nature can bring in miss detection if the appearance of a foreground pixel accidentally falls into

the background distribution of a nearby position. In other words, small moving objects may be ignored and discarded.

Another way to make use of the spatial information is to assume that the observation of a pixel is the convolution of several pixels from its surroundings as the camera moves, thus it can be modelled as a random process which was generated by a mixture of  $K$  random processes from the surrounding pixels[21]. In this way, foreground pixels can be detected by thresholding the PDF of the so-called Mixel Distribution (MD). This approach is claimed to handle various sources of error, including motion blur, sub-pixel camera motion, mixed pixels at object boundaries, and also uncertainty in background stabilization caused by noise, unmodelled radial distortion and small translations of the camera.

### **Background Subtraction with Appearance and Motion Cues**

Background Subtraction approaches that incorporate motion and appearance cues can be divided into three categories: a) motion is used as sparse labelling ([43, 16]), b) Motion is modelled jointly with appearance cue ([33, 18]), and c) Motion and appearance cues are incorporated but used separately ([56, 17]).

- **Motion cue is used as sparse labelling ([43, 16]).** The intuition is that the movements of moving objects and the background objects obey different geometric constraints that are useful to tell them apart. A sparse-to-dense background subtraction framework is usually employed. In these approaches, motion is used for separating background and foreground at the sparse level which is useful for later dense segmentation since the BG/FG movement trajectories lie in different spaces. Thus, the sparse labelling offers information for further dense segmentation. In [43], sparse appearance model is built and used for dense (pixel-level) segmentation, while in [16] dense appearance model is built which is propagated from sparse appearance model and used for dense segmentation.

In [43], the dependence on long-term trajectories for sparse modelling may make it discard the new-incoming moving objects or objects with short-term trajectories. Also sparse modelling may fail to detect small objects as well. [16] overcomes these problems by using different geometric constraints and dense appearance modelling. Even though, both methods may generate false positives when the orthographic projection assumption is not satisfied since the background objects with large depth variation might not be separable to the foreground objects using these geometric constraints. Meanwhile, both algorithms only model the appearance cue that is sensitive to illumination change, which is a common problem of appearance-based background subtraction approaches.

- **Motion is modelled jointly with appearance cue ([33, 18]).** The assumption is that the background pixels (the same for misaligned background pixels) are static after compensating the camera motion, whereas the foreground pixels move. Thus motion and appearance can be incorporated to provide higher discriminative power. One way is to model them jointly. For example, support vector machine ([18]) and kernel density estimation ([33]) model the background incorporating both appearance and motion cues via kernel functions.

The effectiveness of incorporating motion cue is illustrated through a simple example in Figure 2.5. Background can be modelled using Gaussian mixture model (GMM) - due to Stauffer and Grimson [44] - in which given the appearance  $I$  of a pixel at location  $(x, y)$  of the image, its probability of being background is described as:

$$P(I_t(x, y)) = \sum_{i=1}^K w_{i,t} \mathcal{N}(I_t | \mu_{i,t}, \Sigma_{i,t}) \quad (2.8)$$

where  $K$  is the number of Gaussians in the mixture weighted by  $w_{i,t}$ ,  $\mu_{i,t}$  and  $\Sigma_{i,t}$  are their means and covariance matrices. For simplicity of discussion and without loss of generality, we use a single Gaussian to describe the background model, and thus given the appearance  $I$ , its probability of being background can be described as:

$$P(I) = \mathcal{N}(I | \mu_I, \Sigma_I) \quad (2.9)$$

Background pixels should be static after camera motion compensation, and thus, their motion  $M$  can also be modelled as a single Gaussian distribution with mean  $\mathbf{0}$ . As motion and appearance cues are independent to each other, the joint probability is computed as:

$$P(I, M) = P(I)P(M) = \mathcal{N}(I | \mu_I, \Sigma_I) \mathcal{N}(M | \mu_M, \Sigma_M) \quad (2.10)$$

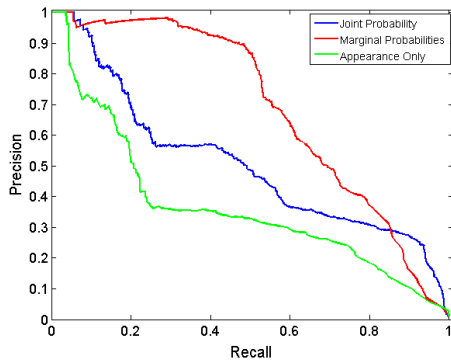
Without loss of generality, we simplify the appearance model by using gray scale and the motion model by using motion magnitude in the toy example. Instead of learning pixel-dependent variations, and in order to illustrate the idea, we compute a global variation for appearance and motion cue respectively after removing the outliers. If motion and appearance are modelled jointly, the classification of a pixel is based on thresholding  $P(I, M)$  or evaluating the exponent of the Gaussian probability in Eq.( 2.10). That is, a pixel is background if

$$(I - \mu_I)^T \Sigma_I^{-1} (I - \mu_I) + (M - \mu_M)^T \Sigma_M^{-1} (M - \mu_M) \leq Th \quad (2.11)$$

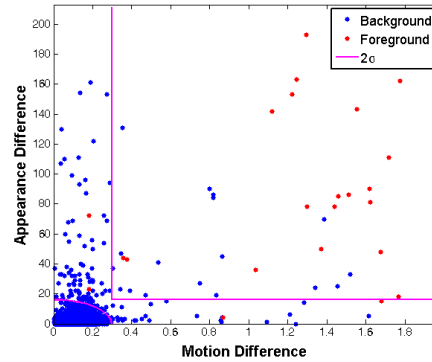




(a)



(b)



(c)

Figure 2.5: A simple example: illustrate two types of outliers that background subtraction approaches may fail to deal with: a) the outliers from pixel misalignment produced by imprecise image registration; b) the outliers from inaccurate motion estimate if motion is incorporated.

where  $Th$  is a threshold.

The Precision-Recall (PR) curve from Fig 2.5 (b) demonstrates that incorporating motion information with appearance information (blue curve) can reduce false positives effectively in terms of its higher precision compared to the appearance-only approach.

However, motion estimate can be inaccurate. These inaccuracies, together with the inaccuracies from appearance cue caused by imperfect image registration, may deteriorate the performance of background subtraction if motion and appearance cue are modelled jointly. To be more specific, a background pixel that exhibits a difference appearance from its model due to pixel misalignment or large motion due to incorrect motion estimate, corresponding to either small  $P(I)$  or small  $P(M)$ , result in a small  $P(I, M)$ , and thus may become a false detection. This is proved by the simple example we discuss above (Fig 2.5 (c)), where some background pixels (blue points) still fall outside the  $2\sigma$  decision boundary, which indicate that these pixels may become false positives.

- **Motion cue and appearance use are incorporated but used separately ([56, 17]).**

The intuition is that only those detected pixels share uniform motion ([17]) or relatively large motion magnitude ([56]) are considered as foreground otherwise they should be the noises.

Both methods use appearance cue for detection and motion cue for filtering out the noises from detection. Compared to those approaches that model motion and appearance jointly, these approaches prefer incorporating motion and appearance cues separately, which can separate the outliers of each cue and reduce the chance of producing false positives. These approaches are similar to modelling the marginal probabilities of these two cues to enhance the power of each cue to filter out the outliers. Returning to the simple example where appearance and motion is modelled with Gaussian distribution, the procedure of evaluating the marginal probabilities of appearance and motion for pixel labelling can be generalized as:

```

if  $(I - \mu_I)^T \Sigma_I^{-1} (I - \mu_I) \leq Th_I$  or
 $(M - \mu_M)^T \Sigma_M^{-1} (M - \mu_M) \leq Th_M$  then
    pixel  $\leftarrow$  background
else
    pixel  $\leftarrow$  foreground
end if

```

Such a classification procedure corresponds to the rectangular decision boundary in Fig 2.5 (c), where its PR curve (red curve) demonstrates that this strategy can suppress the false positives and outperforms the other two approaches.

However, as these approaches consider pixels individually, they may be too conservative and thus lower the number of retrieved foreground pixels, i.e., the recall rate. For example, a pixel with inaccurate motion estimate could become false negative when the motion estimate is incorrectly small though it may look different from the background appearance model. In fact, neighbouring pixels should share similar motion or appearance as long as they are not on the borders (of the moving objects), and thus should be given the same label.

## 2.4 Summary

In this chapter, we briefly review the literature of background subtraction from a static camera and a pan-tilt camera. Parametric background modelling is not suitable to model complex background. Non-parametric background modelling methods may require higher memory than parametric background modelling, but it can handle complex and dynamic

background where parametric methods fail easily. Most of these approaches only consider pixels individually, which make them sensitive to noises. This problem can be solved by defining a graph model where spatial constraints can be imposed to filter out the noises and maintain good recall rate. Existing background subtraction methods that relax the camera stationary assumption face the problem of pixel misalignment when compensating the camera motion. This problem can be alleviated by taking the advantage of the spatial information which is used commonly in appearance-based background subtraction approaches. They work well on dealing with small misalignments but fail when the misalignments are large. Motion can be incorporated with appearance which is useful to suppress the false detections. It can be used to model the background jointly with appearance, but the inaccuracies from motion, together with inaccuracies from appearance, may contribute to the false positives. A conservative rule which models the marginal probabilities of motion and appearance is robust to these two types of outliers, but may generate miss detections when inaccuracies from either cue occurs. The strength of filtering false positives by exploiting marginal probabilities of motion and appearance, and also the benefit of considering different spatial constraints via graph model, motivate us to design a novel background subtraction method, that evaluates the marginal properties of appearance and motion through a two-layer graph model with imposed spatial constraints on pixel labels. More details are discussed in Chapter 3.

# Chapter 3

## Methodology

In Chapter 2, we introduced the state-of-the-art background subtraction approaches from a static camera and a pan-tilt camera. Existing methods that extend background subtraction to a moving platform are vulnerable to pixel misalignment when registering consecutive images or motion inaccuracies if motion is incorporated to classify foreground and background pixels. In this chapter, we introduce a novel background subtraction algorithm which is robust to these two types of outliers at the minimum expense of recall. To be more specific, we evaluate the marginal probabilities of appearance and motion cues and impose spatial constraints to reduce false positives at the minimum expense of false negatives via a two-layer graph model. In the following sections, we present the graph model in detail.

