SSI-DT79125042

Doctoral Thesis

Dynamic Scenes and Appearance Modelling for Robust Object Detection and Matching Based on Co-occurrence Probability

Dong LIANG

February, 2015

Division of Systems Science and Informatics
Graduate School of Information Science and Technology
Hokkaido University

Doctoral Thesis
submitted to Graduate School of Information Science and Technology,
Hokkaido University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Dong LIANG

Thesis Committee: Shun'ichi KANEKO Professor
Masahiko ONOSATO Professor
Takayuki TANAKA Associate Professor

Dynamic Scenes and Appearance Modelling for Robust Object Detection and Matching Based on Co-occurrence Probability*

Dong LIANG

Abstract

Detecting moving objects plays a crucial role in an intelligent surveillance system. Object detection is often integrated with various tasks, such as tracking objects, recognizing their behaviours and alerting when abnormal events occur. However, it suffers from non-stationary background in surveillance scenes, especially in two potentially dynamic cases: (1) sudden illumination variation, such as outdoor sunlight changes and indoor lights turning on/off; (2) burst physical motion, such as the motion of indoor artificial background, which include fans, escalators and auto-doors, and the motion of natural background, which include fountain, ripple on water surface and swaying tree. If the actual background includes a combination of any of these factors, it becomes much more difficult to detect objects. Traditional algorithms, i.e. Gaussian Mixture Model (GMM) and Kernel Density Estimation (KDE) handle gradual illumination changes by building the statistical background models progressively using long-term leaning frames. In practice, however, this kind of independent pixel-wise model often fail to avoid mistakenly integrating foreground elements into the background, and it is difficult to adapt to sudden illumination change and burst motion. On the other hand, spatial-dependence model, i.e. Grayscale Arranging Pairs (GAP) and Statistical Reach Feature (SRF), shows promising performance under illumination change and other dynamic background.

This study proposes a novel framework to build a background model for object detection, which is evolved from GAP method and SRF method. It is brightness-invariant and able to tolerate burst motion. We name it Co-occurrence Probability based Pixel Pairs (CP3). In order to model the dynamic background, spatial pixel pairs with high temporal co-occurrence probability are employed to represent each other by using the stable intensity differential increment between a pixel pair which is much more reliable than the intensity of a single pixel, especially when the intensity of a single pixel changes dramatically over time. The model performs robust detection under outdoor and indoor extreme environments. Compared with the independent pixel-wise background modelling methods, CP3 determines stable co-occurrence pixel pairs, instead of building the parameterized/non-parametrized model for a single pixel. These pixel pairs maintain a reliable background model, which can be used to capture structural background motion and cope with local and global illumination changes. As a spatial-dependence method, CP3 does not predefine/assume any local operator, subspace or block for an observed pixel, but

*Doctoral Thesis, Division of Systems Science and Informatics, Graduate School of Information Science and Technology, Hokkaido University, SSI-DT79125042, February 14, 2015.

it does its best effort to select those qualified supporting pixels which could maintain reliable linear relationship with the target pixels. Moreover, based on the single Gaussian model of the differential value of the pixel pair, it provides an accurate detection criterion even the gray-scale dynamic range is compressed under weak illumination.

The proposed method can be used for modelling the appearance of an image to realize image matching. Theoretically speaking, both the object detection and image matching can be seen as a model matching problem. The differences between the two tasks is that the object detection is to seek the regions of interest (ROI) which violate/mismatch the background model, while the image matching is to seek the ROI which can match the image model optimally. Therefore, in this study, we further extend the use of CP3 to the robust image matching task.

This thesis is organized into the following chapters:

Chapter 1 introduces related works in object detection and image matching. Some general problems are involved and discussed. Furthermore, the motivations and contributions of this research are described.

Chapter 2 presents the details of CP3 background model based on co-occurrence pixel pairs for object detection. We test it on several surveillance video datasets for both qualitative and quantitative analyses. Experiments using several challenging datasets (Heavy fog, PETS-2001, AIST-INDOOR, Wallflower and a supermarket surveillance application) prove the robust and competitive performance of object detection in various indoor and outdoor environments. For quantitative analysis, Precision (also known as positive predictive value), Recall (also known as sensitivity) and F-measure (a weighted harmonic mean of the Precision and Recall) are utilized. The three evaluation metrics measure the exactness, fidelity and the completeness of foreground. We compare our algorithm with three methods: (1) GMM method, which is a standardized method among independent pixel-wise models; (2) Sheikh's KDE method as a representative method among spatially dependent models; (3) our previous method GAP. In addition, we also propose an accelerated version of CP3, which can effectively reduce the time cost of the background modelling stage.

Chapter 3 proposes the framework of CP3 for modelling the appearance of an image to realize image matching. We detail the learning phase and present the similarity measure procedure and present the experimental results. Although an additional learning stage is necessary, the experiment results show that the proposed method is robust under several imaging cases and it also outperforms SRF.

Chapter 4 presents the discussions of the proposed methods, concludes the main contributions of our study, and shows the future works of this study.

Keywords: intelligent surveillance, dynamic scenes, robustness, object detection, image matching

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Research background

In the field of computer vision, high-dimensional data from the real world is acquired, processed, analysed, in order to produce decisions in the form of numerical or symbolic information. A theme in the development of this field has been to duplicate the abilities of human vision by electronically perceiving and understanding a static image or a vivid video.

A stream of research within computer vision which has gained a lot of importance in the past few years is the understanding of scene. Scene understanding has been applied to various fields, the most popular application of which is an intelligent surveillance system. Before the complexity of a surveillance scene can be understood, we first need automatic methods for detecting objects from it. Once the object is detected, depending on the application, the system can do further processing to go into the details of understanding the activity. In an intelligent surveillance system, detecting objects often integrates with various tasks, such as tracking objects [1, 2], recognizing their behaviours [3, 4] and alerting when abnormal events occur [5]. Other applications of object detection include character animation for games and movies, avatars for teleconferencing, advanced intelligent user interfaces, biomechanical analysis of actions for sports and medicine, etc.

It is understandable that, object detection can be seen as a pixel level image processing problem which closes to the imaging sensing device. The problem we hold is that, is the pixel belong to the object or belong to the background? Consequently in the field of computer vision and statistical pattern recognition, it can be seen as a typical binary classification problem.

On the basis of visible spectrum imaging technology, surveillance video camera could be the most widely-used imaging sensing device for an intelligent system [6, 7, 8]. Compared with more sophisticated thermal camera and stereo imaging device [9], its flexibility and low-cost will guarantee a long-term and extensive utilization. Nevertheless, such video camera itself has several weaknesses which influence its imaging quality: first, since the frame rate is fixed (typically 30 fps for a NTSC system), the shutter speed cannot delay longer than the frame interval. Thus, under a low-light condition, the sensor would not keep sufficient exposure, so that the intensity of signal easily submerge into background noise. Second, surveillance task requires large depth of field (i.e. capture a globally sharp scene), the light-gathering aperture is relatively small (i.e. with a large Focal ration of the aperture), which also leads to underexposure. All these problems are affecting the result of object detection.

On the other hand, object detection suffers from non-stationary scenes in surveillance videos, especially in two potentially serious cases: (1) sudden illumination variation, such as outdoor sunlight changes and indoor lights turning on/off; (2) burst physical motion, such as the motion

(a) Outdoor sun light change

(b) Indoor light turing on and off

(c) Ripple on water suface

(d) Escalator in operation

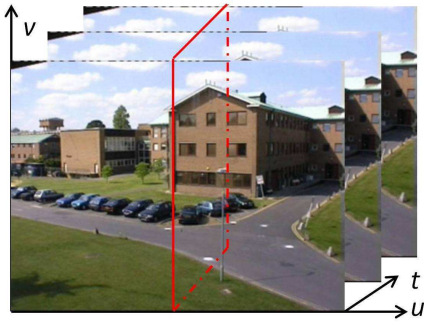
Figure 1.1: Four typical examples of dynamic background.

of indoor artificial objects, which include escalators and auto-doors, and outdoor natural dynamics, including swaying tree and water ripple. If the actual background includes a combination of any of these factors, it becomes even more difficult to perform detection. Figure 1.1 shows four typical examples of dynamic background. State-of-the-art algorithms [10, 11, 12, 13, 14] can handle gradual illumination changes by updating the statistical background models progressively as time goes by. In practice, however, this kind of model update is usually relatively

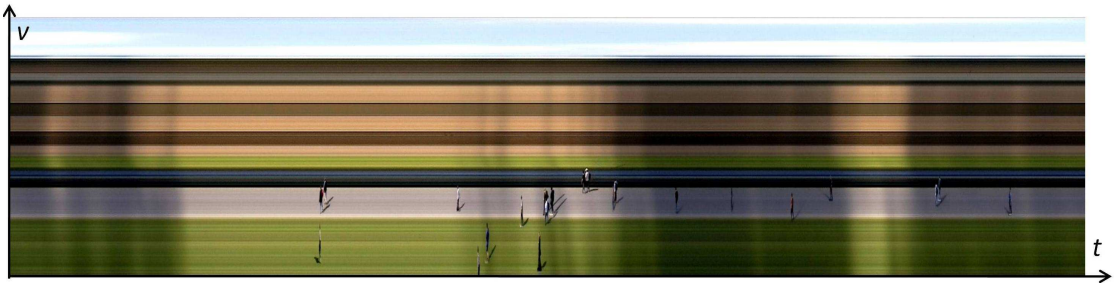
slow to avoid mistakenly integrating foreground elements into the background model, making it difficult to adapt to sudden illumination changes and burst motion.

1.2 Background modelling methods

The rationale in the approach of detecting the moving objects is that finding the difference between the current frame and a reference background scene, often called background image (a specific frame in simple background subtraction method), or background model (a description of the its abstract in a more sophisticated statistical method). A robust background model should be able to handle lighting changes, repetitive motions from clutter and long-term scene changes. Figure 1.2 shows a typical example under sunlight change.



(a) PET2001-d3-c1 Dataset



(b) 2D spatial-temporal image using pixels along a vertical line to create visible spatial-temporal slice.

Figure 1.2: 2D spatial-temporal image to observe sudden illumination variation.

1.2.1 Simple background subtraction

Simple background subtraction [15] of images acquired at different instants in time makes object detection possible, assuming a stationary camera position and constant illumination. A difference image $d(u, v)$, where (u, v) is the location of a pixel, is a binary image where non-

zero value represent image areas with object, that is, areas there was a substantial difference in intensity between a background images f_B and current one f_C.

$$d(u,v) = \begin{cases} 0 & \text{if } |f_C - f_B)| < th \\ 1 & \text{otherwise} \end{cases} \tag{1.1}$$

where th is a small positive number.

In real-world scene, due to outdoor and indoor illumination changes, the simple background subtraction cannot detect object effectively. When utilizing the simple background subtraction under sunlight change in Figure 1.2, the detection results could be like Figure 1.3. The results appear a large area of the background noise, resulting in some targets submerged in the background. Therefore, the background subtraction face the issue that how to fully exploit the temporal information, to build a more accurate background model.

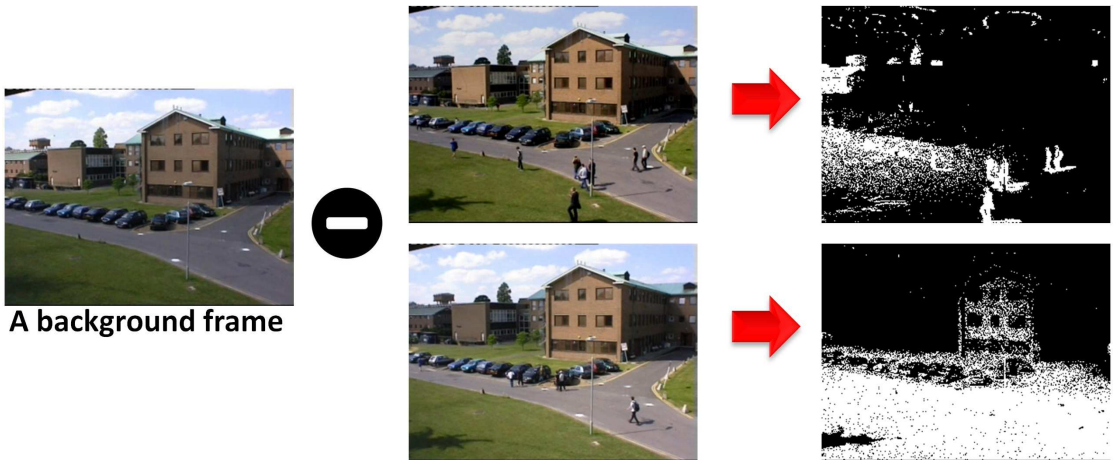


Figure 1.3: Object detection based on simple background subtraction

Since observations of the background in image sequences can be considered stochastic events, many statistical approaches have been employed to model effective backgrounds. The former background modelling approaches can be classified into two categories: (1) Independent pixel-wise modelling, which employs the statistical processing of time-domain observations to each pixel. (2) Spatial-dependence modelling, which employs principles to exploit spatial-dependence among pixels to build a local or global model.

1.2.2 Independent pixel-wise modelling

Independent pixel-wise modelling employs the statistical processing of time-domain observations to each pixel. Most of the earlier background modelling approaches tend to fall into this category. The representative algorithms in this approach are the Single Gaussian background model (SG), Gaussian mixture model (GMM), Kernel density estimation (KDE), Hidden Markov models (HMMs), Codebooks (CBs) [10, 11, 16, 12, 17, 18, 13]. The following is the key contribution of these algorithms and some discussions.

Single Gaussian background model (SG)

Wren [10] modelled the observations (YUV) of each pixel as a single Gaussian probability density function. The proposed algorithm models every background pixel with μ_0 to be the mean (on YUV colour space) of a point on the texture surface, and σ_0 to be the covariance of that pixel distribution. Obviously, the SG model can tolerate noise well, but it is not a fine model for outdoor scenes because it should be capable of dealing with movement through cluttered areas in the visual field, lighting changes, effects of moving elements of the scene (e.g., swaying trees).

Gaussian mixture model (GMM)

A typical example is that of an outdoor scene with trees partially covering a building: a same (u, v) pixel location will show values from tree leaves, tree branches, and the building itself. Other examples can be easily drawn from snowing, raining, or watching sea waves from a beach. In these cases, a single valued background is not an adequate model. To cope with multiple background patterns, the Gaussian mixture model (GMM) [11, 16, 19] was proposed, which use multiple Gaussian distributions deemed to describe only one of the observable background or foreground objects. In practical cases, setting the proper number of Gaussian distributions is a difficult task. modelling the background variations with a small number of Gaussian distribution will not be accurate. Furthermore, the very wide background distribution will result in poor detection because a large range of the gray level spectrum would be covered by the background model.

Kernel density estimation (KDE)

Elgammal [12] employed kernel density estimation (KDE) as a data-driven modelling method. Even with a small number of samples, kernel density estimation leads to a smooth, continuous, and differentiable density estimate. Since kernel density estimation is a non-parametric model, it does not assume any specific underlying distribution and the estimate can converge to any density shape with enough samples, this approach is suitable to model the colour distribution of regions with patterns and mixture of colours. If the underlying distribution is a mixture of Gaussians, kernel density estimation converges to the right density with a small number of samples. Unlike parametric fitting of a GMM, kernel density estimation is a more general approach that does not require the selection of the number of Gaussians to be fitted. Therefore, KDE is closer to the real probability distribution than GMM.

Hidden Markov models (HMMs)

Hidden Markov models (HMMs) [17, 18] have also been applied to model the background; topology free HMMs were described and several state splitting criteria were compared in the context of background modelling in [17], and a non-adaptive three-state HMM was used to

model the background in [18].

Codebooks (CBs)

The recent notable pixel-wise method by Kim [13] presented a real-time algorithm, which sampled background pixel values and quantized them into compressed codebooks (CBs). To improve the processing efficiency of the codebooks, Guo [20] presented a hierarchical scheme.