### 3.1 Introduction

Background subtraction detects the moving objects (foreground) that are significantly different from the background where motion can be incorporated with appearance in background modelling to provide higher discriminative power. Motion inaccuracies can deteriorate the performance of background subtraction when motion estimate is incorrectly large for a background pixel leading to a false detection or when motion estimate is incorrectly small for a foreground pixel which may become a miss-detection. Image registration is necessary for compensating the camera motion so that conventional background subtraction can be performed. Yet perfect registration is not trivial and misaligned pixels may turn out to be false detections when they are compared to unmatched background models. As these two types of outliers tend not to co-occur simultaneously, background subtraction approaches that marginally model appearance and motion cues can reduce many false positive foreground pixels committed by approaches that jointly model these two cues. Yet this decision rule might be conservative towards the foreground pixels and give rise to false negatives when foreground pixels look similar to background or move slowly. This classification procedure can be implemented as two one-layer graph models, one of which models appear-

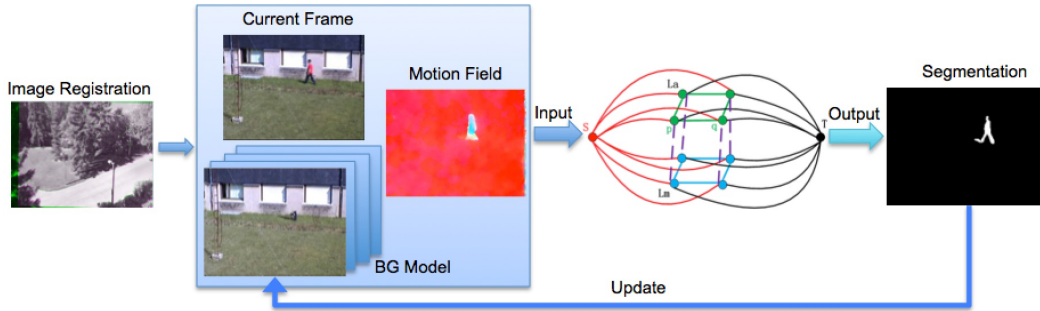


Figure 3.1: The main steps of our approach

ance, the other models motion, and thus the detection result can be achieved by anding the graphs' labelling solved by graph cuts. Yet this decision rule might be conservative towards the foreground pixels as it consider pixels individually, and give rise to false negatives when foreground pixels look similar to background or move slowly. In fact, labelling from the two graphs should satisfy some topological constraints. Intuitively, a pixel should achieve identical labels from the perspective of both appearance and motion. Moreover, neighbouring pixels should also share the same label if they are similar to each other. These constraints should also be considered when labelling pixels. Instead of building two one-layer graph models, we exploit the marginal probabilities of appearance and motion via a two-layer graph model where two types of spatial constraints are imposed to reduce false positives at the minimum expense of false negatives. The main steps of our approach include: image registration, background modelling, graph construction and graph cut (Figure 3.1).

## 3.2 Image Registration

Image registration is the process of estimating an optimal transformation between two images, where a transformation can be rigid or non-rigid. Particularly, for a pinhole camera model, an image can be registered to another through a homography matrix  $H$  if any of the two cases is satisfied:

- the optical centre does not move (pure rotation and possibly change of camera settings), or
- the viewed scene is planar.

$H$  can be estimated as similarity transformation, affine transformation, projective transformation, etc., in terms of the degrees of freedom that are considered. Considering model complexity and accuracy, affine transformation which considers rotation, translation and scale change, is often used as a satisfactory approximation of the projective camera model,

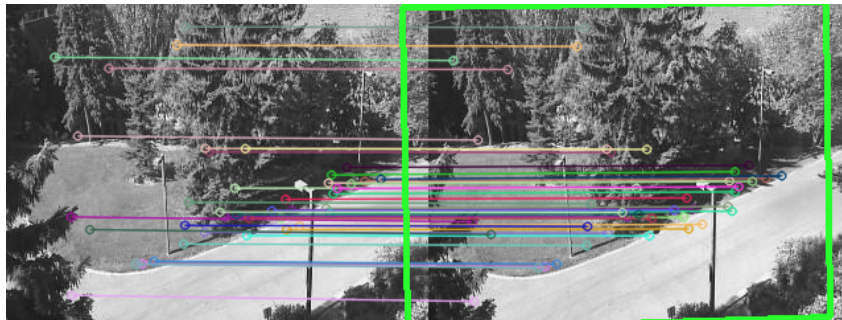
provided that the background can be well approximated by a plane or where the camera motion is restricted to pan, tilt, or zoom, i.e., camera movements without change of optical centre. In our experiments, the camera is hinged at a fixed point, and thus we use affine transformation model to spatially register the images (and also the background models) .

Feature-based image registration is used widely to estimate the homography  $H$  by establishing a correspondence between a number of especially distinct points in images. A transformation is determined to map the target image (the previous frame) to the reference image (the current frame) based on the correspondence between a number of points in images. Here, the feature correspondences can be extracted by Kanade Lucas Tomasi (KLT) tracker which estimates the motion of a feature iteratively. KLT derives from two assumptions: one is the brightness constancy constraint, i.e., the brightness does not change a lot within a period of time; the other is common motion heuristic, i.e., neighbouring pixels move with the same speed. As features can be mismatched if the iterative framework falls into a local solution due to clutters in the appearance, cross-cutting trajectories and scene parallax, we apply RANSAC to determine a set of inliers so that the homography can then be estimated in an optimal fashion from these matched feature correspondences. The mapping function is constructed with the selected feature correspondences and the target frame is transformed through the estimated  $H$  and overlaid over the reference image with interpolation methods such as nearest neighbour function. A simple example of feature based image registration is illustrated in Figure 3.2.

Perfect image registration is not trivial and pixel misalignment is generated from two perspectives. One is the parallax problem due to depth variance, i.e., an apparent displacement of difference of orientation of an object viewed along two different lines of sight. Nearby objects have a larger parallax than further objects when observed from different positions, thus a rigid transformation is inappropriate to compensate displacements in terms of different depths yet a non-rigid transformation needs heavy computation which may be impractical in real-time. Another reason that gives rise to pixel misalignment comes from the assumption of the interpolation process, i.e., the smoothness of the neighbourhood pixels, which does not hold on the edges or object boundaries. The consequence is that the misaligned pixels are compared to the unmatched background model resulting in false detections. A further refinement after image registration can reduce small misalignments (here we use nearest neighbour search to locate the most matched background model), even though, it is not effective when the camera keeps moving.



(a)



(b)

Figure 3.2: Feature-based global image registration by applying Kanade-Lucas-Tomasi Tracker for tracking detected features (features matching) and RANSAC to reduce the features that are outliers. (a) Left: the target frame. Right: the reference frame which is shifted to the right. (b) Matched feature points. Green rectangle: the location of the target frame in the reference frame.

## 3.3 Background Modelling

Background modelling here includes the initialization and maintenance of appearance model and motion model which will be discussed in detail in the following sections.

### 3.3.1 Appearance Model

The basic principle of background subtraction is to compare the background model with the current frame in order to detect regions referred to as the foreground where a significant difference occurs. Generally, most background subtraction approaches build a background model based on a sequence of frames. It makes sense from a statistical point of view that the temporal distribution of the background pixels is a good estimation of the background model. Yet constructing the background model temporally does not satisfy the need of quick initialization and quick object detection in the new field of view when the camera moves. Meanwhile, pixels are considered independently for maintaining and updating their background models based on the observations. Thus individual pixels affected by perturbations, e.g., lens distortion, are easily misclassified. In return, these approaches heavily rely on the post-processing algorithms to clean up these misclassifications. Instead, neighbouring pixels belonging to the same cluster should follow similar variations over time, and thus background subtraction algorithms should also consider the spatial information of neighbouring pixels.

#### Visual Background Extractor(ViBe)

Visual Background Extractor (ViBe) is a universal non-parametric background subtraction technique that has a quick initialization of background model and high resilience to noises by exploiting the spatial distribution of pixels. The idea of ViBe comes from an assumption that the neighbouring pixels share a similar temporal distribution which has been proved in [23]. Particularly, ViBe substitutes the estimation of temporal distribution with the estimation of spatial distribution which enables it to initialize the background model within one frame.

- **Background Model Initialization**

To be more specific, the temporal distribution of the background model for each pixel can be approximated by sampling from its neighbourhoods in the first frame, i.e., its background can be modelled non-parametrically with a sets of samples collected from its neighbourhoods. Mathematically, the background model of pixel  $p$  can be formulated as:

$$B_p = \{s_j | j \in N_G(p)\} \quad (3.1)$$

where  $N_G(p)$  defines the neighbourhood of  $p$ .



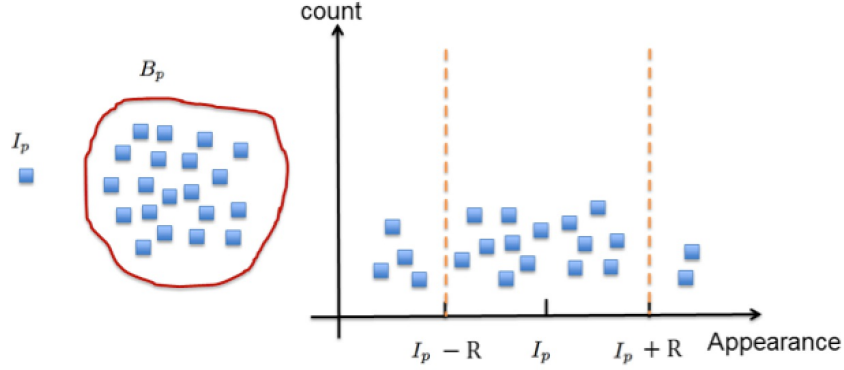


Figure 3.3: Background/Foreground Classification with ViBe

Compared to parametric approaches that rely on probability density functions (pdfs) or statistical parameters, ViBe does not require a statistical estimation of the pdfs as it is a sample-based approach and it is free from the problem of finding an appropriate shape for the pdfs which might be sensitive to outliers.

- **Background/Foreground Classification**

Normally, pixel  $p$ , with its appearance  $I_p$  (here we use the RGB colour space) can be classified as background if the probability of belonging to background is high. In ViBe, this probability can be measured as the number of votes that  $I_p$  is close to its collection of background samples  $B_p$ . To be more specific, the label of  $p$  depends on whether the number of its background samples that are within a distance  $R$  to  $I_p$  is larger than or equal to a given threshold  $\#_{min}$  (Figure 3.3). Mathematically, the label  $f_p$  (BG or FG, which are the abbreviations of background and foreground) can be determined by:

$$f_p = \begin{cases} \text{BG,} & \#\{|s_j - I_p| < R \mid j \in N_G(p)\} > \#_{min} \\ \text{FG,} & \text{otherwise} \end{cases} \quad (3.2)$$

The accuracy of the ViBe model is determined by two parameters only:  $\#_{min}$  and  $R$ . With the assumptions that a) there is no moving object in the field of view when initializing the background model, b) only background pixels should populate the background model and c) the spatial distribution collected from its neighbourhoods simulates the temporal distribution, the underlying idea is that it is more reliable to estimate the statistical distribution of a background pixel with a small number of close values than a large number of samples. The closeness can be measured by the distance threshold  $R$  and the confidence can be measured as  $\#_{min}$  in a sense of voting.

- **Background Model Maintenance**

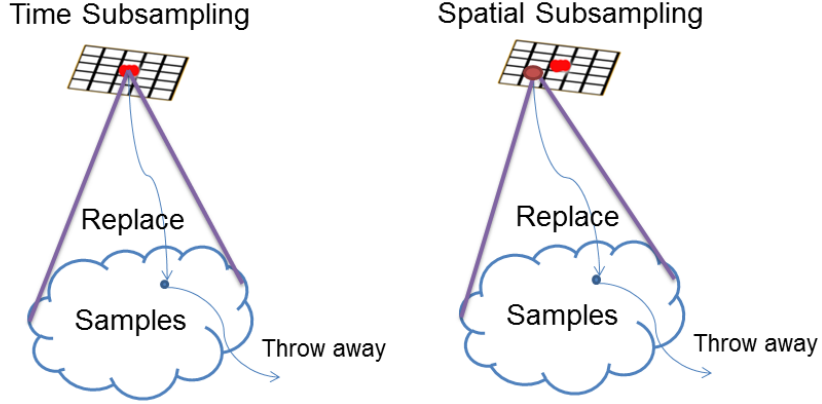


Figure 3.4: Background Model Update Scheme of ViBe

Updating the background model is a crucial step to adapt to lighting changes and handle new objects that appear in the scene. The update scheme should extend the time window covered by the background models and ensure the spatial consistency by propagating background samples spatially. Here ViBe adopts a memoryless update policy.

Put it simply, the update scheme of ViBe includes time subsampling and spatial subsampling. Time subsampling extends the size of the time window covered by a pixel model of a fixed size by randomly replacing one of the background samples for pixel  $p$  with its observation  $I_p$  in the current frame. Spatial subsampling maintains the spatial consistency of the background model by randomly replacing one of the background sample for pixel  $p$  with one of its neighbourhoods which is selected randomly (Figure 3.4).

### Usage of ViBe in Our Graph Model

To initialize the background model, we borrow the idea of ViBe benefiting from its quick background model initialization and low time complexity. Instead of classifying pixels by comparing the observation and the background model which might not be good at dealing with inaccuracies of image registration, we leave the classification to graph cut which can integrate appearance and motion cues and incorporate some topological constraints to refine the segmentation. In appearance layer which evaluates the appearance model, we compute the confidence of belonging to its background as the similarity of pixel  $p$  to its closest background sample. Mathematically,  $Pr(I_p|BG)$  is defined as:

$$Pr(I_p|BG) = \max_{s_j \in \mathbf{B}_p} \exp\left(-\frac{(I_p - s_j)^2}{2R^2}\right) \quad (3.3)$$

Thus we can compute the probability of a pixel belonging to foreground in the appear-

ance layer as  $Pr(I_p|FG)$ :

$$Pr(I_p|FG) = 1 - Pr(I_p|BG) \quad (3.4)$$

The background model is updated using the same mechanic of ViBe.

### 3.3.2 Motion Modelling

In the motion layer, a background pixel should be static after compensating the camera motion, while a foreground pixel, i.e., a pixel belonging to a moving object, has a non-zero motion. Considering the noise of motion estimation, the background motion model can be formulated as Gaussian distribution  $\mathcal{N}(\mathbf{m}_i; \mathbf{0}, \Sigma)$  with mean  $\mathbf{0}$  and the covariance matrix  $\Sigma$ . For computational reason, the covariance matrix is assumed to be of the form:

$$\Sigma = \sigma^2 \mathbf{I} \quad (3.5)$$

This assumes that the noises from  $x$  and  $y$  directions are independent and have the same variances.

Thus in the motion layer, given the motion  $\mathbf{m}_p$  of pixel  $p$ , we can evaluate the probability belonging to background  $Pr(\mathbf{m}_p|BG)$  and that belonging to foreground  $Pr(\mathbf{m}_p|FG)$  as:

$$Pr(\mathbf{m}_p|BG) \sim \mathcal{N}(\mathbf{m}_p; \mathbf{0}, \Sigma), \quad (3.6)$$

$$Pr(\mathbf{m}_p|FG) = 1 - \mathcal{N}(\mathbf{m}_p; \mathbf{0}, \Sigma) \quad (3.7)$$

To approximate the motion field, we compute the dense optical flow for each pixel after compensating the camera motion. Given two frames  $I_0$  and  $I_1$  which are registered on the same coordinate system, the objective of optical flow estimation is to find the motion field  $\mathbf{m}$  such that the image-based error after shifting  $I_0$  with  $\mathbf{m}$  can be minimized together with a regularization force. Formally,  $\mathbf{m}$  can be solved by minimizing

$$\min_{\mathbf{m}} \int_{\Omega} \{\lambda |I_0(\mathbf{x}) - I_1(\mathbf{x} + \mathbf{m}(\mathbf{x}))| + |\mathbf{m}|\} dx \quad (3.8)$$

where  $\lambda |I_0(\mathbf{x}) - I_1(\mathbf{x} + \mathbf{m}(\mathbf{x}))|$  is the image data fidelity, and  $|\mathbf{m}|$  depicts the regularization term.  $\lambda$  balances the weights between the data fidelity and the regularization force.

The pixel-wise optical flow can be computed through TV-L<sup>1</sup> flow algorithm [54] based on total variation (TV) regularization and the robust L<sup>1</sup> norm in the data fidelity term, which can preserve discontinuities in the flow field and offer an increased robustness against illumination changes, occlusions and noises. This method results in an efficient numerical scheme which is based on a dual formulation of the TV energy and which can be speeded up by graphics processing units (GPU).

Even though, perfect optical flow estimate is not trivial and it can be affected by camera noises, quantization noise, weak texture, reflections and illumination changes. Meanwhile,

in the case of a pan-tilt camera, the estimation computed based on the current frame and the warped frame, can also be affected by pixel misalignment ending up with imprecise motion estimation on the misaligned pixels. Inaccurate motion estimation generates false positives and false negatives. In the first case, the motion estimate of background pixels may be imprecisely large (e.g., blue dots along  $x$  axis with values much larger than zero in Figure 1.2 (c)) which might be excluded from the decision boundary and turn out to be false detections. In the second case, miss detections happen when the motion estimate of foreground pixels are imprecisely small (e.g., some red dots with motion estimate close to zero in Figure 1.2 (c)) and thus might be falsely labeled as background pixels.

## 3.4 Graph Construction and Graph Cuts

In Section 3.4.1, we introduce the graph cut algorithm. We then define our graph model in detail in Section 3.4.2.

### 3.4.1 Graph Cuts Algorithm

Background subtraction can be solved by graph cuts efficiently and effectively which not only consider regional properties, i.e., the individual penalties for assigning pixel  $p$  to foreground and background, but also integrate some topological constraints, e.g., spatial constraints or temporal constraints. The latter property is of great importance to reduce false labelling on those individual pixels that are affected by noises. Graph cut approaches are robust and reliable as they provide global optimal solutions. Furthermore, graph cuts can also be applied to solve N-D problems [6].

Mathematically, both the regional properties and the topological constraints can be encoded in an energy function  $E(f)$  which combines of the data term  $E_{data}(f)$  and the smoothness term  $E_{smooth}(f)$ ,

$$E(f) = E_{data}(f) + E_{smooth}(f) \quad (3.9)$$

where labelling  $f = \{f_p | p \in \mathcal{P}\}$  defines a segmentation for an arbitrary set of data elements (pixels or voxels)  $\mathcal{P}$ .

The data term  $E_{data}$  defines the individual penalties of assigning pixel  $p$  to either foreground (FG) or background (BG) and can be modelled as the negative log-likelihoods of given foreground and background models.  $E_{smooth}$  captures the "boundary" properties interpreted as the penalties for discontinuity between neighbouring pixels where a variety of topological constraints can be imposed. Such energy minimization problems can be further formulated into a maximum flow problem in a graph which can be solved efficiently by graph cuts.

Suppose  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is an undirected graph with a set of nodes  $\mathcal{V}$  and a set of undirected edges  $\mathcal{E}$ . Each pixel in the image corresponds to a node in the graph and there are two more special vertices (terminals), namely, the source  $s$  and the sink  $t$ . In the background subtraction problem,  $s$  is usually referred to as the foreground terminal and  $t$  the background terminal. The set of edges  $\mathcal{E}$  consist of two types of undirected edges: n-links, i.e., links connecting neighbouring pixels, and t-links, i.e., links connecting to the terminals. Each node  $p$  has two t-links  $\{p, s\}, \{p, t\}$ , with non-negative costs, which encode the regional bias, representing dissimilarity to the corresponding foreground/background model. Each pair of neighbouring nodes  $p, q$  are connected by an n-link encoding the topological constraints. Usually, spatial constraint is imposed here assuming that similar neighbourhoods should be assigned with the same label. In this way, n-link is given the weight defined in terms of the similarity of neighbouring pixels.

Graph cut is applied here to find a cut  $C$  with the smallest cost, which is equivalent to computing the maximum flow from the source to the sink. Such a cut  $C$  defines a partition that segments  $\mathcal{P}$  to two disjoint sets  $S$  and  $T$  such that  $S$  is referred to as background cluster while  $T$  as the foreground cluster (Figure 2.4).

### 3.4.2 A Two-layer Graph Model

To incorporate motion information and appearance information, we define a two-layer graph which is robust towards two types of outliers illustrated in Figure 1.2.

To be more specific, given a set of pixels  $\mathcal{P}$ , we construct a two-layer graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  with a set of nodes  $\mathcal{V}$  and a set of undirected edges  $\mathcal{E}$  (Figure 3.5). Each pixel  $p$  in  $\mathcal{P}$  is mapped to two corresponding nodes in the graph, one (green node) in the appearance layer  $L_a$  and the other (blue node) in the motion layer  $L_m$ . In this graph,  $\mathcal{E}$  consists of two types of undirected weighted edges: a) t-links (red and black edges) which connect each node in two layers to the terminals,  $S$  and  $T$ , encode the regional properties, i.e., how pixels fit into the appearance model / motion model, b) n-links, consist of two types of edges, each of which encodes one topological constraint that will be discussed later. The first type of edges connect neighbouring nodes in each layer (blue and green edges), while the second type of edges connect two corresponding nodes in two layers (purple edges). Given a labelling  $f$  over the two-layer graph model, the final labelling would be achieved by anding the labels of corresponding nodes in two layers.