All the approaches model independently at a single pixel's location. When detecting objects under a dynamic background, both the object occupation and the illumination change (or other forms of dynamics) can affect on the current intensity of a pixel. Therefore, a background model only modelling the independent intensity change cannot distinguish the object from the dynamic background well. On the other hand, it is intuitive that spatial neighbouring locations will exhibit statistical correlation in the background model, which could be an evidence for the classification of pixel's values. To exploit this property, various methods have been used for detecting the foreground pixels.

1.2.3 Spatial-dependence modelling

The second category uses spatial information to exploit the spatial dependencies of pixels in the background. The representative algorithms in this approach are Co-occurrence block, Wallflower, Eigenspace decomposition, Sheikh's Spatial KDE, Auto-regressive moving average (ARMA), LBP feature, Tensor subspace learning [21, 22, 23, 14, 24, 25, 26]. The following is the key contribution of these algorithms and some discussions.

Co-occurrence block

Matsuyama [21] proposed a regional block matching method against varying illumination, and Seki [27] proposed a co-occurrence-based block correlation method. Their main statement is that neighbouring blocks of pixels belonging to the background should experience similar variations over time. Although this assumption proves true for blocks belonging to a same background object (such as an area with tree leaves), it will evidently not hold for blocks at the border of distinct background objects, which means the above two methods can only yield coarse region-level detection.

Wallflower

Toyama et al. [22] proposed a three layers (pixel, region, and frame level) algorithm framework in which Weiner filters were employed. It used region and frame-level information to verify the pixel-wise background model.

Sheikh's Spatial KDE

Sheikh [14] used the joint representation of image pixels in a local spatial distribution (proximal pixels) and colour information to build both background and foreground KDE models competitively in a decision framework. In the background modelling stage, this method converts each pixel's intensity and location into a feature vector, and then models a joint probability distribution of all the pixels on a feature space. The method is based on the assumption that the pixels' correlation degree depends on the spatial distance between them, (i.e. each observed pixel has higher correlation with its neighbouring pixels), but ignores the localized discrimination between pixels. Such a mechanism can deal well with global illumination changes, but is not robust to local illumination change.

Eigenspace decomposition

Oliver [23] employed eigenspace decomposition in which the background was modelled by the eigenvectors corresponding to the largest eigenvalues. It applied to the whole image instead of blocks. Such an extended spatial domain can extensively explore spatial correlation and avoid the tiling effect of block partitioning.

Auto-regressive moving average (ARMA)

Monnet [24] and Zhong [28] built an auto-regressive moving average (ARMA) model in dynamic scenes, which is used to incrementally learn (using Principal Component Analysis) and then predict motion patterns in the scene.

LBP feature

Heikkilä and Pietikäinen [25] used a local binary pattern (LBP) to subtract the background and detect moving objects in real time. This method models each pixel as a group of adaptive LBP histograms that were calculated over a predefined circular region around the pixel.

Tensor subspace learning

A recent spatial-dependence approach [26] utilized a tensor subspace learning algorithm to represent spatial correlations between pixel values, and modelled appearance changes by incrementally learning a tensor subspace representation by adaptively updating the sample mean and an eigenbasis for each unfolding matrix of the tensor.

All the approaches exploit the spatial dependencies of pixels in the background to build the background model according to neighbouring pixels, defined blocks, fixed directions, local operator, or subspace. Essentially, these methods are based on the assumption that the pixels'

correlation degree depends on the spatial distance between them that each observed pixel has higher correlation with its neighbouring pixels, but ignores the potential localized discrimination between pixels, and also ignores comprehensive spatial information out of the defined range. Therefore, we have a question that, can we have the freedom to break such kind of constraint when we are designing a spatial-dependence background model?

1.2.4 Grayscale arranging pairs

Based on the above consideration, in our previous research, the Grayscale Arranging Pairs (GAP) [29, 30] was first proposed for background modelling, and then it was extended for object tracking [31], face recognition [32], and industrial defect detection [33]. In the training stage of GAP, unlike many traditional spatial-dependence background modelling methods which consider the predefined fixed spatial structure, it does this by considering the relationship between this pixel and several statistically chosen reference pixels, without any constraint of specific spacial location of them. For a target pixel P, suppose we have N positive reference pixels which statistically have higher intensity than P, and N negative reference pixels which statistically have lower intensity than P (refer to [29] for the details of statistically choosing the reference pixels). In a new frame, P is classified as the background when its intensity is normal: lower than those of the positive reference pixels and higher than those of the negative reference pixels; contrarily, P is classified as the background when its intensity is abnormal: higher than those of the positive reference pixels or lower than those of the negative reference pixels. In details, two probabilities that concern whether P belongs to the background are calculated as follows: positive probability $\xi^+ = n^+/N$ (The number n^+/N denotes that the intensity of P is higher than that of n^+ out of N positive reference pixels) and negative probability $\xi^- = n^-/N$ (The number n^-/N denotes that the intensity of P is lower than that of n^- out of N negative reference pixels). These two probabilities compose the GAP feature, and $0 \leq \xi^{\pm} \leq 1$. Although this method is quite effective to deal with unstable background, it still have some open problems: as a threshold-based method in both training and detection stage, the detection results is sensitive to the change of threshold. This is a motivation we would like to improve it.

1.3 Image matching methods

Image matching constructs an important foundation for robot vision systems in a variety of fields such as navigation, object locating and recognition, and computer-aided diagnosis and treatment [34, 35]. It searches an image in various scenes based on a rational quantitative similarity measure and locates its position at an optimum value. Many methods have been proposed [36, 37, 38, 39, 40, 41] to achieve such an objective. In a traditional industrial machine vision system, the robustness of the similarity measure can be guaranteed by simplifying the external environment, for instance, maintaining steady and bright illumination. However, most of the real-world applications suffer from uncontrollable disturbances, e.g. illumination variation, low illumination, sensor noise, occlusion and highlight, which often lead to a high false matching

rate. Generally speaking, approaches for image matching can be roughly classified into two major groups: feature-based approaches and intensity-based approaches.

1.3.1 Feature-based approaches

The feature-based approaches [42, 36, 37, 43] aim to firstly find out salient features, such as sharp edges, structured lines, or salient points. Because the selected features are generally quite sparse, these methods are computationally efficient. Nevertheless, detecting effective features generally implements pipeline framework, using preprocessing steps that need some heuristic scheme for the selection of threshold values. Such pipeline methods can result in a fragile architecture which may suffer from a domino effect as an error can propagate to the subsequent processing stages, especially under various ill-conditions. Therefore, these methods highly depend on the quality of specific images, once involving image degradation, such as low illumination, high noise, saturation and blurry, they hardly seek robust features for matching.

1.3.2 Intensity-based approaches

Relatively more straightforward, the intensity-based approaches [38, 39, 40, 41] are more general and flexible compared with the former. Normalized Cross Correlation (NCC) and its improvement [38, 40] are well-known methods and have been very widely employed because they can cope with noisy images and uniform illumination changes. However, as shown in Figure 3.4 to Figure 3.12 in the experiment section, when a scene involves low light, partial occlusion of objects or saturation/highlight, this sum of multiplication correlation technique sometimes fails to capture true positions. SSD or sum of absolute difference [39] evaluates a sum of squared or absolute differences in the brightness defined between two distinct images. Although it is an efficient algorithm, it is very feeble for changes in brightness caused by illumination fluctuation, occlusion and saturation. ISC [41, 34] can cope with occlusion, shadow, saturation, or highlight. This method is based on a similarity measure evaluated by aggregating increment sign codes at each pixel, which represents Boolean signs between any two adjacent pixels. It is proved to be robust because it does not directly utilize the brightness of pixels, but their incremental tendency, and it is efficient even in comparison with NCC. Due to the information reduction through neglecting brightness values, nevertheless, there have been some problems such as the size of a template cannot be reduced and it cannot discriminate slight changes in the brightness of images. The robust similarity measure using Rank Statistics [44] is based on an alternative approach for brightness variation. This type of similarity measure is based on the fact that the brightness itself changes in the case of a rather uniform variation in illumination, but their ranks or orders can remain unchanged. However, it is largely influenced by significant change of local perturbation of ranks caused by occlusion, shading and saturation.

1.3.3 Statistical reach feature

The Statistical Reach Feature (SRF) [45] builds a local texture model for each target pixel to be brightness-invariant for background subtraction, it is also proper for image matching task [46]. Different from the above methods, SRF method, uses sign coding of the intensity relationship between pixel pairs selected from an image. In detail, the intensity $I(P)$ of a pixel P in the reference image is converted to a code string of the intensity differences from adjacent pixels Q with sufficiently large absolute value of the intensity difference $|I(P) - I(Q)| \geq T_P$, which is used as a local feature at a pixel P, where T_P is a threshold to tolerate image noise. The binary value of $srf(P \rightarrow Q, T_P)$ for an ordered pair (P, Q) consisting of two pixels in reference image I is defined as

$$srf(P \rightarrow Q, T_P) = \begin{cases} 1 & I(P) - I(Q) \geq T_P \\ 0 & I(P) - I(Q) \leq -T_P \\ \varnothing & \text{otherwise} \end{cases} \tag{1.2}$$

SRF introduces a mechanism using noise-resistant signs as observations, which provides robustness to various disturbances. In addition, multiple pixels Q are employed, and the resulting code string is used to calculate the local similarity of the pixel P. Then the similarity between the reference image and the input image is defined as the normalized sum of the local similarities [46].

1.3.4 Some other approaches

Besides these algorithms, some mathematical manipulations were intended for geometric invariance, for instance, rotation invariance [47, 48], scale invariance [49, 36], affine invariance [50] and elastic image registration [51, 52]. In this study, since we focus on the robustness of similarity measure, we will not detailed discuss them which can be readily integrated as a middle-ware for specific applications.

1.4 Introduction for the study

1.4.1 Motivation

For background modelling

In our previous research, we proposed a background model called GAP [29, 30]. GAP employed an alignment of supporting pixels for the target pixel which held a stable intensity subtraction in training frames without any restriction of locations. The intensity subtraction of the pixel pairs allowed the background model to tolerate noise and be illumination-invariant. However, this fixed intensity subtraction influenced the sensitivity of the background model, especially when the dynamic range was compressed due to low illumination; it was also not an optimal way to search for supporting pixels by using a fixed intensity subtraction in that most co-occurrence pixels were not considered. In addition, the GAP method mainly focused

on illumination-invariance, so that the dynamic background caused by burst motion was not discussed sufficiently.

In this study, we propose a novel framework to build a background model for object detection, which is brightness-invariant and able to tolerate burst motion. We name it Co-occurrence Probability based Pixel Pairs (CP3) [53, 54]. The proposed method addresses these open problems. Compared with GAP, the proposed method employs a co-occurrence histogram to describe the relationship of a pixel pair, which is free from any intensity differences, and calculates normalized correlation coefficients for measuring the degree of co-occurrence which can deal with a dynamic background. It also introduces a spatial clustering operation to select optimal supporting pixels and then provides a more accurate parameterized detection criterion instead of a fixed double-sided threshold.

For image matching

An ideal similarity measure for image matching should be discriminative, producing a conspicuous correlation peak as well as suppressing false local maxima. However, the image matching tasks in practice often involve complex situations, such as blurry scene and illumination fluctuation, causing the similarity measure is not discriminative enough. In SRF method, the sign coding is produced by a global threshold T_P, which would not be accurate enough to describe the intensity relationship of a specific pixel pair. In detail, T_P is a trade-off parameter, small T_P cannot maintain a reliable relationship between a pixel pair because the image noise always exists. On the other hand, large T_P would cause a large number of ineffective pixel pair, i.e. the sign code $srf(P \rightarrow Q; T_P) = \varnothing$. This will make SRF lose some details of the image, leading to an insensitive matching. This problem cannot be ignored when the image is a degenerated by defocusing or low illumination. Actually, the intensity difference of a pixel pair can be changed along with the illumination change, e.g. the intensity difference will be smaller under a low illumination case than it under a normal illumination, and vice versa.

In this study, we further extend the use of CP3 to the robust image matching task. In detail, the proposed method is used for modelling the appearance of an image to realize image matching. Similar to the SRF method, the proposed method based on CP3 also relies on the intensity relationship between pixel pairs. Different from SRF method, CP3 builds the intensity difference as a single Gaussian model for each specific pixel pair via a spatial-temporal learning stage, rather than a global threshold T_P in SRF, which is a more accurate description of a pixel pair. The proposed image matching framework also removes the definition of $\varnothing$ in SRF, allowing the pixel pair with similar intensity also building their Gaussian model, which maintain a constant number of supporting pixels Q for each target pixel P.

1.4.2 Contributions

In this study, we propose a novel background model for object detection based on co-occurrence pixel pairs. The model performs robust detection under outdoor and indoor ex-

treme environments. CP3 determines stable co-occurrence pixel pairs, instead of building the parameterized/non-parametrized model for a single pixel. These pixel pairs maintain a reliable background model, which can be used to capture structural background motion and cope with local and global illumination changes. As a spatial-dependence method, CP3 does not predefine any local operator, subspace or block, and it provides an accurate detection criterion even though the gray-scale dynamic range is compressed under weak illumination. Compared with GAP, the proposed method employs a co-occurrence histogram to describe the relationship of a pixel pair, which is free from any intensity differences, and calculates normalized correlation coefficients for measuring the degree of co-occurrence which can deal with a dynamic background. It also introduces a spatial clustering operation to select optimal supporting pixels and then provides a more accurate parameterized detection criterion instead of a fixed double-sided threshold.

In this study, an image matching method based on CP3, which has the nature of robustness for severe imaging conditions, has been also proposed. By learning a set of reference images, the spatial and statistical models effectively improved the quality of similarity measure, and the pixel pair structure effectively reduces the disturbances of the illumination fluctuation. CP3 builds the intensity difference as a single Gaussian model is a more accurate description of a pixel pair than a global threshold in SRF.

1.4.3 Overview

Chapter 1 introduces related works in object detection. Some general problems are discussed and considered. Furthermore, the motivations and contributions of this research are described.

Chapter 2 presents the details of CP3 based on co-occurrence pixel pairs. We tested it on video datasets for both qualitative and quantitative analyses. Experiments using several challenging datasets (PETS-2001, AIST-INDOOR, Wallflower and a surveillance application) prove the robust and competitive performance of object detection in various indoor and outdoor environments. For quantitative analysis, Precision (positive predictive value), Recall (sensitivity) and F-measure (a weighted harmonic mean of the Precision and Recall.) are utilized to measure the exactness, fidelity and the completeness of foreground. We compares our algorithm with: (1) GMM method (2) Sheikh's KDE; (3) our previous method GAP. In addition, the accelerated version of CP3 can effectively reduce the time cost of the background modelling stage.

Chapter 3 proposes the framework of CP3 for modelling the appearance of an image to realize template matching. We detail the learning phase and present the similarity measure procedure and present the experimental results. Although an additional learning stage is necessary, the experiment results show that the proposed method is robust under several imaging cases and it also outperforms other two well-known robust similarity measure Normalized Cross Correlation (NCC) and Statistical Reach Feature (SRF).

Chapter 4 presents the discussions of the proposed methods, concludes the main contributions of our work, and shows the future works of this study.

Chapter 2. CP3 for object detection

2.1 Overview

In this chapter, we propose a novel framework to build a background model for object detection, which is brightness-invariant and able to tolerate burst motion. We name it Co-occurrence Probability based Pixel Pairs (CP3) [53, 54]. It is inspired by the previous work in [55, 56]. In the work of R. M. Haralick et al. [55], gray-level co-occurrence matrices (GLCM) were employed to measure the spatial co-occurrence of pixels to produce an image texture feature (Haralick feature). In the work of Hashimoto and Saito [56], pixels with low spatial co-occurrence probability and with high temporal co-occurrence probability were preferentially extracted as spatially distinctive and temporally stable features to reduce computational complexity for template matching.