As we have two layers in the graph,  $E_{data}(f)$  now composes of two summation terms, i.e.,

$$E_{data}(f) = \sum_{p \in L_a} D_p^a(f_p) + \sum_{p \in L_m} D_p^m(f_p) \quad (3.10)$$

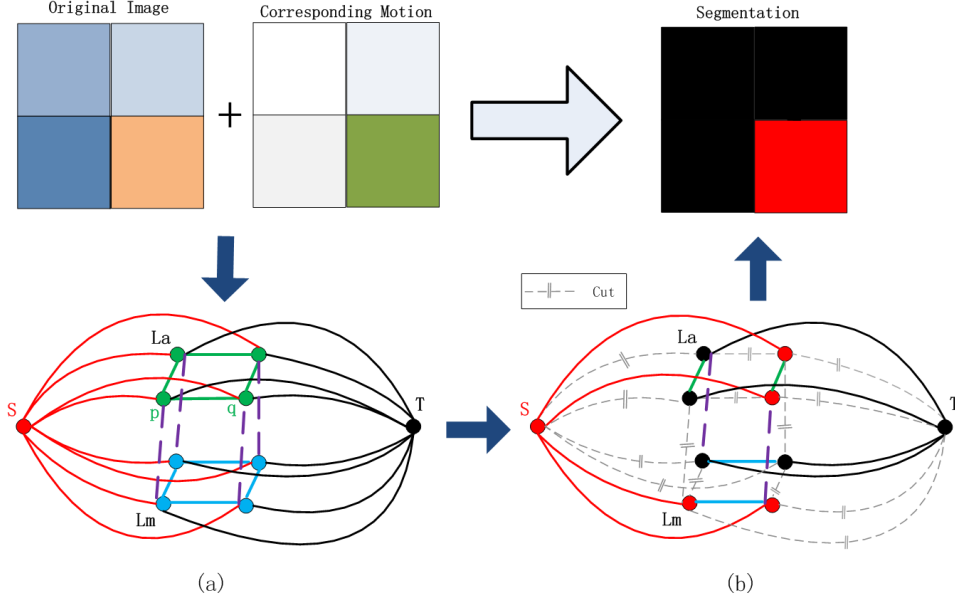


Figure 3.5: The two-layer graph model of our proposed approach.

$D_p^a(f_p)$ , which corresponds to the weight of a t-link in appearance layer, estimates the regional bias of the given appearance model in ratio of the negative log-likelihoods which is originally motivated by the MAP-Markov Random Field formulation, while  $D_p^m(f_p)$  corresponding to the weight of a t-link in motion layer, is defined in a similar fashion in terms of the motion model.

In our graph model, we consider two types of constraints: label consistency constraint and spatial coherence constraint.

- **Label consistency constraint**

This constraint implies that the labels of two corresponding nodes in two layers should be identical. Intuitively, the appearance of foreground should be different from the learned background appearance model according to the assumption of background subtraction. Meanwhile, the moving objects have their own motion patterns which also deviate significantly from the background. Label consistency constraint is also satisfied for a background pixel which is similar to its background appearance model and whose motion is small compared to moving objects after image registration. To impose label consistency constraint, we connect each node in each layer with its spatial neighbourhoods in the other layer by vertical links. In this thesis, we use five-connected neighbourhoods, i.e., each node in one layer is connected to the node in the same position and nodes that are spatially on the left, right, up and down from the other layer. For simplicity, in Figure 3.5, we only illustrate the vertical connections but not the connections to other spatial neighbourhoods. A large constant

weight  $\tau$  is given to these vertical links to prevent a cut through these links so that the corresponding nodes receive the same label. One advantage of label consistency constraint is that we can lower the chance of miss-detections. Two-layer graph model without label consistency constraint will increase miss-detections caused by either small motion or insignificant appearance difference to background model. Imposing label consistency constraint can push a cut to group the two corresponding nodes to the foreground cluster if the other cue has a strong confidence to be foreground, in return, maintaining a high recall.

- **Spatial coherence constraint**

This constraint derives from the pairwise Markov property, i.e., a variable is conditionally independent of all other variables given its neighbours. In graph cut, it can be interpreted as any two neighbouring pixels with high similarity should keep the same label while the discontinuity should be preserved. As we have two layers, the spatial coherence constraint can be measured in terms of the similarity of two neighbouring pixels with respect to its appearance in  $L_a$  or its motion in  $L_m$  and imposed through n-links. Imposing spatial coherence constraint can smooth out the noises caused by small pixel misalignment and motion inaccuracies, thus in return, maintaining a high precision.

In the two-layer graph model, the smoothness term can be extended as:

$$E_{smooth}(f) = \sum_{p,q \in \mathcal{N}^a} V_{p,q}^a(f_p, f_q) + \sum_{p,q \in \mathcal{N}^m} V_{p,q}^m(f_p, f_q) + \sum_{p,q \in \mathcal{N}^{a,m}} V_{p,q}^{a,m}(f_p, f_q) \quad (3.11)$$

The first two terms encode spatial coherence constraint which is imposed through n-links while the last term encodes label consistency constraint which is imposed through vertical links.  $\mathcal{N}^a, \mathcal{N}^m, \mathcal{N}^{a,m}$  contain unordered pairs of neighbouring nodes in  $L_a, L_m$ , and between two layers respectively.

In a word, the weights of edges in  $\mathcal{E}$  are summarized in Table 1, where  $\sigma_a$  and  $\sigma_m$  are two constants,  $dist(p, q)$  defines the distance between  $p$  and  $q$ ,  $\delta_{f_p \neq f_q}$  is a delta function which returns 1 if  $f_p \neq f_q$ , otherwise 0 if  $f_p = f_q$ .

### 3.5 Summary

In this Chapter, we introduce a background subtraction method which is robust to outliers of appearance and motion cues while maintaining high recall. A two-layer graph model

Table 3.1: Weights definition of the edges in the two-layer graph model  $\mathcal{G}$

edge	notation	weight(cost)
$\{p, S\}$	$D_p^a(f_p = BG)$	$-\ln \Pr(I_p   BG)$
	$D_p^m(f_p = BG)$	$-\ln \Pr(\mathbf{m}_p   BG)$
$\{p, T\}$	$D_p^a(f_p = FG)$	$-\ln \Pr(I_p   FG)$
	$D_p^m(f_p = FG)$	$-\ln \Pr(\mathbf{m}_p   FG)$
$\{p, q\}$	$V_{p,q}^a(f_p, f_q)$	$e^{-\frac{(I_p - I_q)^2}{2\sigma_a^2}} \text{dist}(p, q) \delta_{f_p \neq f_q}$
	$V_{p,q}^m(f_p, f_q)$	$e^{-\frac{(\mathbf{m}_p - \mathbf{m}_q)^2}{2\sigma_m^2}} \text{dist}(p, q) \delta_{f_p \neq f_q}$
	$V_{p,q}^{a,m}(f_p, f_q)$	$\tau \delta_{f_p \neq f_q}$

is proposed with one layer evaluating the appearance model and one layer evaluating the motion model. Each pixel is mapped to two corresponding nodes, one in each layer, thus the evaluation of belonging to background/foreground in terms of appearance and motion cues can be exploited individually. This graph model can be solved efficiently by graph cut algorithms. The final labelling can be achieved by anding the results from two layers based on the assumption that foreground should be different from background with respect to both appearance and motion. This strategy can effectively reduce the outliers caused by pixel misalignment and motion inaccuracies. To reduce false detections at the minimum expense of miss detections, we consider two types of constraints. One is label consistency constraint by introducing a vertical n-link between the spatial neighbouring nodes of the two layers, inferring the labelling in two layers should be identical; the other is spatial coherence constraint by adding edges between neighbouring pixels within each layer, inferring that the labels of similar neighbouring pixels should be the same. The background appearance model can be built within one frame by substituting temporal distribution of spatial distribution which enables it to detect objects immediately when the camera moves to a new position. The background/foreground motion model can be built assuming that the background motion is zero after compensating the camera motion. To demonstrate the effectiveness of our approach and the usefulness of the imposed constraints, experimental results on comparing different state-of-the-art background subtraction methods are shown quantitatively and qualitatively in Chapter 4.



## Chapter 4

# Experimental Results and Discussion

In this chapter, we conduct three sets of experiment. The first set of experiment is to evaluate the importance of motion cue when applying background subtraction in a moving platform via comparing the performance of [40] and [56]. Both use GMM to model background, the difference is that [40] only relies on appearance cue while [56] incorporates motion cue to filter out the appearance outliers. [56] and our approach derive from the same assumption that the foreground objects should be different from the background in terms of both appearance and motion cues and the rest should be the background. Yet as we discussed in chapter 2 through a simple example, [56] only considers pixels individually thus may result in false negatives. Different from [56], our approach concerns the spatial information of neighbouring pixels and also the constraints existing in corresponding pixel labels in terms of motion and appearance via a two-layer graph model. The second set of experiment is to evaluate the effectiveness of these spatial constraints we impose to reduce false negatives while maintaining high recall. The last set of experiment is to illustrate which way is better to exploit motion and appearance cues in background modelling, either jointly ([18]) or marginally (our approach). The implementation of our approach which runs in real-time, uses the GPU and the NVIDIA CUDA framework.

Six competing video sequences with resolution of  $240 \times 320$  at 30 frames per second were used in order to analyze the performance of the background subtraction approaches in different environments. These six video sequences can be divided into two categories: three video sequences are taken outdoors and the other are taken indoors. Among these six video sequences, two outdoor sequences are from [3], and the rest are taken from a pan-tilt-zoom camera (AXIS Q6034-e PTZ Dome Network Camera) at a fixed point.

The detection results are presented qualitatively and quantitatively. The parameters for each algorithm were determined experimentally. For each sequence, several representative frames, the ground truth and detection results produced by each algorithm are

presented. The detection results are shown as black and white images where white pixels represent foreground objects while black pixels represent background. The performance of each approach is also evaluated quantitatively using a) the traditional pixel wise evaluation metrics (precision, recall, F-measure) which are used commonly in evaluating background subtraction approaches and b) the component-based evaluation metrics which are designed from the perspective of object detection, here we use the correct detection rate, miss detection rate and false alarm rate defined in [34]. The extensive experiments show that our solution is superior to the competing background subtraction algorithms designed for dealing with a pan-tilt camera.

Section 4.1 overviews the video sequences we use. In Section 4.2, we will introduce the evaluation metrics. A brief introduction of the state-of-the-art algorithms that we compare with is given in Section 4.3. The qualitative detection results of the three set of experiments are shown in Section 4.4 while the quantitative evaluations are illustrated in Section 4.5.

## **4.1 Datasets**

Six video sequences involving camera panning and tilting (with three from outdoors and three from indoors) are used to evaluate the performance of our approach. The first two outdoor sequences are used in [3] while the rest are captured by the AXIS camera which is mounted in our lab. The frame rate of all of the three sequences is 30 frames per second.

### **4.1.1 Outdoors Sequences**

The first two sequences show a person walking along a terrace house captured by a tripod-mounted camera. The camera pans to the left in these two sequences while the person moves in different directions. The third sequence involves both panning and tilting movements which may result in depth variations in the field of view which is more challenging than the previous two sequences. Meanwhile, as the sequence was captured in a snowy weather leading to some untextured area in the camera view, it is more difficult to register images correctly, i.e., the detection of this sequence is more vulnerable to pixel misalignment and inaccurate motion estimate (Figure 4.1).

### **4.1.2 Indoors Sequences**

These three indoor sequences are captured with a pan-tilt-zoom camera mounted in our lab which moves under the same focal length with different speeds and different angles, meanwhile, the lighting conditions are also different. The background in these three sequences is complicated where image registration may be fragile to align those edges properly, meanwhile, the depth variation of background objects due to camera movements also gives rise



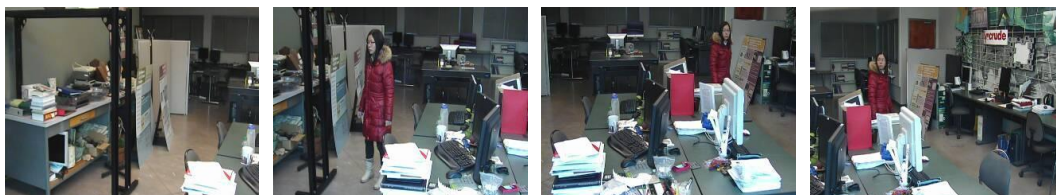
(a) Outdoor Sequence 1



(b) Outdoor Sequence 2



(c) Outdoor Sequence 3



(d) Indoor Sequence 1



(e) Indoor Sequence 2



(f) Indoor Sequence 3

Figure 4.1: The datasets used for evaluation

to pixel misalignment (Figure 4.1).

## 4.2 Evaluation Metrics

In this thesis, we use two types of measurements to evaluate the performance of different approaches, one defines in pixel-level, the other in component-level.

The first type of evaluation metrics defined in pixel level is the most direct measure which is used often to evaluate the performance of background subtraction approaches, including precision, recall and F-measure. They are defined as follows:

$$Precision = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \quad (4.1)$$

$$Recall = \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives} \quad (4.2)$$

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4.3)$$

These evaluation metrics measure the accuracy of the approach at the pixel level, however, in some cases, people are not interested in the detection of point targets but object regions instead. Thus, we also use the object-based evaluation metrics proposed in [34]. To be more specific, we consider three cases mentioned in [34] which are shown as follows:

- **Correct Detection (CD) or 1-1 match:** the detected region corresponds to one and only one ground truth region.
- **False Alarm (FA):** the detected region has no correspondence in the ground truth.
- **Detection Failure (DF):** the ground truth region is not detected.

According to the definitions, we need to determine the correspondence of the foreground region in the detection result and in the ground truth, i.e., whether the foreground region in the ground truth is matched with the segmentation. Based on the correspondences, we can evaluate a selected approach in terms of the correct detection rate, the false alarm rate and the detection failure rate.

Object matching can be defined by a binary correspondence matrix  $C^t$  that describes the correspondence of the detected components and the ground truth. The match can be determined by the overlap between the detected region and the ground truth region. Let us assume that we have  $N$  ground truth regions  $\tilde{R}_i$  and  $M$  detection regions  $R_j$ . Under these conditions  $C^t$  is a  $N \times M$  matrix, defined as follows

$$C^t(i, j) = \begin{cases} 1, & \text{if } \frac{\#(\tilde{R}_i \cap R_j)}{\#(\tilde{R}_i \cup R_j)} > T \\ 0, & \text{if } \frac{\#(\tilde{R}_i \cap R_j)}{\#(\tilde{R}_i \cup R_j)} < T \end{cases} \quad \forall i \in \{1, \dots, N\}, j \in \{1, \dots, M\} \quad (4.4)$$

where  $T$  is a threshold determining whether two regions are matched in terms of their overlap area.

According to the definition of correct detection, a correct detection event is associated to a one-to-one match between the  $i^{th}$  ground truth region and the  $j^{th}$  detected region if and only if  $\mathcal{C}(i, j) = 1$ . No other detected region is matched to the same ground truth region (otherwise indicating that the ground truth region is split in the detection) and no other ground truth region is matched to the same detected region (otherwise indicating the detected regions are merged). A false alarm event is associated to a detected region  $R_j$  if no match can be found in the ground truth regions. Respectively, a detection failure event corresponds to a ground truth region  $\tilde{R}_i$  if no detected region is matched to it. The three events can be defined mathematically with two auxiliary vectors shown as follows

$$L(i) = \sum_{j=1}^M \mathcal{C}(i, j) \quad i \in \{1, \dots, N\} \quad (4.5)$$

$$C(j) = \sum_{i=1}^N \mathcal{C}(i, j) \quad j \in \{1, \dots, M\} \quad (4.6)$$

$$\begin{aligned} \text{CD} \quad & \exists_i : L(i) = C(j) = 1 \wedge \mathcal{C}(i, j) = 1, \\ \text{FA} \quad & \exists_i : C(j) = 0 \\ \text{DF} \quad & \exists_i : L(i) = 0 \end{aligned} \quad (4.7)$$

The region-based measures described here depends on an overlap requirement  $T$  between the region of the ground truth and the detected region. A higher  $T$  indicates a more conservative evaluation rule. Here we set  $T = 10\%$  as [34] does.

In the evaluation, we use the percentage of each event for comparison. The percentage of correct detection, detection failures were obtained by normalizing the number of each type of event by the total number of moving objects in the whole sequence. The percentage of false alarms which is a number in the range 0-100% is defined by normalizing the number of false alarms by the total number of detected objects.

### 4.3 Approaches for Comparison

Our approach is compared to three state-of-the-art methods: [40], [18] and [56].

[40] proposes an improved GMM algorithm to deal with the outliers from image registration. The key idea is that the matched background model should exist in neighbourhoods, and thus a local search can locate the matched background model. A spatial distribution of Gaussians model is proposed to deal with moving object detection where the motion compensation is not exact but approximated. To be more specific, instead of comparing a pixel  $p$  with its mapped background model, [40] seeks a matched background model

among its neighbourhood which is effective to rectify small misalignment in image registration. However, as we will see later, the performance is sensitive to the size of the defined local region.

[56] observed that the motion information can be used for reducing the noises generated from appearance-based approaches. The intuition is that only those detected pixels sharing large motion magnitude can be considered as foreground otherwise they should be the background pixels. Here motion is used for post-processing, to filter out those detected blobs that do not move. In other words, motion and appearance cues are exploited separately. Theoretically, [56] can be approximated by evaluating the marginal probabilities of appearance and motion cues in background modelling, which can be implemented as a two-layer graph model without spatial constraints.

[18] incorporates motion cue and appearance cue through a Markov Random Field (MRF) graphical model where spatial constraints can be imposed for further refinement on the segmentation. To be more specific, a Gaussian kernel function is proposed to combine these two cues with a trade-off weight  $\lambda$  to balance their contributions. Formally, given the appearance observation  $I_p$ , the motion estimate  $\mathbf{m}_p$ , and the background model which is composed of a few background samples  $B_i(p), i = 1..n$  at pixel  $p$ , the Gaussian kernel function in terms of background is defined as:

$$F_B(p) = \sum_{i=1}^n B_i(p) \cdot \omega \cdot \exp^{-\frac{\|I_p - B_i(p).a\|^2 + \lambda \|\mathbf{m}_p - B_i(p).m\|^2}{2\sigma^2}} \quad (4.8)$$

Each background sample contains two components, i.e., the appearance component  $B_i(p).a$  and the motion component  $B_i(p).m$  in  $x$  and  $y$  directions, which are collected from the first  $n$  consecutive frames. The importance of each background sample is measured by  $B_i(p).\omega$ , which can be computed in terms of the sample's contribution to the Gaussian kernel function. The segmentation is obtained by applying graph cut on a Markov Random Field (MRF) graphical model where the input is the pixel-wise score map computed from the Gaussian kernel function.

As [18] models both appearance and motion jointly, it could be vulnerable to the inaccuracies from these two cues, and thus, may contribute to false positives.

## 4.4 Qualitative Results

In this section, we illustrate the performance of different background subtraction algorithms qualitatively of those three sets of experiments. Table 4.2 - Table 4.7 show the detection results over outdoors and indoors sequences. Six representative frames are selected from each sequence and shown in each table followed by their ground truths (GT). The results produced by background subtraction using appearance only ([40]), background subtraction

	Appearance	Motion	Marginally modelled	Jointly modelled	Spatial Constraints
GIR [40]	Y				
MBGS [56]	Y	Y	Y		
JBGS [18]	Y	Y		Y	Y
Ours	Y	Y	Y		Y

Table 4.1: Summarization of our approach and three other approaches for comparison. "Marginally modelled" and "Jointly modelled" refer to as the two ways of incorporating appearance cue and motion cue in background modelling.

of evaluating marginal probabilities of appearance and motion without considering spatial information ([56]), or evaluating the joint distribution of these two cues in an MRF framework ([18]) are illustrated in row four to six with the abbreviations GIR, MBGS, JBGS respectively. The performance of our approach is shown in the last row. Table 4.1 summarizes the characteristics in each method.

#### 4.4.1 Ex.1 - Appearance Only vs Appearance and Motion

Here, we compare the performance of [40] which only considers appearance cue and that of [56] which combines appearance and motion cues, i.e., modelling background with appearance and filtering out the noises with the use of motion cue (row "GIR" and row "MBGS" in Table 4.2 to Table 4.7). Several conclusions can be made from the detection results:

- Combining motion information is useful to reduce false positives caused by the outliers from image registration.** From the detection results of the six sequences, the spatial distributed Gaussian models proposed in [40] by searching among local neighbourhoods for matched background model may fail to label those misaligned pixels correctly. In fact, the performance of [40] heavily depends on the size of searching region. A small searching region will lower the chance of locating the match background model and thus increasing false positives while a large searching region will lead to high computation. The performance is worse when dealing with complex background where lots of edges occur (Table 4.5 to Table 4.7). Combining motion information, on the other hand, can effectively and robustly reduce these outliers as their motion is usually separable from foreground pixels.
- Combining two cues separately without considering spatial constraints is conservative with respect to the foreground pixels.** Accurate motion estimation through dense optical flow algorithm is not trivial, i.e., the motion estimate of static background pixels could be incorrectly large or that of moving objects could be incorrectly small. The first case would possibly not affect the performance as pixel misalignment and

inaccurate motion estimate tend not to occur simultaneously, while the second case may increase the false negatives, i.e., labelling foreground pixels as background due to their small motion estimate. The detection results in the row of "MBGS" from Table 4.2 to Table 4.7 where some pixels are usually mis-detected in the detected objects also demonstrate that the strategy of [56] may be too conservative.