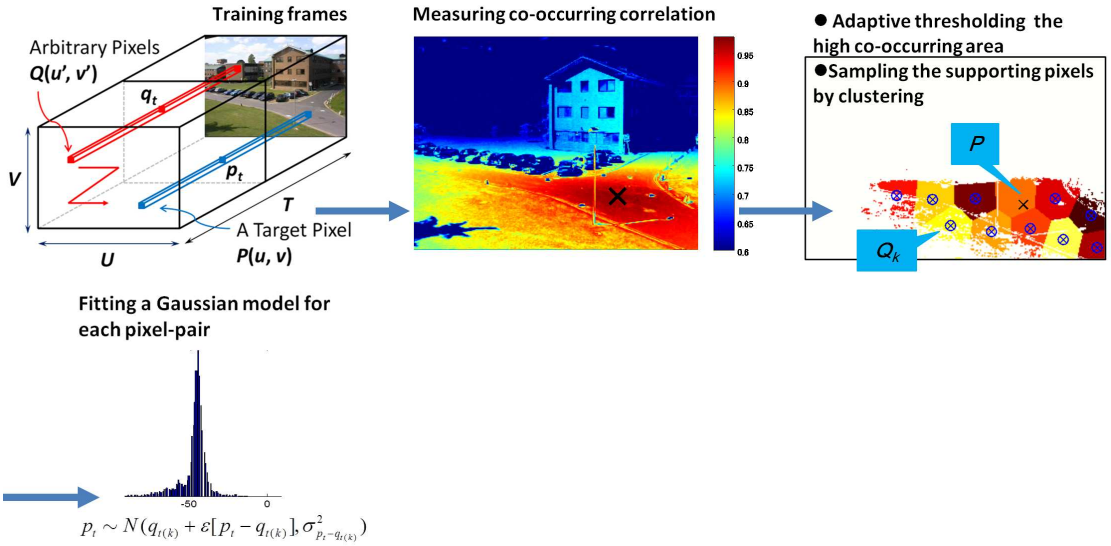


Figure 2.1: Schematic diagram of CP3 background model.

In this study, in order to model the dynamic background, spatial pixel pairs with high temporal co-occurrence probability are employed to represent each other by using the stable intensity differential increment between a pixel pair which is much more reliable than the intensity of a single pixel, especially when the intensity of a single pixel changes dramatically over time. Figure 2.1 shows the schematic diagram of CP3 background model. CP3 method first conducts an off-line learning stage to build the parametrized background model, and then utilizing the background model for object detection.

A pixel pair consists of each pixel itself (called *target pixel* hereafter) and a selected pixel (called *supporting pixel* hereafter). As a pixel-wise background model, the target pixel P refers to all pixels in a scenario. The supporting pixels are neither arbitrary pixels in the scene, nor predefined fixed local structures around each target pixel; instead, the supporting pixels are selected based on their statistical stability with the target pixels.

The remainder of this chapter is organized as follows. Section 2.2 details the background model. Section 2.3 presents the object detection procedure. Section 2.4 presents the experimental results, and Section 2.5 summarizes this chapter.

2.2 Background modelling

The algorithm is described for gray-scale imagery; however, it can also be used for colour or multi-modality imagery with minor modification.

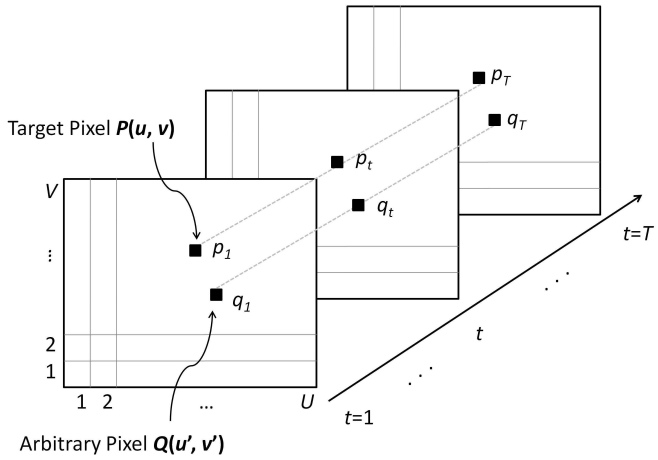


Figure 2.2: Fundamental definitions of the image data. Target pixel P and an arbitrary pixel Q with their intensity sequence over time p_t and q_t.

Figure 2.2 shows the fundamental definitions of the image data. Suppose we are given a training image sequence $B = \{I_1, I_2, ..., I_T\}$ with a total of T images, and each image has $M = U \times V$ pixel positions. In the three-dimensional space $\Gamma = \{(u, v, t)|1 \leq u \leq U, 1 \leq v \leq V, 1 \leq t \leq T\}$, we have $U \times V \times T$ intensity values within a gray-scale level range $[0, L-1]$.

In the following, the intensities over time at each pixel position are regarded as samples from a stochastic process. We define P as a *target pixel* at location (u, v). The location of P varies to cover all pixels of a frame, and its intensity sequence over time is denoted as $\{p_t(u, v)\}_{t=1,2,...,T}$. In the same way, we define $Q(u', v')$ as an *arbitrary pixel* with intensity sequence $\{q_t(u', v')\}_{t=1,2,...,T}$ at location (u', v'). For simplicity, we have omitted most of the (u, v) and (u', v') in the following discussion.

2.2.1 Bivariate statistical property of a pixel pair

To analyze the bivariate statistical property of a pixel pair, the joint histogram of intensity for a pixel pair is defined.[1] The i,jth bin of the joint histogram for an arbitrary pixel pair (P, Q) in T training images can be expressed as

$$h_{PQ}(i,j) = \sum_{t=1}^{T} \delta(p_t, q_t, i, j), \tag{2.1}$$

where $\delta(p_t, q_t, i, j)$ represents the two-dimensional discrete Kronecker delta function:

$$\delta(p_t, q_t, i, j) = \begin{cases} 1 & if\ (p_t = i) \cap (q_t = j) \\ 0 & otherwise \end{cases}.$$

The bins $h_{PQ}(i,j)$ corresponding to $i,j \in [0, L-1]$ represent the co-occurrence probability of $p_t = i$ and $q_t = j$. The joint histogram $\boldsymbol{h}_{PQ}$ can be written compactly as an ordered array,

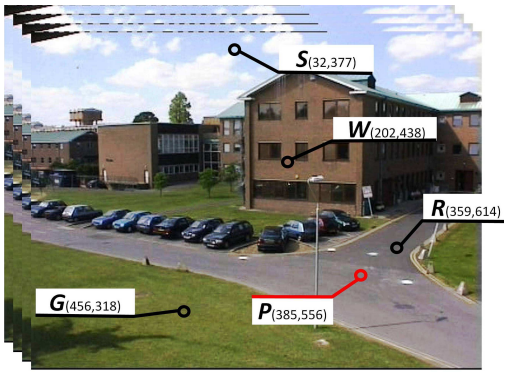
$$\boldsymbol{h}_{PQ} = \{h_{PQ}(i,j)\}_{i,j=0}^{L-1}. \tag{2.2}$$

As an example using PET2001-dataset3 camera1[2] shown in Figure 2.3, we selected a target pixel P located on the "road", and four arbitrary pixels S, G, W and R from the "sky", "grass", "wall" and "road" respectively. The section $h_{PQ}(i,j) > 0$ of joint histograms are illustrated as a few representable examples in Figure 2.3 (c-f). $\boldsymbol{h}_{PS}$ shows the most irregular distribution, while $\boldsymbol{h}_{PG}$, $\boldsymbol{h}_{PW}$ and $\boldsymbol{h}_{PR}$ reveal a more regular distribution. It is obvious that the most regular distribution is the histogram $\boldsymbol{h}_{PR}$ shown in Figure 2.3 (f), and its co-occurrence bins are parallel to a diagonal line running downwards through the histogram. The corresponding intensity time sequences of the four pixel pairs are shown in Figure 2.4. In Figure 2.4 (b, c), the intensity changes show an unexpected phase difference between the two time sequences, which disturbs the stable distribution of the joint histograms $\boldsymbol{h}_{PG}$ and $\boldsymbol{h}_{PW}$ shown in Figure 2.3 (d, e). For example, when a light source is shifting repeatedly, the shift speed would vary between different times, which would cause such a phase difference.

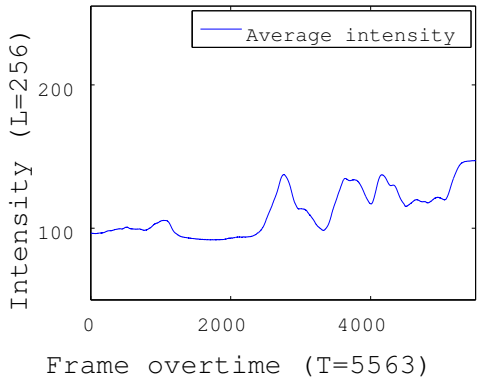
The corresponding joint histograms in Figure 2 (d, e) indicate that, under illumination fluctuation, the intensity relationships of the pixel pair (P,G) and (P,W) follow multiple varying modes, which essentially releases the relationship constraint of the pixel pair. If we employ such a pixel pair to model the background, the sensitivity to detect objects will be low, because the background model would be so flexible that a large range of intensity pairs would gather to model illumination changes. On the other hand, plenty of pixels have a simpler and more explicit statistical relationship with the target pixel as shown by the pixel pair (P,R) in Figure 2 (f). Therefore, R can be selected as a *supporting pixel* Q^P to observe the intensity of P even under illumination changes. The training stage of the proposed method starts from selecting supporting pixels Q^P for each target pixel P, the associated calculation procedure is presented in the next subsection.

[1] CP3 can also be called the pixel correlation model. However, we call it Co-occurrence Probability based Pixel Pairs because it is designed on the basis of co-occurrence histogram of a pixel pair, and the following calculations also rely on the probability of the histogram bins.

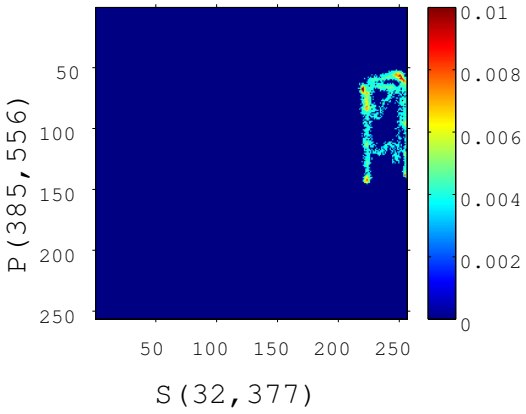
[2] `ftp://ftp.pets.rdg.ac.uk/pub/PETS2001/DATASET3/`



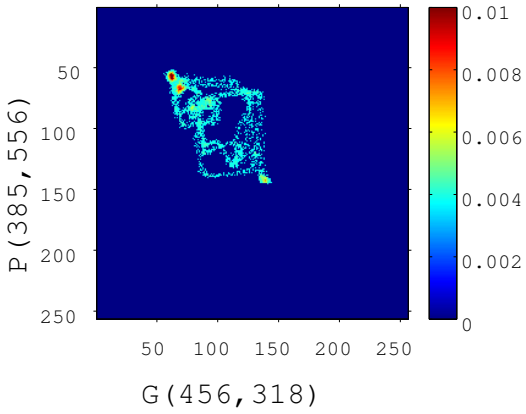
(a) Location of the target pixel P selected from the road, and four arbitrary pixels S, G, W, and R, selected from the sky, grass, wall, and road.



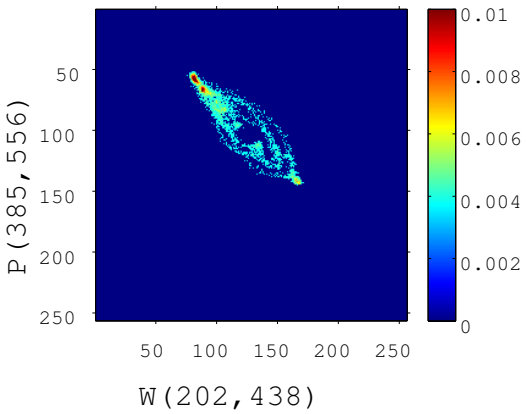
(b) The average intensity change of each frame in this scenario.



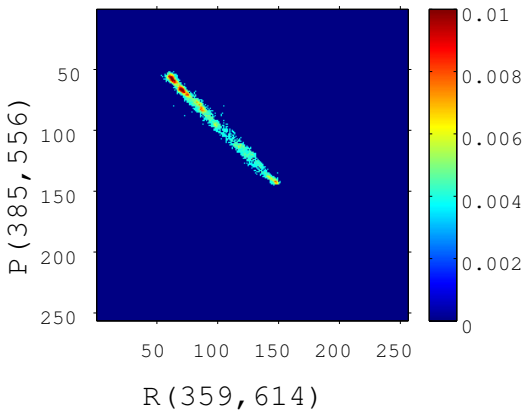
(c) The joint histogram $\boldsymbol{h}_{PS}$



(d) The joint histogram $\boldsymbol{h}_{PG}$

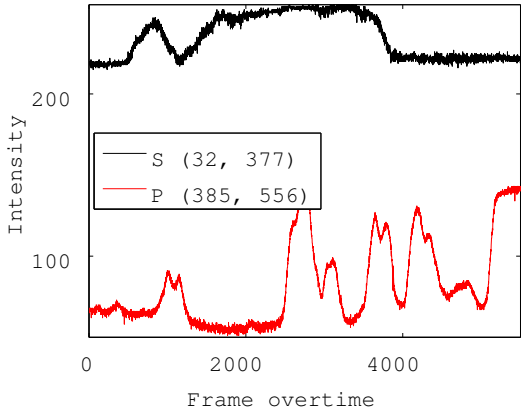


(e) The joint histogram $\boldsymbol{h}_{PW}$

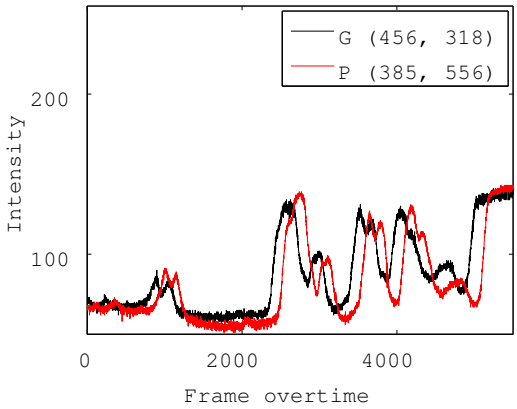


(f) The joint histogram $\boldsymbol{h}_{PR}$

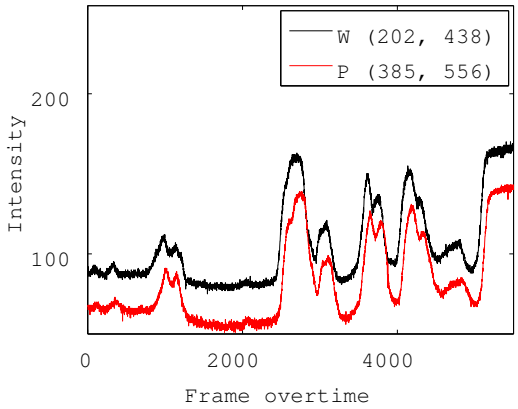
Figure 2.3: Co-occurrence joint histograms use PETS2001-dataset3 camera1 training dataset with 5563 frames ($T = 5563$) and the gray-scale level range is $[0, 255]$ ($L = 256$).



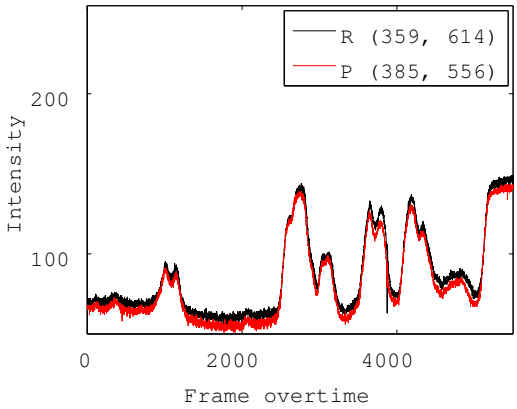
(a) The intensity change of P and S



(b) The intensity change of P and G



(c) The intensity change of P and W



(d) The intensity change of P and R

Figure 2.4: The intensity change of target pixel P and four arbitrary pixels corresponding to Figure 2.3 (c-f).