#### **4.4.2 Ex.2 - Importance of Spatial Constraints: with vs without**

To illustrate the importance of spatial constraints, we conduct the experiment of comparing [56] and our approach (row "MBGS" and row "Ours"). Both methods evaluate the marginal probabilities of appearance and motion cues, yet our approach takes two types of spatial constraints into account while [56] considers pixels individually. From the detection results, our approach can detect the moving objects with better silhouette and fewer holes, i.e., fewer false negatives.

More true positives pixels can be retrieved through imposing the two types of spatial constraints we define in Chapter 3. More specifically, label consistency constraint imposed through the high-cost vertical links can prevent a cut through the connection of two corresponding nodes, thus the same label can be given to these two nodes depending on their higher confidence of either labelling as foreground or background. This constraint can reduce the false negatives whose motion might be small but largely different from background appearance, or whose appearance is similar to background but who are actually moving. Spatial coherence constraint, can prevent a cut through similar neighbouring pixels which should share similar labels while preserving discontinuities by defining weights based on the similarities of the neighbouring pixels in terms of appearance layer and motion layer. Some false positives may be generated but, from Table 4.2 to Table 4.7, these constraints have much improved the recall at the minimum expense of precision.

#### **4.4.3 Ex.3 - Modelling Appearance and Motion: Jointly vs Marginally**

This set of experiment involves the comparison of background modelling with appearance and motion cues jointly (row "JBGS", [18]) and our approach (row "Ours"), which evaluates the marginal their statistical models separately. Both approaches consider spatial constraints through an MRF framework which is solved by graph cut algorithm, and the difference is the way of modelling appearance cue and motion cue: either jointly or separately.

Modelling motion and appearance jointly may be more sensitive to the outliers from appearance cue and motion cue than modelling them marginally.

- **Modelling appearance and motion cues jointly is vulnerable to the inaccuracies**



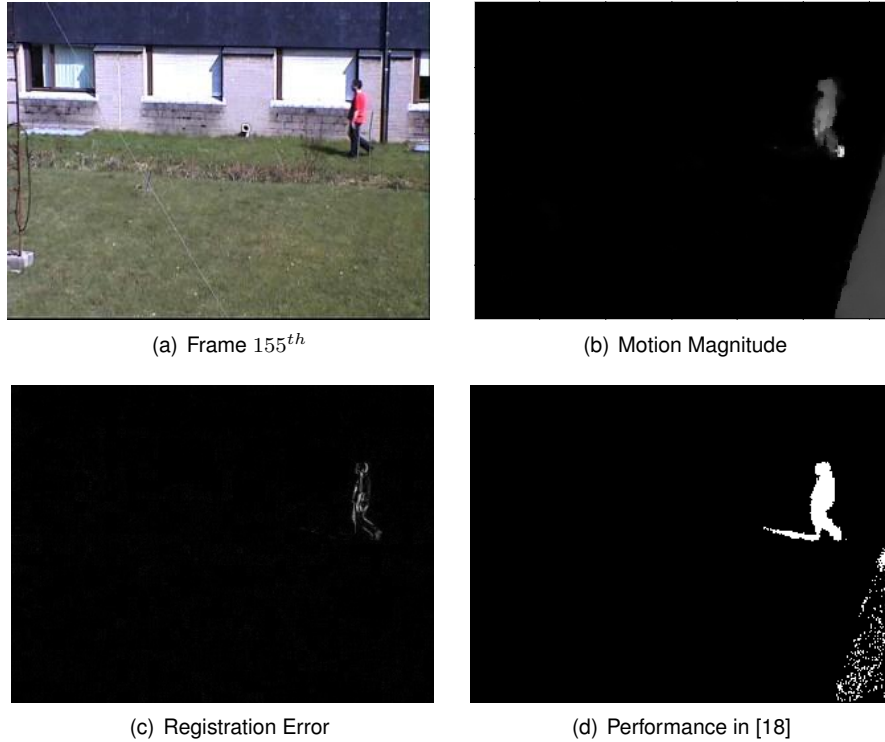


Figure 4.2: Illustration of the motion field of Frame 155<sup>th</sup> in outdoor sequence 1.

**of motion estimate.**

Motion estimate in the pixel domain may be unreliable because of the lack of the edge information, especially in untextured regions. Thus the inaccuracies in motion estimate will cause a bias in the Gaussian kernel estimation, which may produce false positives if the background motion estimate is incorrectly large (Table 4.2, frame 155<sup>th</sup>, 591<sup>st</sup> and 811<sup>st</sup>, Table 4.4 frame 181<sup>st</sup>) or false negatives if the foreground motion estimate is incorrectly small. Here a detailed explanation is given based on the frame 155<sup>th</sup> in Table 4.2. The pixel wise motion magnitude is shown in Figure 4.2, where the whiter the appearance, the larger the motion magnitude. We notice that the motion magnitude on the bottom right corner is comparable to walking person as they share similar appearance. However as we can see, this region is only part of the grass field, it should be static as other grass region does not move. Meanwhile from the image difference between the current frame and the wrapped frame (Figure 4.2(c)), we know that this region can be registered well using the KLT+RANSAC registration method, i.e., these false alarms are not produced by registration error. Thus, we can guess that these false alarms in Figure 4.2(d) result from the inaccuracies of motion estimate.

- **Modelling appearance and motion cues jointly is vulnerable to the inaccuracies**

**from appearance cue.**

These inaccuracies may come from reflection (Table 4.2, frame 1063<sup>rd</sup>), or pixel misalignments (Table 4.2 frame 221<sup>st</sup>, Table 4.3 frame 362<sup>nd</sup>, Table 4.4 frame 214<sup>th</sup>, Table 4.5 frame 345<sup>th</sup>).

As one cue can help rectify the error brought by the other cue, the performance can be improved by tuning the parameter  $\lambda$  of these two cues in the Gaussian kernels. However, it is hard to find a suitable  $\lambda$  to fit all of the situations. Our approach instead, is robust to these outliers by modelling appearance and motion cues separately, meanwhile, it can suppress the false negatives by considering the spatial information compared to [56].

## 4.5 Quantitative Results

In this section, we give the quantitative evaluation over the six video sequences (Table 4.8 to Table 4.13) at both the pixel level using precision, recall, and F-measure, and at the object level using correct detection (CD), false alarms (FA), and detection failure (DF).

First of all, in the pixel level evaluation, our algorithm enjoys similar recall compared to those competing approaches but a high precision which indicates much fewer false positives and thus a significant higher F-measure value. Second, in the component level evaluation, our approach shares similar correct detection (CD) rate, which is similar to recall but at the component level, and similar detection failure (DF) rate, i.e., the number of moving objects that our approach fails to detect. However, the false alarms (FA) rate is much lower than other approaches.

## 4.6 Summary

In this chapter, experimental results comparing our approach and other three competing background subtraction algorithms designed for a pan-tilt camera are presented. Experiments are conducted on both outdoors sequences and indoors sequences, both of which demonstrate that our approach outperforms among these algorithms and it is robust to the outliers from inaccurate motion estimate and pixel misalignment when registering consecutive images. The result of comparing the appearance-based approaches with that of incorporating motion and appearance demonstrates that, motion can provide higher discriminative power than using appearance cue alone, which can improve the robustness to the outliers from image registration, yet modelling motion and appearance cues jointly is vulnerable towards these outliers from either cue, since these outliers may be introduced into the joint kernel function, which will deteriorates its accuracy. Evaluating marginal probabilities is useful to deal with the outliers and provides higher precision, yet the recall may









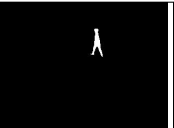









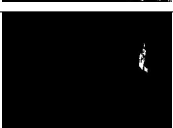
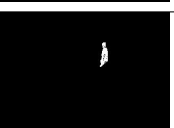
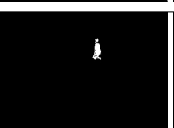
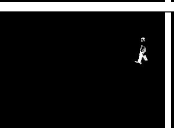
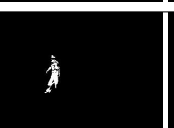



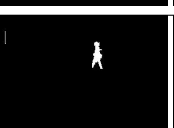




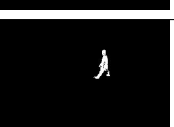
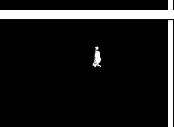



$j^{th}$	155 <sup>th</sup>	221 <sup>st</sup>	336 <sup>th</sup>	591 <sup>st</sup>	811 <sup>st</sup>	1063 <sup>rd</sup>
Img						
GT						
GIR						
MBGS						
JBGS						
Ours						

Table 4.2: Qualitative comparison on outdoor sequence 1.

be much lower since it may be overly conservative. The two types of spatial constraints we consider in this thesis (label consistency constraint and spatial coherence constraint) can alleviate this problem, and thus, by imposing these constraints through a two-layer graph cut model which models appearance and motion separately in two layers can effectively reduce the false positives at the minimum expense of miss detections.



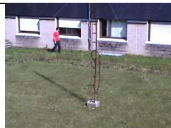
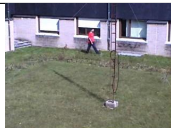
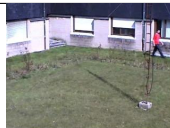

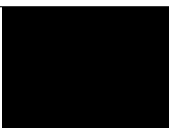



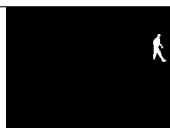
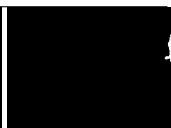

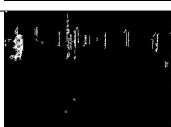

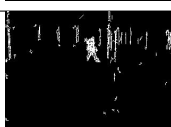


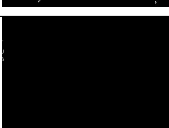


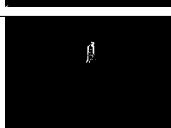



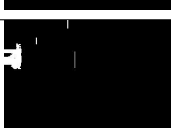



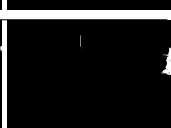

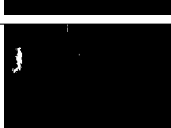
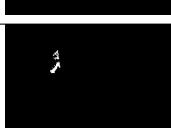



$i^{th}$	125 <sup>th</sup>	237 <sup>th</sup>	272 <sup>nd</sup>	305 <sup>th</sup>	362 <sup>nd</sup>	372 <sup>th</sup>
Img						
GT						
GIR						
MBGS						
JBGS						
Ours						

Table 4.3: Qualitative comparison on outdoor sequence 2.