As a grayscale/single-channel image, the change in pixel intensity is proportional to the illumination increment. Hence, the statistical linearity of a pixel pair reduces to a stable intensity differential increment $\Delta(p_t, q_t)$ just as the example in Figure 2.3 (f), in which the slope of the regression line approaches 1.

For one target pixel P, it is natural to expect that one or more pixels Q^P which maintain a stable intensity differential increments $\Delta(p_t, q_t)$ during training frames, exist even though P and Q^P might be at quite different locations. When detecting objects under a dynamic background, both the object occupation and the illumination change (or other forms of dynamics) can affect on the current intensity of a target pixel P. A background model only modelling the independent intensity change of P cannot distinguish the object from the dynamic background. Instead, Q^P pixels could be employed to estimate the intensity of the target pixel P in a current detection frame, i.e. $\hat{p} = \Delta(p_t, q_t) + q$, where q is the intensity of Q^P in a current detection frame. When the illumination changes on a target pixel P but no object exists on it, $\hat{p}$ simultaneously changes with q, so that the current intensity p will fall into the estimated range $\hat{p}$, then P will be considered to be a background element. If P is occupied by an object, the current intensity p will be out of the estimated range $\hat{p}$, then P will be considered to be occupied by an object.

In our training stage, $\Delta(p_t, q_t)$ is modelled as a single Gaussian model, with a unique mean and variance. Using a Gaussian model allows the pixel pair to tolerate noise. More importantly, for different pixel pairs, the mean of $\Delta(p_t, q_t)$ could vary and the noise effect could also be at different ranges, so the mean and variance of a Gaussian calculated on specific pixel pair provide an accurate statistical constraint between them.

For robust detection, it is necessary to maintain a sufficient number of *supporting pixels*, denoted $\{Q_k^P\}_{k=1,2,...,K}$, where K is the total number of supporting pixels. In our detection stage, a double layer probability-based decision is used to detect objects: the first layer is to identify whether an individual pixel pair $(P,\ Q_k^P)$ matches its Gaussian model; and the second layer is to observe a matched ratio, i.e. to count the number of pixel pairs $(P, \{Q_k^P\})$ that can match their Gaussian model from total K pixel pairs for each P. Once the matched ratio decreases, P will be regarded as a foreground element. The associated calculation procedure is presented in Section 2.3.

2.2.2 Statistical measurement of co-occurrence pixel pairs

In this section, we introduce how to determine the supporting pixels from $M - 1$ candidate pixels for each target pixel P. For an arbitrary pixel pair $(P,\ Q)$, the one dimensional histograms corresponding to their marginal probability distributions are

$$h_P(i) = \sum_{j=0}^{L-1} h_{PQ}(i, j) \tag{2.3}$$

and

$$h_Q(j) = \sum_{i=0}^{L-1} h_{PQ}(i, j). \tag{2.4}$$

The expectation values of P and Q are

$$\mathcal{E}(p_t) = T^{-1} \sum_{i=0}^{L-1} i h_P(i) \tag{2.5}$$

and

$$\mathcal{E}(q_t) = T^{-1} \sum_{j=0}^{L-1} j h_Q(j) \tag{2.6}$$

respectively, and their variances are

$$\sigma_{p_t}^2 = \frac{1}{T} \sum_{i=0}^{L-1} [i - \mathcal{E}(p_t)]^2 h_P(i) \tag{2.7}$$

and

$$\sigma_{q_t}^2 = \frac{1}{T} \sum_{j=0}^{L-1} [j - \mathcal{E}(q_t)]^2 h_Q(j). \tag{2.8}$$

The covariance of a $(P,\ Q)$ pair can be defined as follows:

$$C_{P,Q} = \frac{1}{T} \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [i - \mathcal{E}(p_t)][j - \mathcal{E}(q_t)] h_{PQ}(i, j). \tag{2.9}$$

$C_{P,Q} > 0$ means P and Q has positive covariance value, which indicates they potentially have a high co-occurrence probability. We call this kind of pixel pair co-occurrence pixel pair hereafter. In order to measure the independent co-occurrence quantitatively, we utilize Pearson product-moment correlation coefficient:

$$\gamma_{(P,\ Q)} = \frac{C_{P,Q}}{\sigma_{p_t} \cdot \sigma_{q_t}} \tag{2.10}$$

where σ_{p_t} and σ_{q_t} are the standard deviations of P and Q respectively. Eq. (2.10) is used to estimate the linear dependence of spatial pixel pairs, and $\gamma_{(P,\ Q)}$ is ± 1 in the case of a perfect positive/negative linear correlation. This normalized correlation coefficient does not involve the restraint between $\mathcal{E}(p_t)$ and $\mathcal{E}(q_t)$. Namely, the pixels Q with any mean values are likely to have large $\gamma_{(P,\ Q)}$ with P.

Figure 2.5 shows four examples of $\gamma_{(P,\ Q)}$ using PETS2001-dataset3, the black crosses stand for the location of P, and the red coloured area have high correlation coefficient values.

In practice, Eq. (2.10) can be calculated based on a correlation matrix instead of calculating pixel-by-pixel serial processing. The correlation matrix is the covariance matrix of the standardized random variables $\tilde{p}_t = p_t / \sigma(p_t)$. With a total of $M = U \times V$ pixel positions, the image sequence can be arranged progressively as a column vector set $\chi^M = \{\tilde{p}_t(m)\}_{m=1,2,\ldots,M}$.

The correlation matrix of size $M \times M$ is

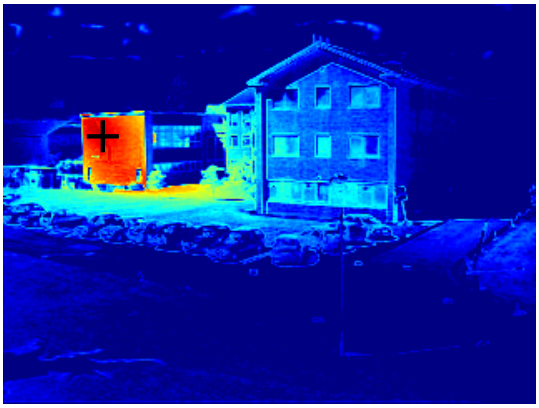
$$\begin{aligned} \Upsilon(\chi^M) &= C(\chi^M, (\chi^M)^T) \\ &= \begin{bmatrix} \tilde{p}_t(1) - \mathcal{E}(\tilde{p}_t(1)) \\ \tilde{p}_t(2) - \mathcal{E}(\tilde{p}_t(2)) \\ \vdots \\ \tilde{p}_t(M) - \mathcal{E}(\tilde{p}_t(M)) \end{bmatrix} \cdot \begin{bmatrix} \tilde{p}_t(1) - \mathcal{E}(\tilde{p}_t(1)) \\ \tilde{p}_t(2) - \mathcal{E}(\tilde{p}_t(2)) \\ \vdots \\ \tilde{p}_t(M) - \mathcal{E}(\tilde{p}_t(M)) \end{bmatrix}^T \end{aligned} \tag{2.11}$$

where $C(\cdot)$ is the covariance operation. The correlation matrix is symmetric so that each row and column of the $\Upsilon(\chi^M)$ is an array of $\gamma_{(P,\ Q)}$ for each $P(u,v)$.

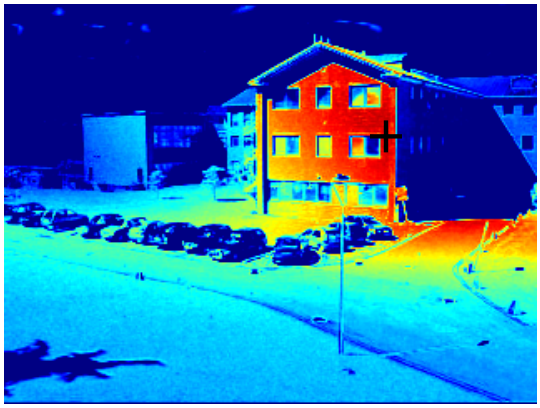
For each target pixel $P(u,v)$, $M-1$ values of $\gamma_{(P,\ Q)}$ need to be calculated at different locations (u',v'). Then Q_n corresponding to the highest N components in the array $\gamma_{(P,\ Q(u',v'))}$ can be selected as the candidates of preferred supporting pixels, namely

$$\{Q_n\} = \{Q(u',v') | \gamma_{(P,\ Q)} > \check{\gamma}\},\ n = 1,2,...,N, \tag{2.12}$$

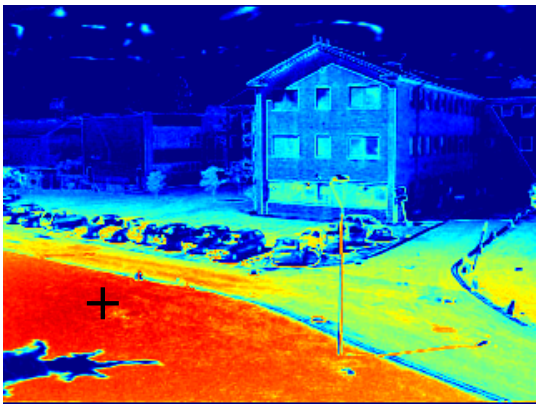
where N is the number of candidate supporting pixels Q_n for each target pixel P, and $\check{\gamma}$ is the lower limit for the co-occurrence pixel pair. $\check{\gamma}$ is an adaptive threshold depending on the variation characteristic of P, which is discussed in the following Section.



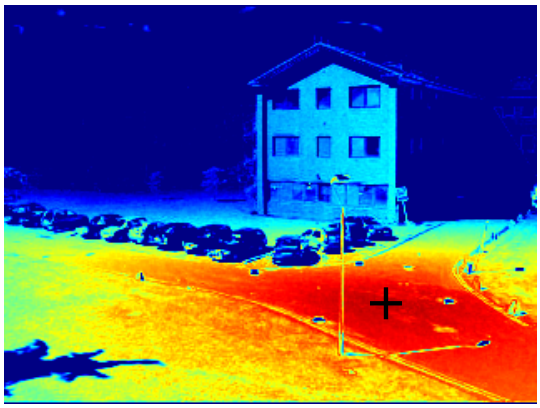
(a) Target pixel 1 and its $\gamma_{(P,\ Q)}$



(b) Target pixel 2 and its $\gamma_{(P,\ Q)}$



(c) Target pixel 3 and its $\gamma_{(P,\ Q)}$



(d) Target pixel 4 and its $\gamma_{(P,\ Q)}$

Figure 2.5: Diagram of correlation coefficients $\gamma_{(P,\ Q)}$ using PETS2001-dataset3-camera1 dataset. The black crosses stand for the locations of P, and the red coloured area have high correlation coefficient values.

2.2.3 Lower limit for a co-occurrence pixel pair

Due to sensor noise and encoding noise, any p_t and q_t cannot maintain a full co-occurrence relation. Therefore, the lower limit $\check{\gamma}$ for choosing the high co-occurrence pixel pairs is a key parameter. This parameter is generally solved through preparatory experiments using typical samples of target images as a training dataset. To estimate the lower limit more reasonably, we need a theoretical formalization as a guidance with a natural prospective view.

Our approach to formalization is to assume that, $p_t = p_t' + e_1$ and $q_t = q_t' + e_2$, where p_t' and q_t' are the intensities without any noise; e_1 and e_2 are the additive noise, independent of each other but with the same density function $\mathcal{N}(0, \sigma_n^2)$. Then we assume p_t' and q_t' have a perfect positive linear correlation with a constant $b = \Delta(p_t', q_t')$, namely $p_t' = q_t' + b$, and analyse $\check{\gamma}$ as a statistic for investigating how large degradation is raised by the noise. For the computation of $\gamma_{(P,\,Q)}$, disconcordance between p_t and q_t can degrade the $\check{\gamma}$ value. The correlation coefficient $\check{\gamma}$ can be represented by the next expression according to Eq. (2.10),

$$\check{\gamma} = \frac{C(p_t' + e_1, p_t' + e_1 - e_2 - b)}{\sigma_{p_t'+e_1} \cdot \sigma_{p_t'+e_1-e_2-b}}. \tag{2.13}$$

According to the properties of covariance and variance, the above formula is developed as

$$\check{\gamma} = \frac{\sigma_{p_t'}^2 + \sigma_n^2}{\sigma_{p_t'+e_1} \cdot \sigma_{p_t'+e_1-e_2-b}}. \tag{2.14}$$

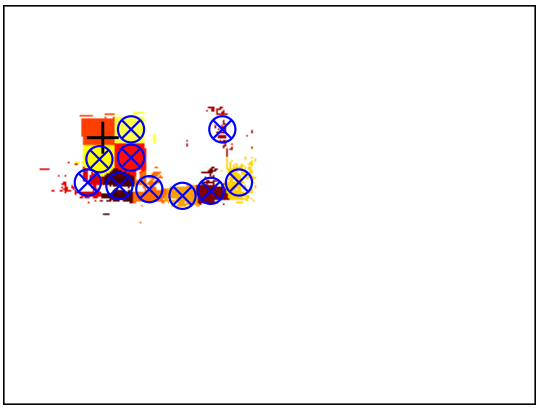
When p_t' is independent of e, Eq. (2.14) can be rewritten as

$$\begin{aligned}
\check{\gamma} &= \frac{\sigma_{p_t'}^2 + \sigma_n^2}{[(\sigma_{p_t'}^2 + \sigma_n^2)(\sigma_{p_t'}^2 + 2\sigma_n^2)]^{\frac{1}{2}}} \\
&= \left(\frac{\sigma_{p_t'}^2 + \sigma_n^2}{\sigma_{p_t'}^2 + 2\sigma_n^2} \right)^{\frac{1}{2}} \qquad (2.15) \\
&= \left(\frac{\sigma_{p_t}^2}{\sigma_{p_t}^2 + \sigma_n^2} \right)^{\frac{1}{2}} \\
&= \left(1 + \frac{\sigma_n^2}{\sigma_{p_t}^2} \right)^{-\frac{1}{2}}. \qquad (2.16)
\end{aligned}$$

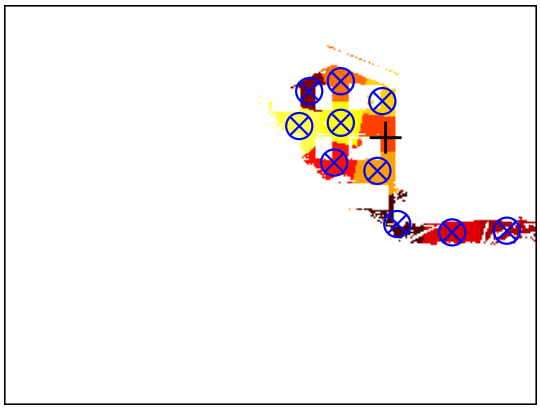
When the noise level is significantly smaller than the dynamic range of p_t, namely $\sigma_{p_t'}^2 \gg \sigma_n^2$, Eq. (2.15) approximates to 1, which reveals that with large-scale intensity variation in the training dataset, the noise effect for correlation measurement can be reduced. On the other hand, if the intensities of P keep steady, namely $\sigma_{p_t'}^2 \to 0$, Eq. (2.15) will level off to $1/\sqrt{2}$, then the candidate supporting pixels can be selected from the stationary elements of the background. We directly use Eq. (2.16) instead of Eq. (2.15), in which $\sigma_{p_t}^2$ can be calculated based on Eq. (2.7) and σ_n^2 can be determined according to the noise level of the image sequence.

From the theoretical analysis, the lower limit can be determined from the comprehensive conditions combining with σ_n^2 which can be easily provided by users and a computable $\sigma_{p_t}^2$.

Figure 2.6 shows four example of selected Q_n (the coloured area), which demonstrate that the rules to choose Q_n according to the lower limit $\check{\gamma}$ allow the spatial distributions of Q_n to follow irregular illumination variation patterns.

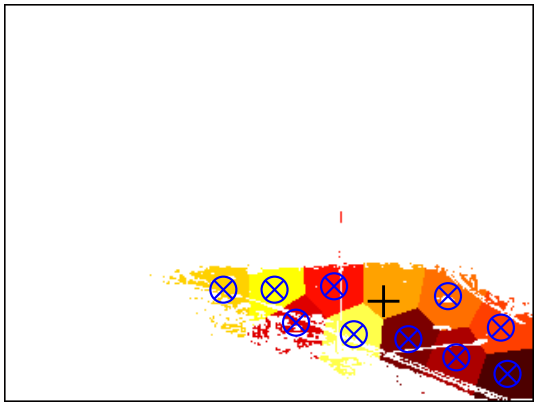


(a) Target pixel 1 and its Q_k^P



(b) Target pixel 2 and its Q_k^P

(c) Target pixel 3 and its Q_k^P



(d) Target pixel 4 and its Q_k^P

Figure 2.6: Diagram of selecting Q_k^P using K-means in spatial domain. The black markers correspond to the P pixels in Figure 2.5 (a-d). The different-coloured areas, which stand for the clustering subsets, come from the high co-occurrence pixel Q_n with total N pixels. The blue circles are the centres of each clustering subset, which are selected as Q_k^P. In these examples, $K = 10$.

2.2.4 Background model of pixel pairs

In this section, we discuss how to produce a limited number of supporting pixels Q_k^P and build the background model.

In Section 2.2.3, the selected Q_n using $\check{\gamma}$ resulted in an indeterminate number of N. To perform an executable algorithm, we need to confirm a series of a limited number of supporting pixels. As the spatial distribution of Q_n follows irregular patterns, we cannot implement any former spatial interpolation approach to select high representative Q_k^P from Q_n. To solve this issue, K-means clustering is employed to partition Q_n into K clusters, depending on the nearest clustering centres [57]. With clustering convergence, the pixel that is closest to the k-th cluster centre is selected as a unique Q_k^P. The details of K-means clustering for optimizing the spatial distribution is described in the Appendix. Four demonstrations of the Q_k^P optimization are shown in Figure 2.6 in which $K = 10$. It is reasonable to assume that selecting more supporting pixels will contribute to a robust result. The supporting pixels are essentially a group of statistical samples. Theoretically speaking, a larger number of samples would estimate a more robust statistical model. On the other hand, the number of supporting pixels directly affects the computation cost for object detection. This issue is discussed in the form of quantitative experiments in Section 2.4.2. Without loss of generality, the number of K for a given video scene is set at 10 for the experimental comparison in Section 2.4.1 and 2.4.3.

Each Q_k^P keeps a bivariate differential increment with P,

$$p_t \sim \mathcal{N}(q_{t(k)} + b, \sigma_\varepsilon^2), \tag{2.17}$$

where σ_ε^2 follows a normalized distribution $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. We use this Gaussian function to model the intensity distribution of a pixel pair instead of a Gaussian mixture model [11] because we find that a single Gaussian works better since the selected pixel pair keeps highly steady difference except for noise. We calculate a noise standard deviation estimation as follows,

$$\hat{\sigma}_\varepsilon = \sigma_{p_t - q_{t(k)}}, \tag{2.18}$$

and the estimation of the differential increment b is,

$$\hat{b} = \mathcal{E}[p_t - q_{t(k)}]. \tag{2.19}$$

After the training step, the above two parameters $\hat{\sigma}_\varepsilon$, $\hat{b}$ are recorded for the following detection procedure. The background model is a look-up table (LUT) consisting of $[\, u', v', \hat{\sigma}_\varepsilon, \hat{b}\,]$ for $\{Q_k^P\}_{k=1,2,\ldots,K}$.

To train the background model of a scene, a training dataset including the dynamic background of the scene is necessary. In this work, we use three public datasets (PETS2001-dataset3 camera1, AIST-INDOOR, Wallflower), and each of them provides a separate training and detection dataset. We use each specified training and detection dataset for training and detection respectively.

2.2.5 Moving background case

In the case of a moving background, the moving parts cover several pixels in the same frame that also present co-occurrence. Therefore, we can search for the supporting pixels Q_k^p if the intensity changes of the pixel pairs are simultaneous. Using the proposed method to evaluate the intensity changes caused by a moving background makes no difference within the case of illumination variation. Hence, the earlier discussion based on illumination change is also appropriate for the case of a moving background.

A typical motion pattern in backgrounds is *burst motion*. This motion pattern can be described as a moving part of the background following regular directions but with an irregularly scheduled occurrence; hence, its speed and frequency cannot be directly predicted. Plenty of moving background elements can be simplified into burst motion. For example, tree swinging caused by random wind, fan speed change, dynamic horizontal lines of a displayer caused by its refresh rate, and auto-induction escalators and doors, all of which often appearing in outdoor/indoor surveillance scenes are examples of the burst motion. In general, applying independent pixel-wise methods (such as GMM [11] or Codebook [13]) to deal with motion backgrounds only employ pixel's history by continuously updating background, using a fixed learning rate as the background updating criterion; these background models are sensitive to burst motion. Our proposed method is a frequency and speed adaptive background model, which employs the spatial-dependence of pixel pairs to keep a stable differential increment regardless of the intensity of a single pixel under any frequency or speed of burst motion. The selected pixel pairs convert the non-stationary scene to a stable background model for offsetting the patterns of motion without any learning rate. Figure 2.7 shows four examples of various burst motion backgrounds without severe illumination change. Figure 2.7 (a) and (b) are from the Wallflower dataset [19], and contain a waving tree and a cathode ray tube displayer respectively; Figure 2.7 (c) and (d) contain an escalator in operation and a fan in operation respectively, which can be downloaded from the author's web links[3] [4].

Compared with the case of severe illumination change shown in Figure 2.5, $\gamma_{(P,\,Q)}$ appear a part of negative values along the vertical direction of the movement locus shown in Figure 2.7.

Figure 2.8 shows examples of the mixture case of burst motion and sudden illumination change using AIST-INDOOR dataset[5].

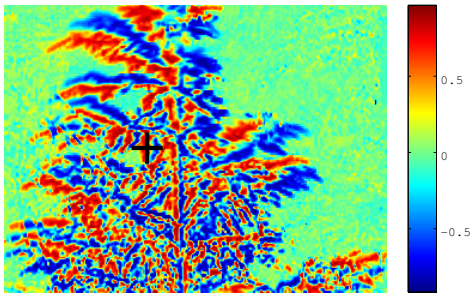
In this scene, a target pixel repeatedly passes by an auto-induction door while a light turning on. The supporting pixels with high co-occurrence are located along the vertical direction of the door to meet the simultaneity of its burst motion (shown in Figure 2.8 (b-d)), rather than around a regular neighborhood. As a comparison, Figure 2.8 (e, f) shows the case of a target pixel in a static part of the same scene. The pseudo-code of CP3 for background modelling is shown in Algorithm 1.

[3] `http://ssc-lab.com/~liang/CP3_project/escalator.rar` (The domain name `ssc-lab.com` would be changed to `www.hce.ist.hokudai.ac.jp`)

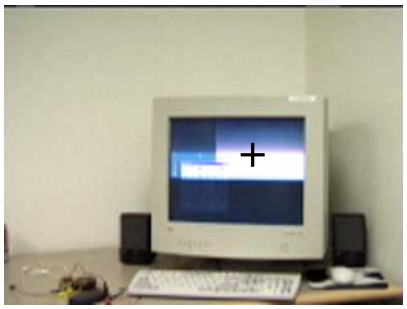
[4] `http://ssc-lab.com/~liang/CP3_project/fan.rar`

[5] `http://ssc-lab.com/~liang/CP3_project/AIST_INDOOR_DATASET.rar`

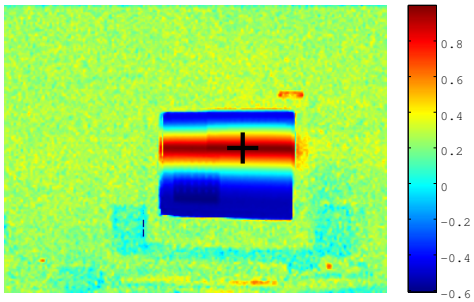
(a) Tree swing.



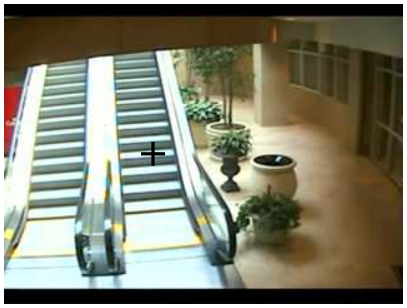
(b) $\gamma_{(P,\ Q)}$ values of a target pixel in (a).



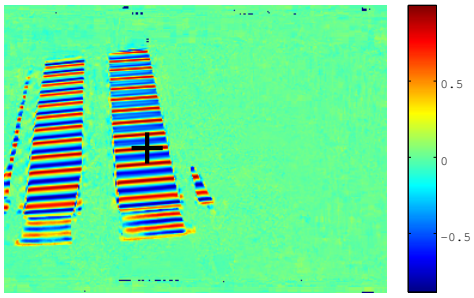
(c) Dynamic refresh of a displayer.



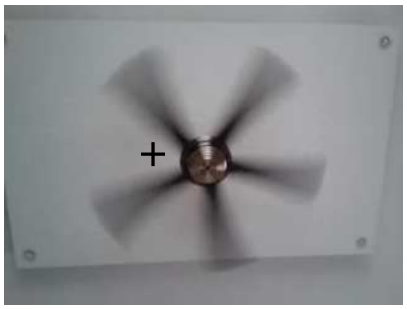
(d) $\gamma_{(P,\ Q)}$ values of a target pixel in (c).



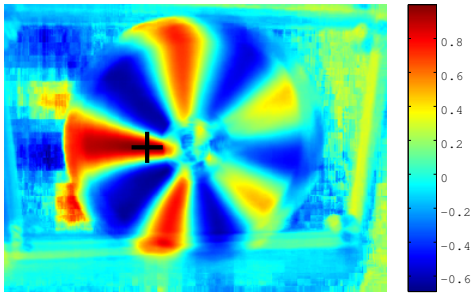
(e) Auto-induction escalator.



(f) $\gamma_{(P,\ Q)}$ values of a target pixel in (e).

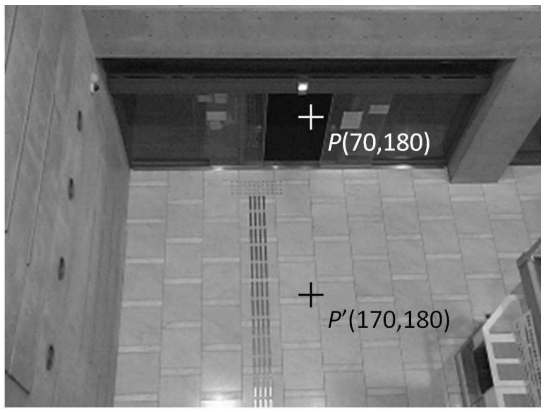


(g) Speed-adjustable fan.

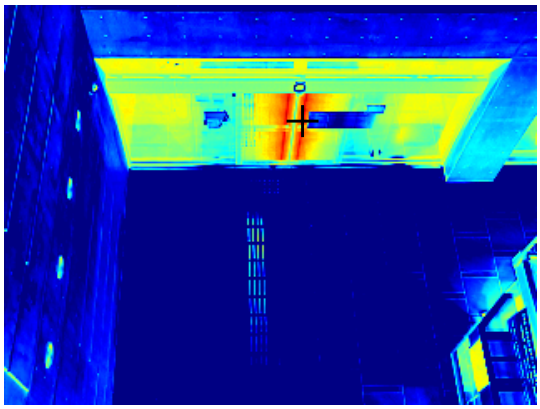


(h) $\gamma_{(P,\ Q)}$ values of a target pixel in (g).

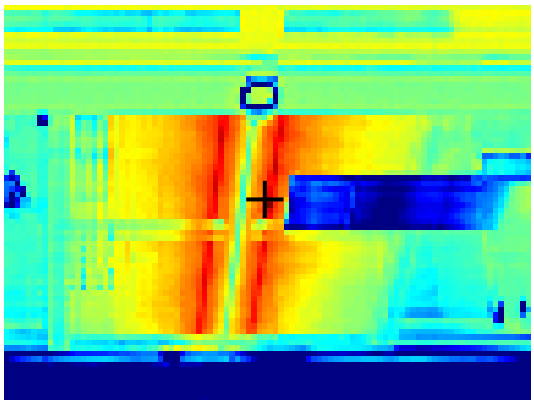
Figure 2.7: Examples of various burst motion. Correlation coefficients $\gamma_{(P,\ Q)}$ values of a selected pixel. The black crosses stand for the locations of P, and the red coloured area have high correlation coefficient values.



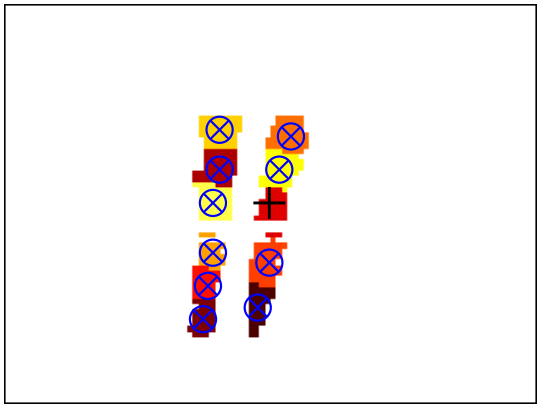
(a) Location of the target pixels $P(70, 180)$ and $p'(170, 180)$.



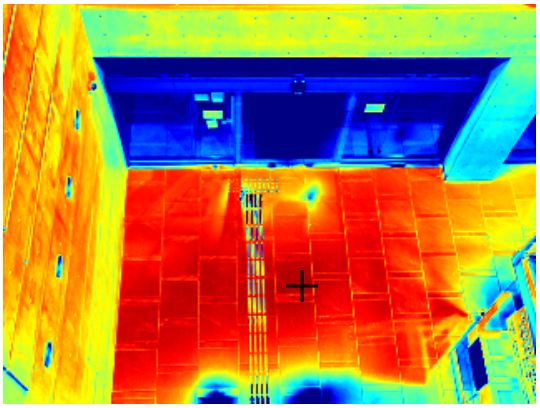
(b) $\gamma_{(P,\ Q)}$ of $P(70, 180)$.



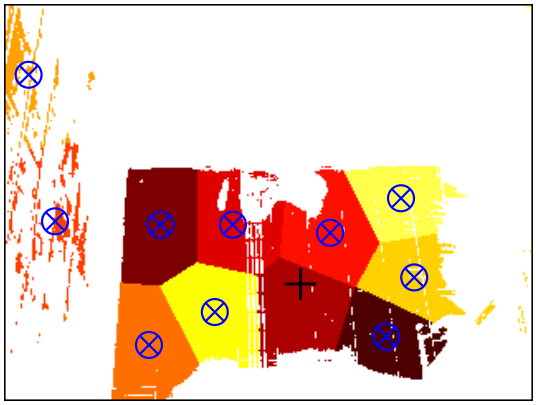
(c) Partial enlarged drawing of (b).



(d) The supporting pixels of (c).



(e) $\gamma_{(P,\ Q)}$ of $P'(170, 180)$.



(f) The supporting pixels of (e).

Figure 2.8: Examples of the mixture of burst motion and sudden illumination change. Correlation coefficients $\gamma_{(P,\ Q)}$ values of a selected pixel. The black crosses stand for the locations of P, and the red coloured area have high correlation coefficient values.

Algorithm 1: CP3 for background modelling

Data: $B = \{I_1, I_2, ..., I_T\}$ with total T images, and σ_n^2.