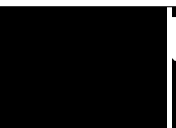



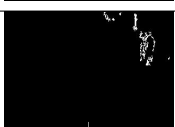
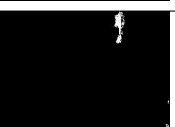





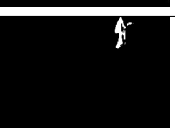
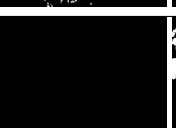










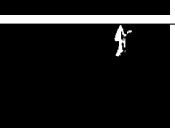




$i^{th}$	$36^{th}$	$74^{th}$	$150^{th}$	$181^{st}$	$214^{th}$	$250^{th}$
						
GT						
GIR						
MBGS						
JBGS						
Ours						

Table 4.4: Qualitative comparison on outdoor sequence 3.


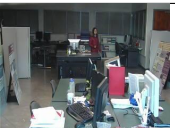
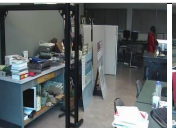


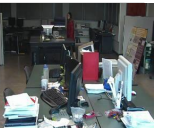









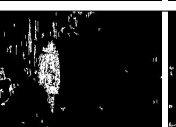




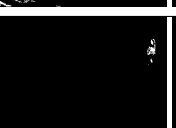
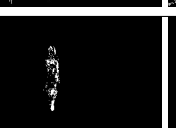














$i^{th}$	62 <sup>nd</sup>	97 <sup>th</sup>	191 <sup>st</sup>	345 <sup>th</sup>	505 <sup>th</sup>	983 <sup>th</sup>
Img						
GT						
GIR						
MBGS						
JBGS						
Ours						

Table 4.5: Qualitative comparison on indoor sequence 1.



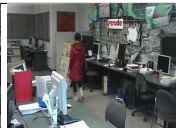

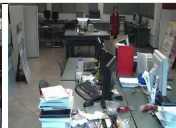
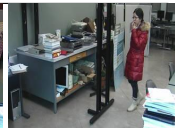









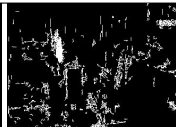





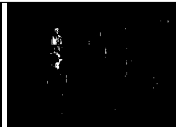











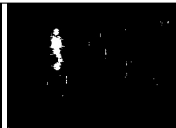


$i^{th}$	61 <sup>st</sup>	300 <sup>th</sup>	359 <sup>th</sup>	500 <sup>th</sup>	588 <sup>th</sup>	945 <sup>th</sup>
Img						
GT						
GIR						
MBGS						
JBGS						
Ours						

Table 4.6: Qualitative comparison on indoor sequence 2.

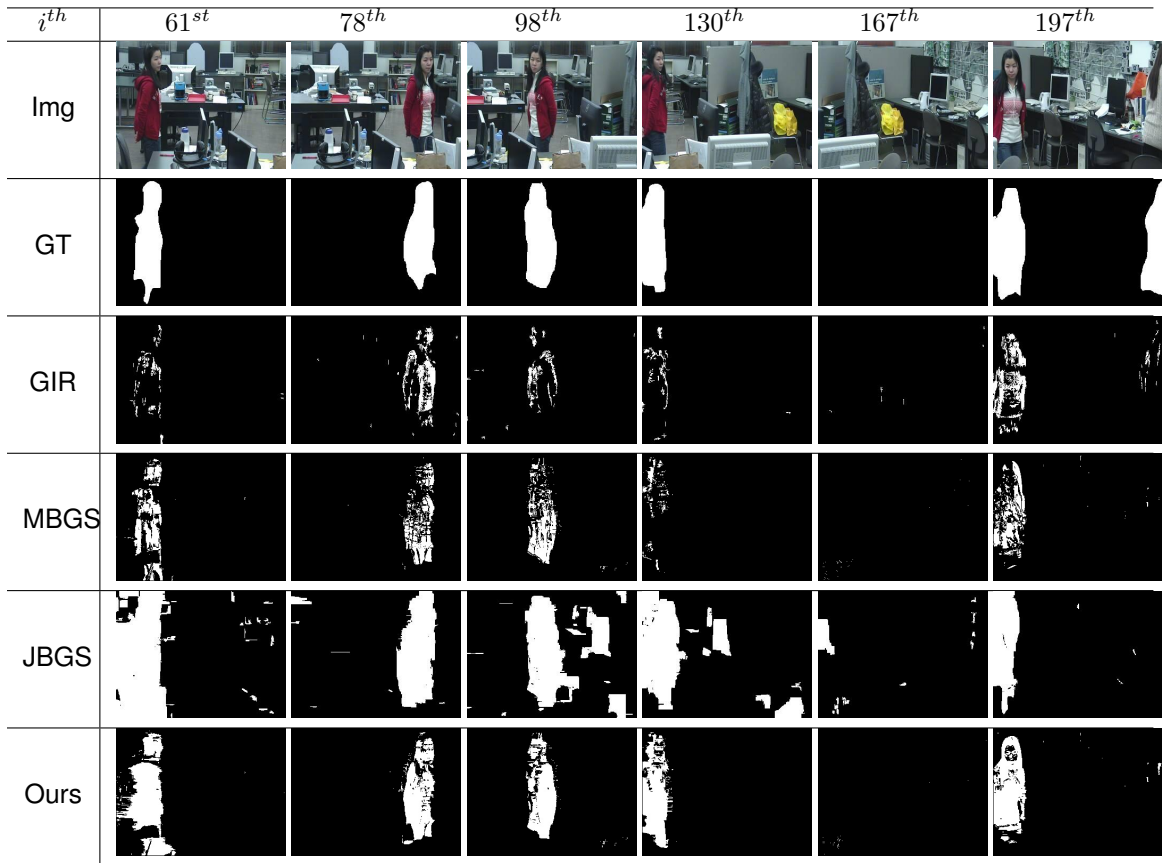


Table 4.7: Qualitative comparison on indoor sequence 3.

Outdoor-1	Precision	Recall	F-measure	CD	FA	DF
GIR [40]	0.46	0.79	0.58	<b>0.86</b>	0.87	<b>0.05</b>
MBGS [56]	<b>0.90</b>	0.42	0.58	0.59	<b>0.02</b>	0.29
JBGS [18]	0.59	<b>0.84</b>	0.69	0.87	0.57	0.08
Ours	0.85	0.74	<b>0.79</b>	0.82	0.15	0.08

Table 4.8: Quantitative evaluation on outdoor sequence 1.

Outdoor-2	Precision	Recall	F-measure	CD	FA	DF
GIR [40]	0.33	<b>0.93</b>	0.49	<b>0.95</b>	0.97	<b>0.03</b>
MBGS [56]	<b>0.91</b>	0.70	0.79	0.39	<b>0.03</b>	0.50
JBGS [18]	0.66	<b>0.93</b>	0.77	0.92	0.64	0.05
Ours	0.87	0.88	<b>0.88</b>	0.88	0.27	0.07

Table 4.9: Quantitative evaluation on outdoor sequence 2.



Outdoor-3	Precision	Recall	F-measure	CD	FA	DF
GIR [40]	0.18	0.78	0.29	0.65	0.90	0.40
MBGS [56]	0.68	0.77	0.72	0.64	0.48	<b>0.07</b>
JBGS [18]	0.29	0.77	0.42	0.60	0.89	0.21
Ours	<b>0.85</b>	<b>0.93</b>	<b>0.89</b>	<b>0.80</b>	<b>0.24</b>	0.20

Table 4.10: Quantitative evaluation on outdoor sequence 3.

Indoor-1	Precision	Recall	F-measure	CD	FA	DF
GIR [40]	0.55	0.72	0.63	0.75	0.86	0.15
MBGS [56]	<b>0.94</b>	0.52	0.67	0.44	<b>0.08</b>	0.45
JBGS [18]	0.33	<b>0.93</b>	0.49	<b>0.98</b>	0.93	<b>0.01</b>
Ours	0.90	0.76	<b>0.83</b>	0.86	0.18	0.07

Table 4.11: Quantitative evaluation on indoor sequence 1.

Indoor-2	Precision	Recall	F-measure	CD	FA	DF
GIR [40]	0.58	0.61	0.60	0.81	0.83	0.12
MBGS [56]	<b>0.91</b>	0.32	0.48	0.44	<b>0.10</b>	0.46
JBGS [18]	0.40	<b>0.89</b>	0.55	<b>0.96</b>	0.90	<b>0.02</b>
Ours	0.83	0.66	<b>0.74</b>	0.78	0.22	0.10

Table 4.12: Quantitative evaluation on indoor sequence 2.

Indoor-3	Precision	Recall	F-measure	CD	FA	DF
GIR [40]	0.50	0.58	0.54	0.42	0.53	0.39
MBGS [56]	<b>0.96</b>	0.59	0.73	0.43	<b>0.06</b>	0.34
JBGS [18]	0.31	<b>0.95</b>	0.47	<b>0.94</b>	0.93	<b>0.03</b>
Ours	0.91	0.75	<b>0.83</b>	0.78	<b>0.22</b>	0.09

Table 4.13: Quantitative evaluation on indoor sequence 3.



## Chapter 5

# Conclusion and Future Work

In this thesis, we have presented a novel background subtraction algorithm incorporating appearance and motion cues for detecting moving objects with a pan-tilt camera. The observation that false positive detections due to pixel misalignment or inaccurate motion estimate tend not to occur simultaneously motivates us to model the appearance and motion separately rather than jointly. Particularly, we evaluate the marginal statistical models of appearance and motion via a two-layer graph - one layer for appearance model and one for motion model where spatial information is considered by imposing label consistency constraints and spatial coherence constraint. Our approach can be applied to detect both large and small moving objects no matter the camera moves fast or slow, as long as the two consecutive frames share sufficient overlap of the field of view, and it can detect the moving objects instantly when the camera moves without a learning process for the new background model. We should also mention that our method is implemented on GPU and we have developed a real-time system to detect objects with a hinged pan-tilt camera.