Result: look-up table of $\{Q_k^P\}_{k=1,2,...,K}$ consisting of $[u', v', \hat{\sigma}_\varepsilon, \hat{b}]$.

Initialization:

Build the vector set $\chi^M = \{\tilde{p}_t(m)\}_{m=1,2,...,M}$.

(I) Compute correlation matrix:

$\Upsilon(\chi^M) = C(\chi^M, (\chi^M)^T)$.

(II) Select supporting pixels:

for *each* $P(u, v)$ **do**

(a) Compute $\check{\gamma}$ based on Eq. (2.16).

(b) $\{Q_n\}_{n=1,2,...,N} = \{Q | \gamma_{(P,\,Q)} > \check{\gamma}\}$.

(c) K-means sampling in spatial domain $\{Q_n\} \Rightarrow \{Q_k^P\}_{k=1,2,...,K}$.

(d) Compute and record $[\, u', v', \hat{\sigma}_\varepsilon, \hat{b}\,]$ for $\{Q_k^P\}_{k=1,2,...,K}$.

2.2.6 Accelerated background modelling

In Section 2.2.2, in order to calculate $\gamma_{(P,\,Q)}$ for each P in a training dataset with $M = U \times V$ pixel positions and T frames, the computational complexity for Eq. (2.11) is $O(TM^2)$. In contrast, in the independent pixel-wise models [11, 12], it is unnecessary to consider spatial dependence, so the computational complexity is only $O(TM)$ in the training stage. With the same memory cost, the time consumption of our proposed method is $M - 1$ times higher than independent pixel-wise models.

Hoping to have a faster version of background modelling, we modified Eq. (2.11) using a hierarchical structure of a covariance-matrix: χ^M can be sampled uniformly using an integral sample interval Λ, the sub-set $\chi^{[M/\Lambda^2]} \subset \chi^M$; thus, we have

$$\Upsilon(\chi^{[M/\Lambda^2]}) = C(\chi^{[M/\Lambda^2]}, (\chi^{[M/\Lambda^2]})^T). \tag{2.20}$$

In order to cover all the target pixels, we have Λ^2 hierarchical correlation matrices $\Upsilon(\chi_\lambda^{[M/\Lambda^2]})$ and

$$\chi_\lambda^{[M/\Lambda^2]} = \{\tilde{p}_t(\omega\Lambda^2 + \lambda)\}_{\omega=1,2,...,[M/\Lambda^2]}, \tag{2.21}$$

where $\lambda = 1, 2, ..., \Lambda^2$. In this way, both the memory cost and time consumption is $O(TM^2\Lambda^{-2})$, which means the hierarchical structure of a covariance-matrix reduces computational complexity by Λ^2.

The speed-up algorithm is based on the pixel sampling operation, which reduces the number of candidates of supporting pixels. However, we are willing to accept the possible loss in information, because it allows us to achieve high speed processing when dealing with high resolution surveillance videos. We recommend that the high resolution applications choose the accelerated background modelling, and the low data-volume applications choose the standard CP3 algorithm. The experimental discussion of the sample interval Λ is presented in Section 5.3.

2.3 Object detection

The proposed background model converts the object detection problem into a competitive binary classification problem [58] by comparing the intensity pairs $(P, \{Q_k^P\}_{k=1,2,...,K})$ in turn. It includes two stages: (1) to identify the normal/abnormal state of the pixel pair $(P,\ Q_k^P)$; (2) to identify the foreground/background state of P.

2.3.1 Pixel pair state

For each pixel pair $(P,\ Q_k^P)$, the binary function $\beta(Q_k^P)$ for discriminating the normal/abnormal state can be estimated as the following condition according to Eq. (2.17):

$$\beta(\{Q_k^P\}_{k=1,2,...,K}) = \begin{cases} 1 & \text{if } \ |(p - q_k) - \hat{b})| < C \cdot \hat{\sigma}_\varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{2.22}$$

where p and q_k are the intensity values of P and Q_k^P in the input frame J respectively, and C is a constant. It is important to note that using the bivariate normal distribution of the differential increment of the pixel pair is different from using the traditional single Gaussian pdf-based identification function of a single pixel [10]; in a single Gaussian pdf-based method, the ideal threshold should be changed following the latest intensity variation. For example, the standard deviation should be larger when the illumination fluctuations become more intense. In our proposed version, the stable differential increment of a pixel pair provides an adaptive observation so that $\hat{\sigma}_\varepsilon$ is only related to the noise acting on each pixel. Therefore, we do not need an adjustable C to adapt to changes caused by illumination changes or background motion. The constant C can be set from 1.0 to 3.0 in order to contain approximately an area of 68-99% of its probability density function. In addition, the recording of $\hat{\sigma}_\varepsilon$ and $\hat{b}$ from the modelling stage provides a more accurate parametrized criterion than a fixed double-sided threshold in GAP method [29]. This will be confirmed in the experiment section.

2.3.2 Decision function

After identifying the normal/abnormal state of the pixel pair, K bits of $\beta(Q_k^P)$ are produced for the following decisions of each P. In order to classify whether P is a foreground pixel, the probability $\xi(P)$ of a pixel being in the background is defined as,

$$\xi(P) = \frac{1}{K} \sum_{k=1}^{K} \beta(Q_k^P). \tag{2.23}$$

Target pixel P in the input image is considered as a foreground pixel only if $\xi(P) < PF$, where PF is a global threshold that can be adjusted to achieve the desired result. Otherwise, P is considered as a background pixel. For instance, if $PF = 0.5$, and the number of abnormal Q_k^P is larger than $K/2$, namely, $\xi(P) < 0.5$, then P should be a foreground pixel in the input frame J. The procedure for calculating $\xi(P)$ for every target pixel is performed by a bits counting

Algorithm 2: Object detection

Data: Testing frame J, parameters C and PF.
Result: $Foreground/Background$
for *Each pixel P in J* **do**
  (I) Initialize:
  Load look-up table $\sim [u', v', \hat{\sigma}_\varepsilon, \hat{b}]$ of $\{Q_k^P\}_{k=1,2,...,K}$.
  (II) Pixel pair identification:
  for $k = 1, 2, ..., K$ **do**
    if $|(p - q_k) - \hat{b})| < C \cdot \hat{\sigma}_\varepsilon$ **then**
      $\beta(Q_k^P) = 1$
    else
      $\beta(Q_k^P) = 0$
  (III) Object identification:
  Compute $\xi(P) = \frac{1}{K} \sum_{k=1}^{K} \beta(Q_k^P)$.
  if $\xi(P) < PF, (0 < PF < 1)$ **then**
    $P \rightarrow Foreground$
  else
    $P \rightarrow Background$

operation, along with a look-up table for calculating $\beta(Q_k^P)$, which is easy to implement on any conventional hardware. The pseudo-code of the object detection is shown in Algorithm 2.

2.4 Experimental results

To evaluate the performance of the proposed method, we tested it on video datasets including a variety of indoor and outdoor environments for both qualitative and quantitative analysis. For all the experiments, $\sigma_n^2 = 100$ in the training stage and the two thresholds were set as $C = 2.5$ and PF=0.5 respectively in the detection stage.

For quantitative analysis, the three evaluation metrics, *Precision* (also known as positive predictive value), *Recall* (also known as sensitivity) and F-measure were utilized. *Precision*, *Recall* and F-measure are widely used in pattern recognition and information extraction with binary classification [59, 60]. Since pixel-level object detection is a typical binary classification problem, the three metrics also have been used for the quantitative analysis of object detection [14, 29, 61]. *Precision* can be seen as a measure of exactness or fidelity, and *Recall* can be seen as a measure of completeness of foreground,

$$Precision = \frac{TP}{TP + FP} \tag{2.24}$$

and

$$Recall = \frac{TP}{TP + FN} \tag{2.25}$$

where TP, FP and FN stand for the number of true positive pixels, false positive pixels and

false negative pixels, respectively. F-measure is a weighted harmonic mean of the *Precision* and *Recall*,

$$F = \frac{2Precision \cdot Recall}{Precision + Recall} \tag{2.26}$$

2.4.1 Comparison with other approaches

We compared our algorithm with three methods: (1) GMM [11, 16] method, which is a standardized method among independent pixel-wise models; (2) Sheikh's KDE method [14] as a representative method among spatially dependent models, which is different from the original nonparametric Kernel Density Estimation method (KDE) in that it employs KDE over the joint domain(location) and range (intensity) representation of image pixels; (3) our previous method GAP [29]. The parameters for the GMM algorithm were set as defaults in the OpenCV tool; the parameters for Sheikh's KDE algorithm were set according to the author's recommendations in [14], and the size of model is [26, 26, 26, 21, 31]; In GAP method, $W_G = 20, W_P = 0.9, W_H = 0.3$.

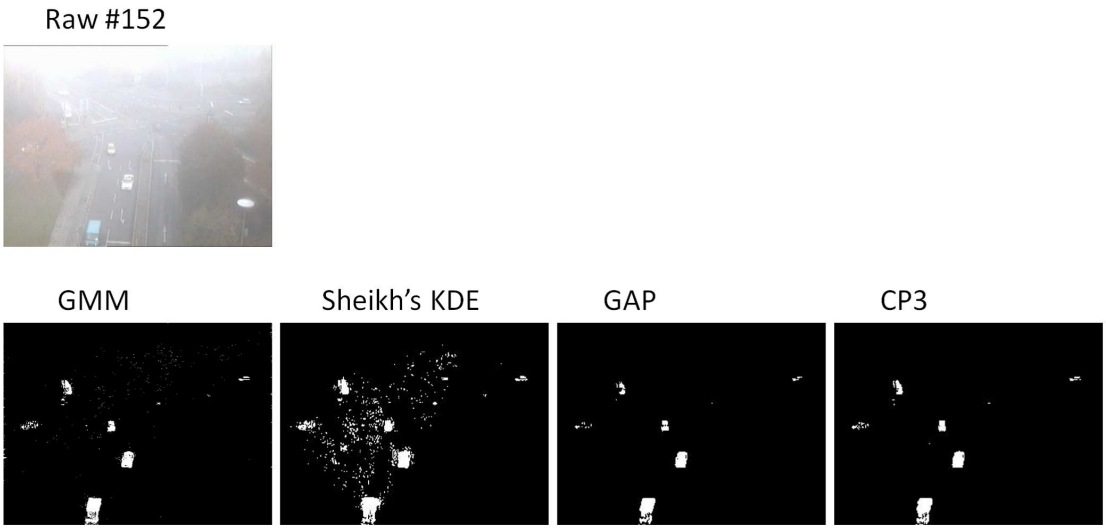
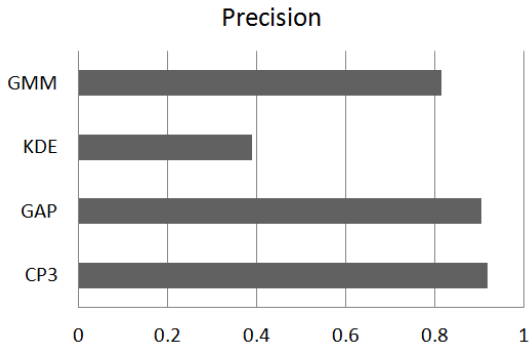


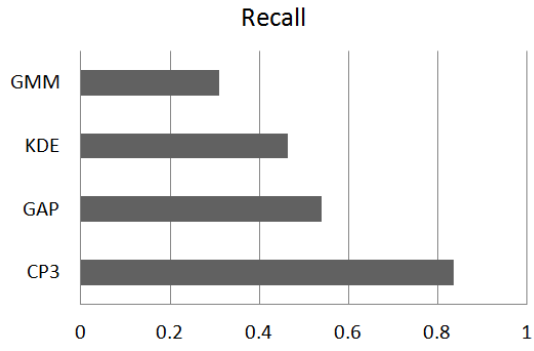
Figure 2.9: Qualitative comparison of GMM, Sheikh'KDE, GAP, and proposed CP3 method using a dataset of traffic sequence with heavy fog. The difficulty is that the heavy fog compresses the dynamic range of the scene.

First, we use a dataset of traffic sequence with heavy fog[6]. In which there is only gradually varied illumination but no burst motion background. The only difficulty is that the heavy fog compresses the dynamic range of the scene. The detection results are shown in Figure 2.9. All the four methods provide similar results in appearance. In the point of view of detection sensitivity, GAP method shows slight weak, because of the fixed threshold during the training and testing phase.

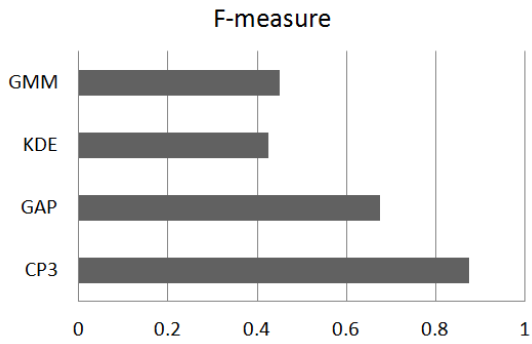
[6] http://i21www.ira.uka.de/image_sequences/



(a) *Precision*



(b) *Recall*



(c) F-measure

Figure 2.10: The mean metrics of GMM, KDE, GAP and proposed CP3 using PETS2001-dataset3 camera1.

Next, We use the sequences from PETS2001-dataset3 camera1 to test the outdoor scenes captured under severe illumination fluctuation. The sudden partial illumination variations caused by moving clouds are obvious in this scene which are clearly represented as average intensity change shown in Figure 2.13 (b). The dataset PETS2001-dataset3 camera1 includes a total of 5563 frames for training and 5336 frames for detection. The available ground truth data of PETS2001 allows us to complete the comparison[7]. The average *Precision*, *Recall* and F-measure of the four methods are shown in Figure 2.10. The *Precision*, *Recall* and F-measure over time of the four methods are shown in Figure 2.12 (a,b) and Figure 2.13 (c). Clearly, both CP3 and GAP show a higher level of *Precision* than the other two methods; CP3 has an obviously higher *Recall* and F-measure value than other methods which means it has higher sensitivity for detecting foreground.

It should be noted that, during the frames from 150 to 200, GAP and Sheikh's KDE methods show clearly decreasing performance of *Recall*, as the test video comes into a darker phase after

[7] `http://limu.ait.kyushu-u.ac.jp/en/dataset/`

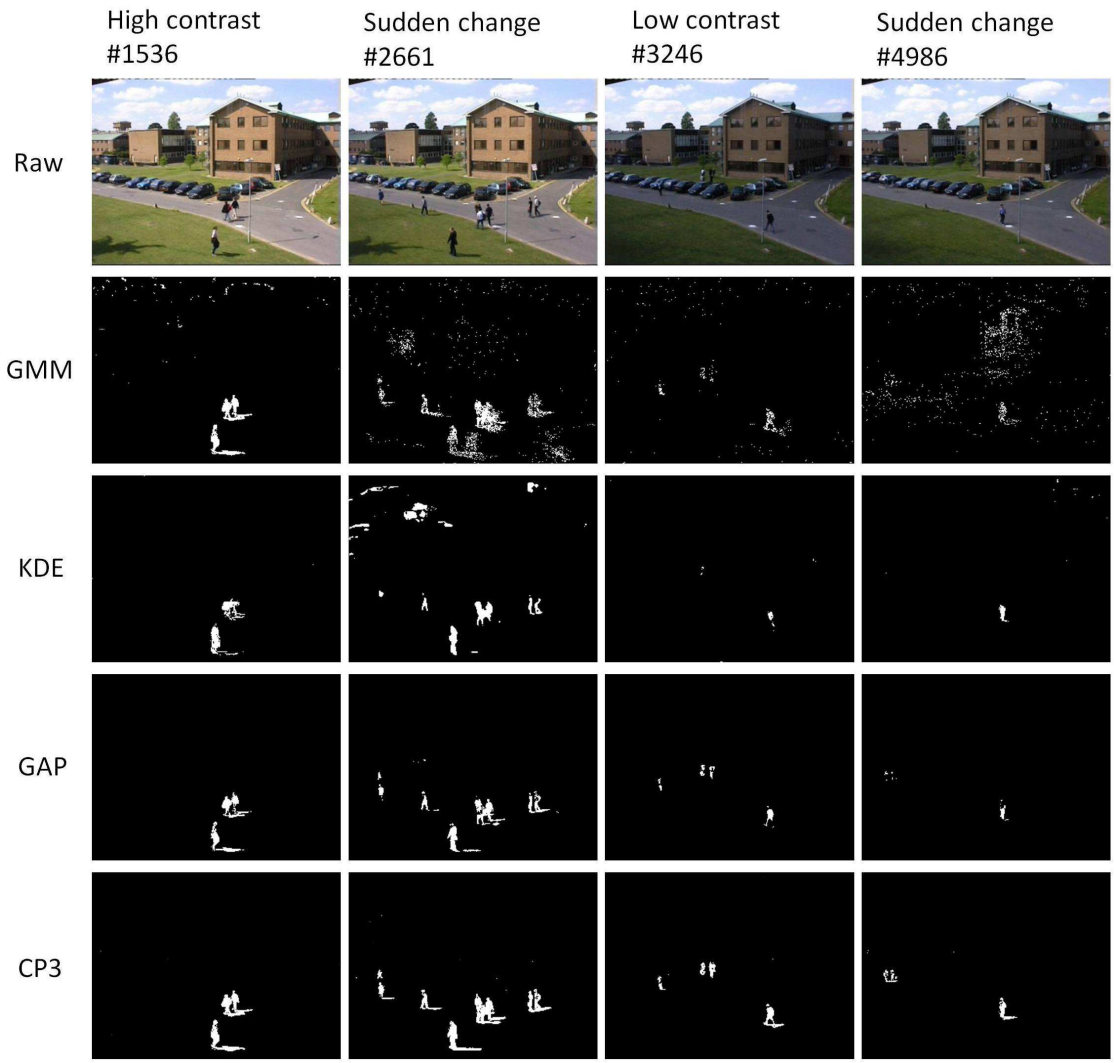
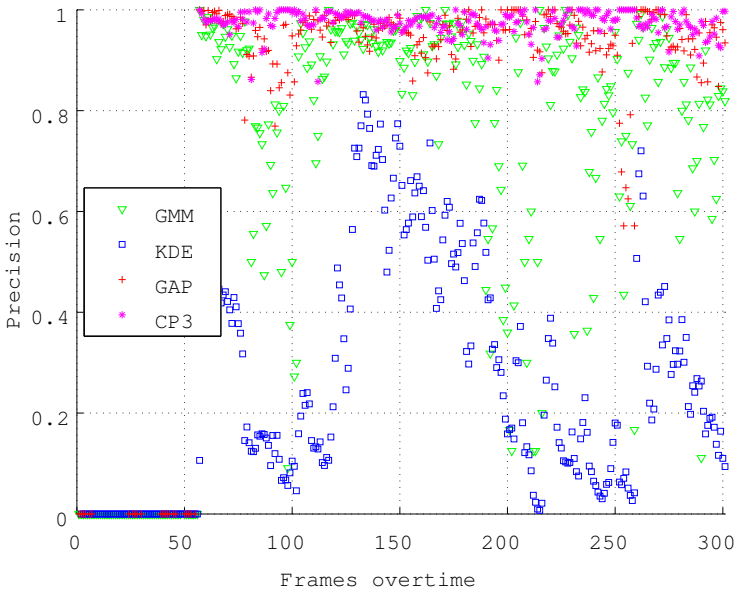


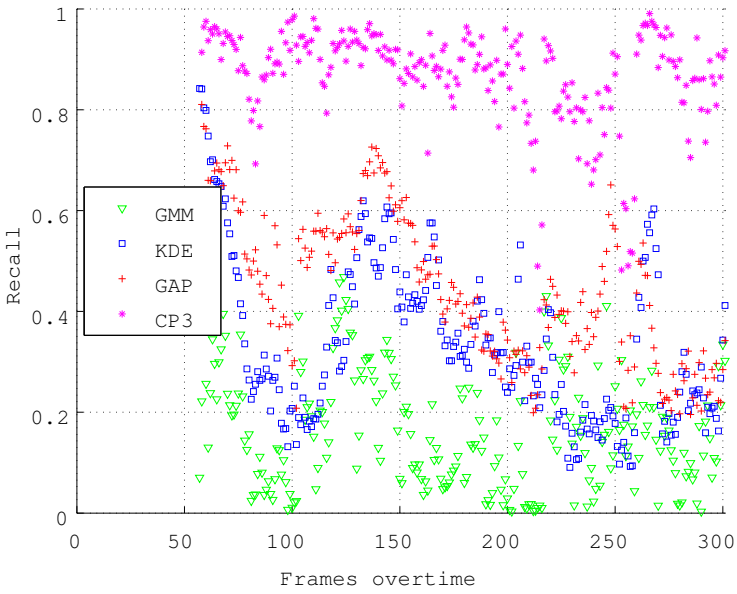
Figure 2.11: Qualitative comparison of GMM, Sheikh's KDE, GAP, and proposed CP3 using PETS2001-dataset3-camera1 dataset with outdoor severe illumination fluctuation.

the frame 150, and the dynamic range of the intensity is compressed, as shown in Figure 2.13 (d). Since the GAP method uses a fixed double-sided threshold, it is more sensitive to changes in dynamic range than CP3, which leads to more false negative detections (incomplete foreground). The performance of Sheikh's KDE and GMM methods show the weakness that they are sensitive to the rapid changes of illumination while updating.

Figure 2.11 shows the qualitative results of the four methods under high contrast, low contrast and rapid changes of illumination, respectively. It is stressed that no morphological operators like erosion/dilation were used in the presentation of these results. The above results indicate that, our proposed method has a better illumination invariance compared with the state of the art methods, even under sudden illumination changes and a low contrast background.

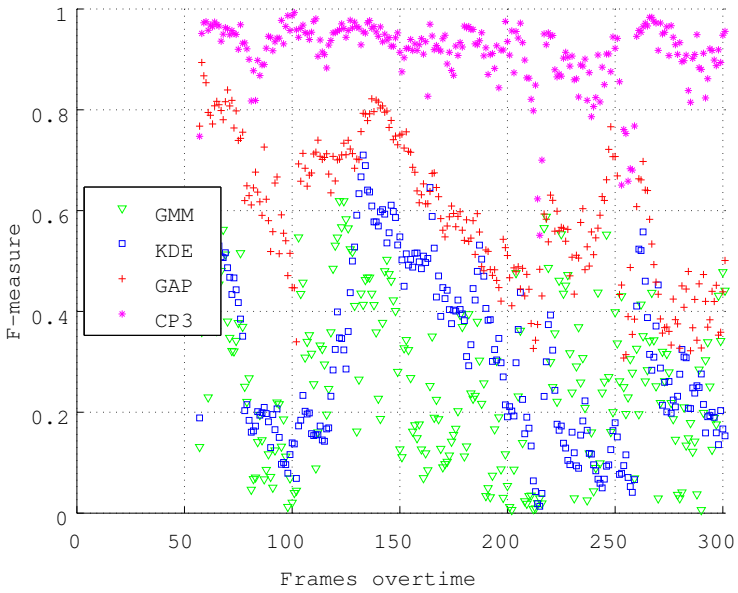


(a) *Precision*

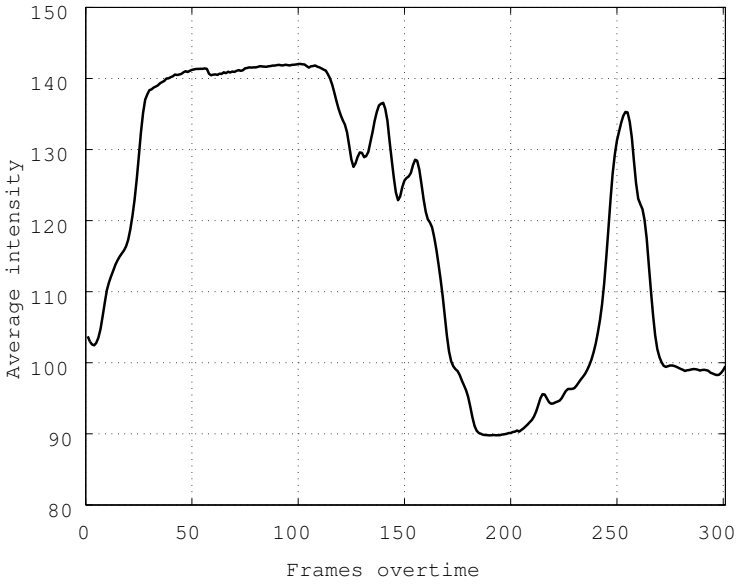


(b) *Recall*

Figure 2.12: *Precision* and *Recall* comparison of GMM, Sheikh's KDE, GAP, and proposed CP3 using PETS2001-dataset3-camera1 dataset with outdoor severe illumination fluctuation over time through 301 testing frames.



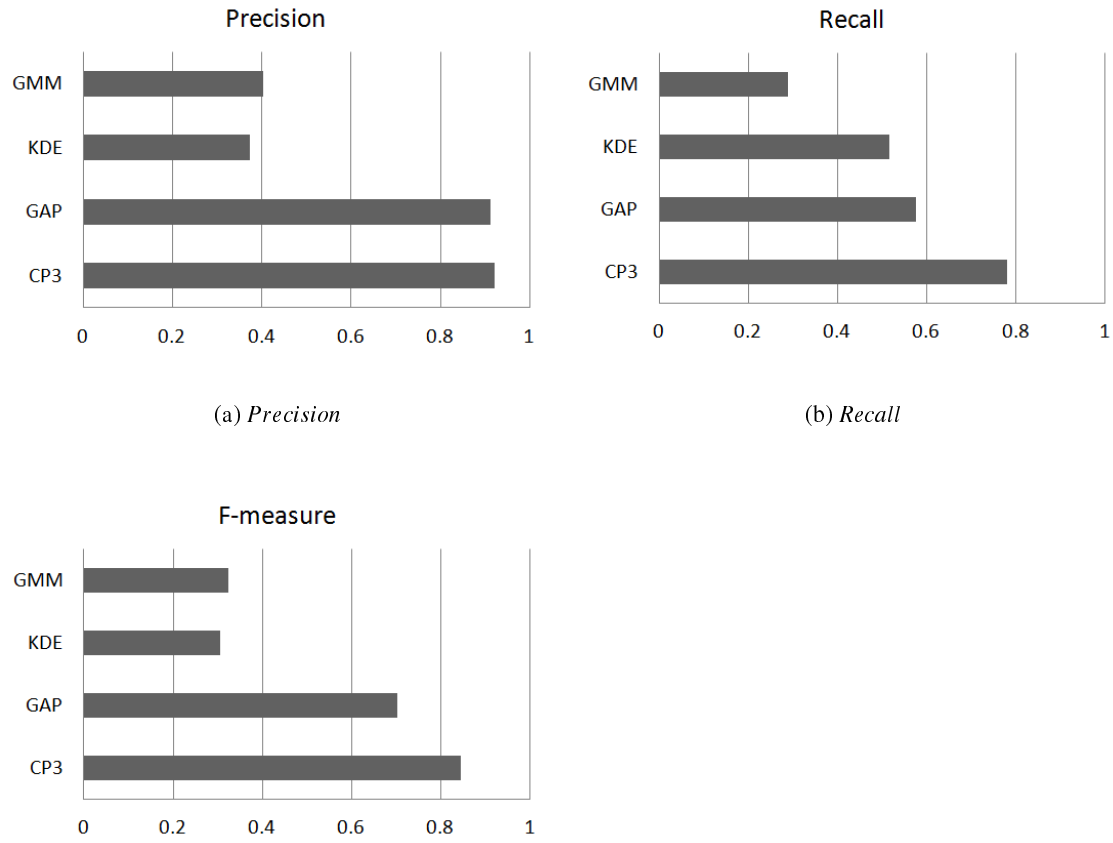
(a) F-measure



(b) Average intensities over time

Figure 2.13: F-measure comparison of GMM, Sheikh's KDE, GAP, and proposed CP3 using PETS2001-dataset3-camera1 dataset with outdoor severe illumination fluctuation, and the average intensities over time through 301 testing frames.

The next dataset for testing indoor environments is the AIST-INDOOR dataset[8]. It contains several indoor extreme conditions: low contrast illumination, low texture, turning on lights and an auto-door rapidly opening and shutting. The dataset AIST-INDOOR includes total 300 frames for training and 4890 frames for detection. The average *Precision*, *Recall* and F-measure of the four methods are shown in Figure 2.15. Figure 2.14 shows four demo frames of detection with CP3, GAP, Sheikh's KDE and GMM respectively. Compared with other approaches, CP3 is not only insensitive to varying illumination but also robust to reciprocating motion of the auto-door. The average *Precision*, *Recall* and $F-measure$ of the above three experiments are shown in Table 2.1.



(a) *Precision*

(b) *Recall*

(c) F-measure

Figure 2.15: The mean metrics of GMM, KDE, GAP and proposed CP3 using AIST-INDOOR dataset.

Here, we would like to emphasize the differences between our method and Sheikh's KDE method from reference [14] because both of them belong to the spatial-dependence model. The most important difference is the difference in their modelling mechanisms. Sheikh's KDE

[8] `http://ssc-lab.com/~liang/CP3_project/AIST_INDOOR_DATASET.rar` (The domain name `ssc-lab.com` would be changed to `www.hce.ist.hokudai.ac.jp`)

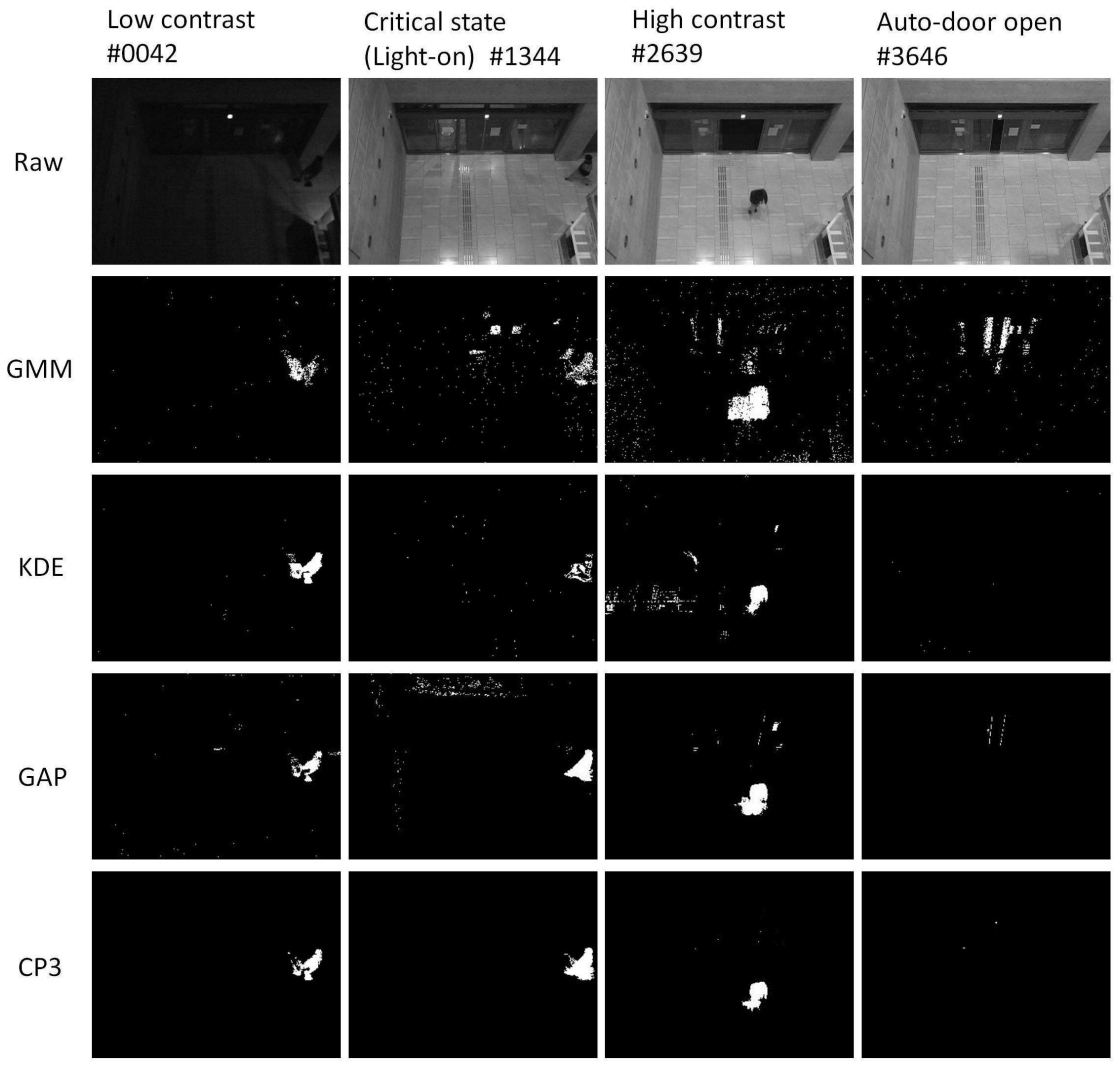


Figure 2.14: Qualitative comparison of GMM, Sheikh's KDE, GAP, and proposed CP3 using AIST-INDOOR dataset. It contains several indoor extreme conditions: low contrast illumination, lights sudden on-off and an auto-door rapid open-shut.