We have demonstrated the effectiveness and the usefulness of our method through extensive experiments. The comparison with three state-of-the-art background subtraction approaches show that our method reduces most of the false positives produced by pixel misalignment and inaccurate motion estimate at the minimum expense of false negatives. Our approach can be applied to both indoors environment and outdoors environment, or even high-textured environment which still maintains high precision and high recall with few miss detections and false alarms. The GPU implementation allows our approach to process each frame quickly, i.e., a real-time frame rate can be accessed, and thus the pan-tilt camera can be moved smoothly towards the object of interest once it is detected.

Our future work includes the extension of our method to a hand-held camera. Up to now, our method can be applied to any of the two cases: a) the optical centre of a camera does not move, and b) the viewed scene is planar. These two cases enable us to simplify the image registration by approximating it as affine transformation. However, this approximation is not practical on a hand-held camera as the assumptions do not hold. It is important to

relax the restriction and apply the proposed method on a hand-held camera as object detection on such moving platforms is more and more important.

What's more, another concern to improve the proposed approach is to improve its precision and recall of object detection when dealing with zoom-in/zoom-out movement of the camera. Thus, once an object of interest is detected, the camera can zoom in to get a closer look at the tracking object or zoom out if necessary, which will be very helpful in a lot of applications.

# Bibliography

- [1] Nicholas Arcolano and Daniel Rudoy. One-class support vector machines: Methods and applications.
- [2] P. Azzara, L. Di Stefano, and A. Bevilacqua. An effective real-time mosaicing algorithm apt to detect motion through background subtraction using a ptz camera. *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 511 – 516, 2005.
- [3] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, 2011.
- [4] S.A. Berrabah, G. De Cubber, V. Enescu, and H. Sahli. Mrf-based foreground detection in image sequences from a moving camera. *IEEE International Conference on Image Processing*, pages 1125 – 1128, 2006.
- [5] Kiran S. Bhat, Mahesh Sapharishi, and Pradeep Khosla. Motion detection and segmentation using image mosaics. *IEEE International Conference on Multimedia and Expo*, pages 1577–1580, 2000.
- [6] Yuri Boykov and Gareth Funka-lea. Graph cuts and efficient n-d image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.
- [7] D. Butler, S. Sridharan, and V. Bove. Real-time adaptive background segmentation. *International Conference on Acoustics, Speech, and Signal Processing*, 3:341–344, 2003.
- [8] Li Cheng and Minglun Gong. Realtime background subtraction from dynamic scenes. *IEEE International Conference on Computer Vision*, pages 2066–2073, 2009.
- [9] Sen ching S. Cheung and Chandrika Kamath. Robust techniques for background subtraction in urban traffic video. *Visual Communications and Image Processing*, 5308:881 – 892, 2004.
- [10] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov. Bilayer segmentation of live video. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 53 – 60, 2006.
- [11] Marco Cristani, Michela Farenzena, Domenico Bloisi, and Vittorio Murino. Background subtraction for automated multisensor surveillance: A comprehensive review. *European Association for Signal Processing*, pages 1–46, 2010.
- [12] Sascha Cvetkovic, Peter Bakker, and Johan Schirris. Background estimation and adaption model with light-change removal for heawily down-sampled video surveillance signals. *IEEE International Conference on Image Processing*, page 1829=1832, 2006.
- [13] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002.
- [14] Ahmed Elgammal, David Harwood, and Larry Davis. Non-parametric model for background subtraction. *6th European Conference on Computer Vision*, pages 751–767, 2000.

- [15] Shireen Y. Elhabian, Khaled M. El-Sayed, and Sumaya H. Ahmed. Moving object detection in spatial domain using background removal techniques - state-of-art. *Recent Patents on Computer Science*, 1:32–54, 2008.
- [16] Ali Elqursh and Ahmed Elgammal. Online moving camera background subtraction. *Europe Conference on Computer Vision*, 7577:228–241, 2012.
- [17] H. Fradi and J. Dugelay. Robust foreground segmentation using improved gaussian mixture model and optical flow. *IEEE International Conference on Systems*, pages 248–253, 2012.
- [18] Minglun Gong and Li Cheng. Incorporating estimated motion in real-time background subtraction. *IEEE International Conference on Image Processing*, 3265 - 3268, 2011.
- [19] Chun hao Wang and Ling Guan. Graph cut video object segmentation using histogram of oriented gradients. *IEEE International Symposium on Circuits and Systems*, pages 2590–2593, 2008.
- [20] Eric Hayman and Jan-Olof Eklundh. Statistical background subtraction for a mobile observer. *International Conference on Computer Vision*, 1:67–74, 2003.
- [21] Eric Hayman and Jan-Olof Eklundh. Statistical background subtraction for a mobile observer. *IEEE International Conference on Computer Vision*, 1:67–74, 2003.
- [22] Nicholas R. Howe and Alexandra Deschamps. Better foreground segmentation through graph cuts. *Tech Report*, pages 1–8, 2004.
- [23] P. M Jodoin, M. Mignotte, and J. Konrad. Statistical background subtraction using spatial cues. *IEEE Transaction on Circuits and System for Video Technology*, 17(12):1758–1763, 2007.
- [24] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Background modeling and subtraction by codebook construction. *IEEE International Conference on Image Processing*, 5, 2004.
- [25] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russel. Towards robust automatic traffic scene analysis in real-time. *The 33rd IEEE Conference on Decision and Control*, 4:3776–3781, 1994.
- [26] Suha Kwak, Taegy Lim, Woonhyun Nam, Bohyung Han, and Joon Hee Han. Generalized background subtraction based on hybrid inference by belief propagation and bayesian filtering. *IEEE International Conference on Computer Vision*, pages 2174 – 2181, 2011.
- [27] B. Lee and M. Hedley. Background estimation for video surveillance. *Image and Vision Computing New Zealand Conference*, pages 315 – 320, 2002.
- [28] Dar-Shyang Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, 2005.
- [29] Horng-Horng Lin, Tyng-Luh Liu, and Jen-Hui Chuang. A probabilistic svm approach for background scene initialization. *International Conference on Image Processing*, 3:893–896, 2002.
- [30] Horng Horng Lin, Tyng Luh Liu, and Jen Hui Chuang. Learning a scene background model via classification. *IEEE Transactions on Signal Processing*, 57(5):1641–1654, 2009.
- [31] Yang Wang Kia-Fock Loe and Jian-Kang Wu. A dynamic conditional random field model for foreground and shadow segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):279–289, 2006.
- [32] Anurag Mittal and Dan Huttenlocher. Scene modeling for wide area surveillance and image synthesis. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:160–167, 2000.
- [33] Anurag Mittal and Nikos Paragios. Motion-based background subtraction using adaptive kernel density estimation. *Computer Vision and Pattern Recognition*, 2:302–309, 2004.

- [34] Jacinto Nascimento and Jorge Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, pages 761–774, 2006.
- [35] A. Ogihara, H. Matsumoto, and A. Shiozaki. Hand region extraction by background subtraction with renewable background for hand gesture recognition. *International Symposium on Intelligent Signal Processing and Communications*, 2006.
- [36] Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland. A bayesian computer vision system for modeling human interactions. *International Conference on Vision Systems*, 1999.
- [37] Massimo Piccardi. Background subtraction techniques: a review. *IEEE International Conference on Systems, Man and Cybernetics*, 4:3099–3104, 2004.
- [38] Richard J. Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.
- [39] Ying Ren, Chin-Seng Chua, and Yeong-Khing Ho. Motion detection with non-stationary background. *International Conference on Image Analysis and Processing*, pages 78–83, 2001.
- [40] Ying Ren, Chin-Seng Chui, and Yeong-Khing Ho. Statistical background modeling for non-stationary camera. *Pattern Recognition Letters*, 24:183–196, 2003.
- [41] Christof Ridder, Olaf Munkelt, and Harald Kirchner. Adaptive background estimation and foreground detection using kalman-filtering. *Proceedings of the International Conference on recent Advances in Mechatronics*, pages 193–199, 1995.
- [42] Lionel Robinault, Stephane Bres, and Serge Miguet. Real time foreground object detection using ptz camera. *International Conference on Computer Vision, Theory and Applications*, pages 609 – 614, 2009.
- [43] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving camera. *IEEE International Conference on Computer Vision*, pages 1219 – 1225, 2009.
- [44] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.
- [45] Yasuyuki Sugaya and Kenichi Kanatani. Extracting moving objects from a moving camera video sequence. *Proceedings of the 10th Symposium on Sensing via Image Information*, pages 279–284, 2004.
- [46] Zhen Tang and Zhenjiang Miao. Fast background subtraction using improved gmm and graph cut. *2008 Congress on Image and Signal Processing*, pages 181–185, 2008.
- [47] A. Tavakkoli. Automatic video object plane extraction using non-parametric kernel density estimation. *Mathematical Methods in Computer Vision*, 2005.
- [48] Alireza Tavakkoli, Micrea Nicolescu, and George Bebis. A novelty detection approach for foreground region detection in videos with quasi-stationary backgrounds. *Proceedings of the Second International Conference on Advances in Visual Computing*, 1:40–49, 2006.
- [49] Kentaro Toyama, John Krumm, Barry Brumitt, and Brian Meyers. wallflower: Principles and practice of background maintenance. *The Seventh IEEE International Conference on Computer Vision*, 1:255–261, 1999.
- [50] Du-Ming Tsai and Shia-Chih Lai. Independent component analysis-based background subtraction for indoor surveillance. *IEEE Transactions on Image Processing*, 18(1):158–167, 2009.
- [51] Berwin A. Turlach. Bandwidth selection in kernel density estimation: A review. *CORE and Institut de Statistique*, pages 1–33, 1993.

- [52] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Real-time tracking of the human body. *IEEE Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.
- [53] M. Xiao, C. Han, and X. Kang. A background reconstruction for dynamic scenes. *IEEE International Conference on Image Processing*, pages 1–7, 2006.
- [54] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv- $l^1$  optical flow. *Proceedings of the 29th DAGM conference on Pattern Recognition*, pages 214–223, 2007.
- [55] J. Zheng, Y. Wang, N. Nihan, and E. Hallenbeck. Extracting roadway background image: A mode based approach. *Transportation Research Board of the National Academies*, 2006.
- [56] DongXiang Zhou and Hong Zhang. Modified gmm background modelling and optical flow for detection of moving objects. *IEEE International Conference on Systems*, 3:2224–2229, 2005.
- [57] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. *Proceedings of the International Conference on Pattern Recognition*, 2:28–31, 2004.