Table 2.1: Mean *precision*, *recall*, and $F-measure$ of GMM, Sheikh's KDE, GAP, and proposed CP3 method using Heavy fog, PATS2001, and AIST-INDOOR datasets.

Methods	Quantitative evaluation	Heavy fog	PATS2001	AIST-INDOOR	Total
GMM	*Precision*	0.614	0.816	0.402	0.611
	Recall	0.747	0.311	0.290	0.422
	$F-measure$	0.674	0.450	0.323	0.482
Sheikh's KDE	*Precision*	0.439	0.390	0.374	0.401
	Recall	0.763	0.464	0.517	0.327
	$F-measure$	0.557	0.424	0.306	0.429
GAP	*Precision*	0.847	0.905	0.912	0.888
	Recall	0.605	0.539	0.575	0.573
	$F-measure$	0.706	0.676	0.703	0.695
Proposed CP3	*Precision*	0.862	0.918	0.922	0.901
	Recall	0.795	0.836	0.780	0.804
	$F-measure$	0.827	0.875	0.845	0.849

method converts each pixel's intensity and location into a feature vector, and then models a joint probability distribution of all the pixels on a feature space. The method is based on the assumption that the pixels' correlation degree depends on the spatial distance between them, (i.e. each observed pixel has higher correlation with its neighbouring pixels), but ignores the localized discrimination between pixels. Such a mechanism can deal well with global illumination changes, but is not robust to local illumination change, which can be clearly observed in the qualitative experimental results in Figure 2.11 and Figure 2.14. Compared with Sheikh's method, our proposed background model avoids any prior assumption of pixel's local correlation, but selects a group of supporting pixels which can maintain a stable statistical relationship with their target pixel. Such a mechanism is robust under both global and local illumination changes.

The next dataset is Wallflower, which is introduced in the work of Toyama et al. [22]. This dataset consists of seven video sequences, each of which addresses a special canonical background subtraction problem. We trained each video sequence using the frames specified by the instruction files in the Wallflower dataset. Our results are shown in Figure 2.16. Compared with the results in [22], our method deals with the illumination changes and background fluctuations well. Note that, some incomplete parts of the foreground exist mainly because the dark colours of both foreground and background come into similar intensities after covering them into gray-scale image. When using a multi-channel colour detection, the performance would be improved.

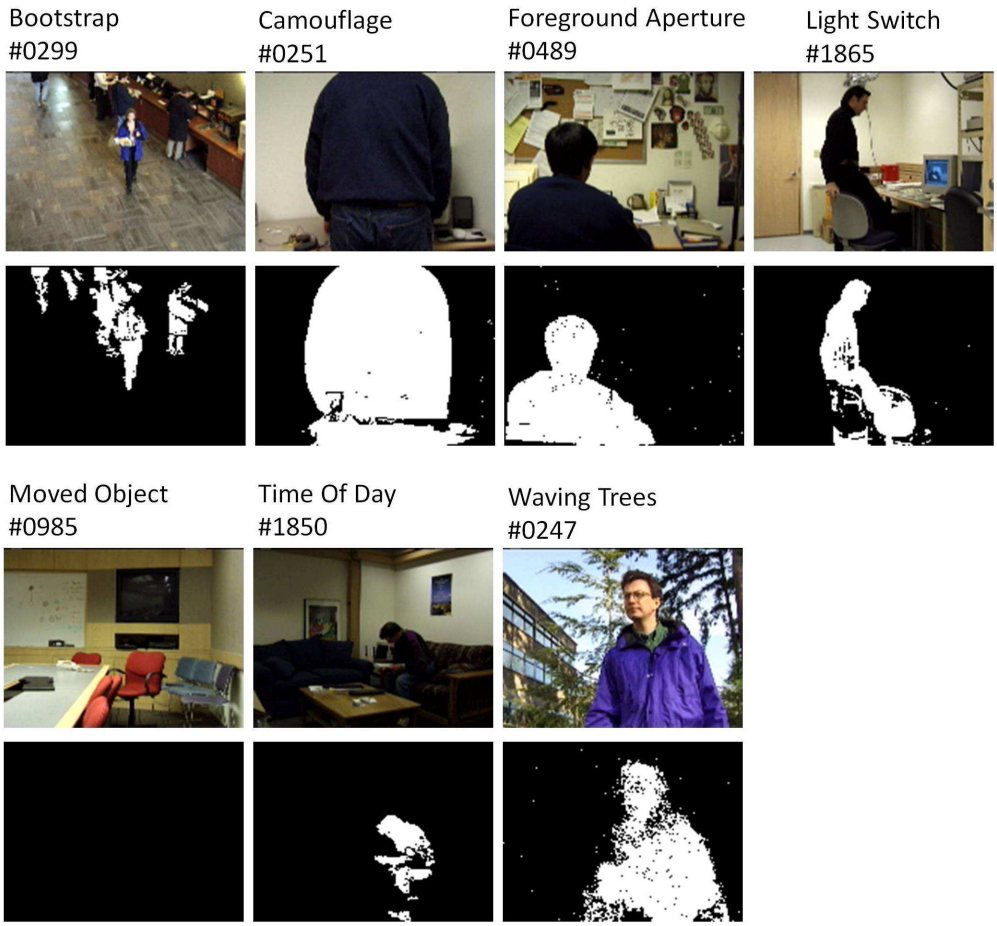
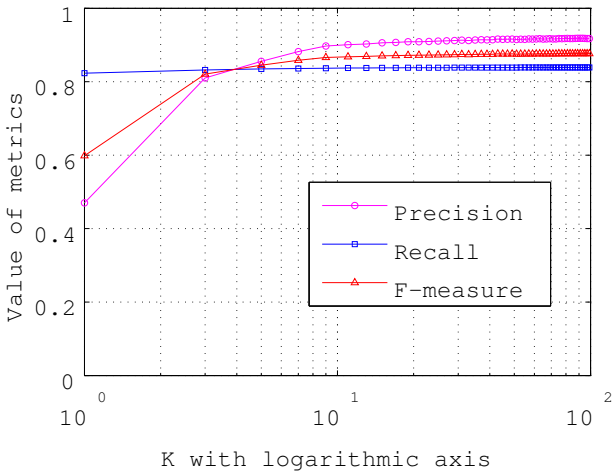


Figure 2.16: CP3 results of Wallflower dataset.

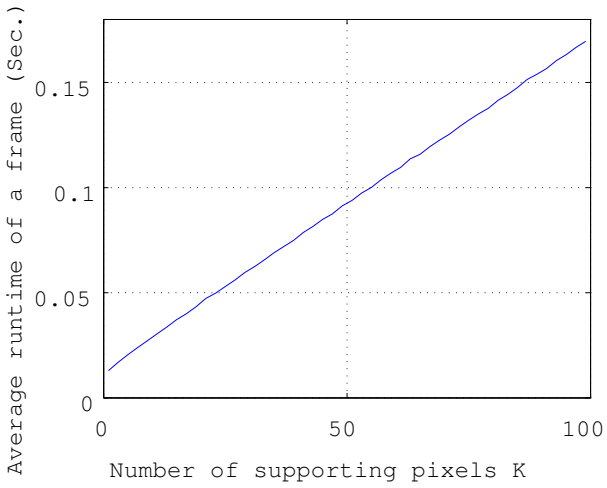
2.4.2 Different number of supporting pixels

To examine the detection performance of the proposed method using a different number of supporting pixels, we use PETS2001-dataset3 camera1 and calculate the mean value of *Precision*, *Recall* and F-measure shown in Figure 2.17 (a). The corresponding average runtime to process each frame is shown in Figure 2.17 (b). The runtime is measured on a computer with a Intel Xeon 3.0 GHz processor with a C language implement. From Figure 2.17 (a), we can observe that K mainly affects *Precision*: when $K = 1$, *Precision* is around 0.5. According to the definition of *Precision*, this result means the number of false positive pixels (FP) and true positive pixels (TP) are similar which indicates that one supporting pixel Q_1^P cannot provide robust detection. When K becomes larger, *Precision* increases dramatically; and when K continues to grow (larger than 9), *Precision* tends to be stable, which means there is a small quantity of false positive pixels (FP). Figure 2.17 (a) also shows that *Recall* changes little under different K, which indicates that the completeness of the object could be preserved even K is small. The results of the above quantitative analysis can also be observed from the detection sample in Figure 2.18. In addition, as shown in Figure 2.17 (b), the runtime of detection linearly

increase with K. In conclusion, considering both the detection performance and computation cost discussed above, we set the number of supporting pixels K between 10 to 20 in practice.



(a) *Precision*, *Recall* and F-measure under different K (To show the variation tendency clearly when K is small, the horizontal axis is logarithmic.).



(b) Average detection runtime of a frame under different K.

Figure 2.17: Work with a different number of supporting pixels K using PETS2001-dataset3 camera1. From (a), when $K = 1$, *Precision* is around 0.5; when K becomes larger, *Precision* increases dramatically; and when K continues to grow, *Precision* tends to be stable; *Recall* changes little under different K. From (b), the runtime of detection linearly increase with K.

(a) Frame #2706.

(b) $K = 1$

(c) $K = 5$

(d) $K = 15$

(e) $K = 99$

Figure 2.18: Detection sample with a different number of supporting pixels K using PETS2001-dataset3 camera1. When K becomes larger, smaller quantity of false positive pixels (FP) exists.

2.4.3 Analysis of accelerated background modelling

We also use the accelerated version to carry out background modelling with a sample interval from $\Lambda = 2$ to $\Lambda = 5$, and compare its time consumption of background modelling and object detection performance with Algorithm 3 shown in Table 3.3. The dataset is a high resolution (1024×1024) surveillance video in a supermarket with 101 training samples, some detection samples are shown in Figure 2.20. The runtime are measured on a computer with a Intel Xeon 3.0 GHz processor.

From Table 3.3, we can observe that compared with the original CP3, the time cost for background modelling of the accelerated algorithm dramatically reduces even when only using a limited sampling interval. For instance, when $\Lambda = 5$, the runtime reduced by 91.96%, while the *Precision*, *Recall* and F-measure only reduced by 5.85%, 4.56%, and 5.12%, respectively. The above three metrics can keep high levels mainly because when the sampling interval is relative small, there are still sufficient qualified supporting pixels to maintain robust detection.

Larger-range quantitative tests are shown in Figure 2.19. We can observe that when the sampling interval Λ continues to increase, *Precision*, *Recall* and F-measure clearly decrease, and finally tend to be stable at low levels.

In summary, when using the accelerated algorithm of CP3, it is suggested to firstly conduct a preliminary work using a relative small sampling interval, which would efficiently reduce the time cost of background modelling and maintain a fairly good detection performance. In addition, an optimized implementation of Algorithm 4 for object detection can process about 17 fps using a frame size of 1024×1024.

Table 2.2: The comparison of CP3 and its accelerated version.

Interval	Runtime (Sec.)	Precision	Recall	F-measure
–	151.57	0.923	0.877	0.899
$\Lambda = 2$	47.02	0.920	0.857	0.887
$\Lambda = 3$	31.51	0.903	0.851	0.876
$\Lambda = 4$	15.73	0.887	0.844	0.865
$\Lambda = 5$	12.19	0.869	0.837	0.853

2.5 Summary

In conclusion, CP3 builds a novel background model for object detection based on co-occurrence pixel pairs. The model performs robust detection under outdoor and indoor extreme environments.

The key contributions of the algorithm are as follows:

- Compared with independent pixel-wise methods, CP3 determines stable co-occurrence pixel pairs, instead of building the parameterized/non-parametrized model for a single

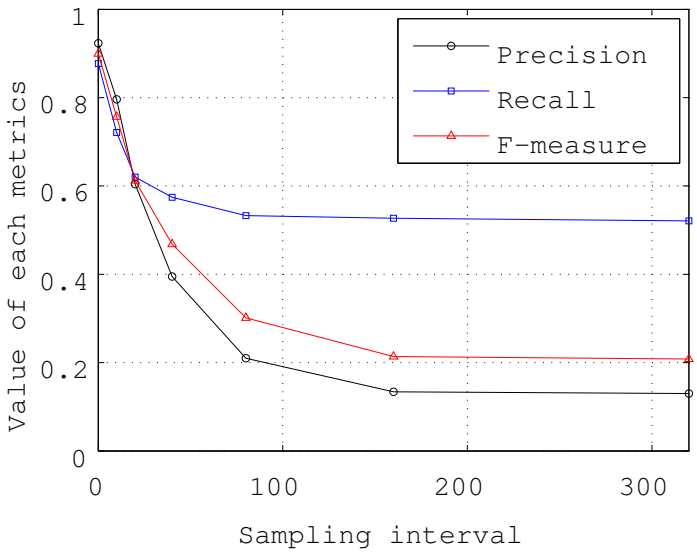


Figure 2.19: *Precision*, *Recall* and F-measure using the accelerated algorithm of CP3 model with the sampling intervals from 0 to 320.

pixel. These pixel pairs maintain a reliable background model, which can be used to capture structural background motion and cope with local and global illumination changes.

- As a spatial-dependence method, CP3 does not predefine any local operator, subspace or block, and it provides an accurate detection criterion even though the gray-scale dynamic range is compressed under weak illumination.

We interpreted our method and conducted experiments using gray-scale data, however, using colour imagery is also an option for object detection. The modification of CP3 for multi-channel use, such as *rgb*, can be realized through vectorized extension of the basic unit for a three-dimensional observation; for the Gaussian model of each pixel pair, the associated mean value and variance evolve to a 3D vector and a 3×3 covariance matrix respectively.

Appendix

K-means clustering for optimizing spatial distribution:

Define a spatial distance between P and Q_n by $d(P, Q_n)$, and define a spatial distance between different Q_n by $d(Q_n, Q_{n'})$, where $n = 1, 2, ..., N$ and $n \neq n'$. There are different approaches for checking the distance between two observed values, such as Euclidean distance, Hamming distance or city-block distance. As a spatial distance measurement, we used Euclidean distance. A set to represent the local spatial distribution of P and Q_n is $\psi(Q_n) = \{d(P, Q_n), d(Q_n, Q_{n'})\}$.
(I) Initialize:

Given random initial clustering centers $Q_k^p(0) \in \{Q_n\}, k = 1, 2, ..., K$, and one fixed clustering center of P, the total $K + 1$ number of initial seed locations, the corresponding initial cluster

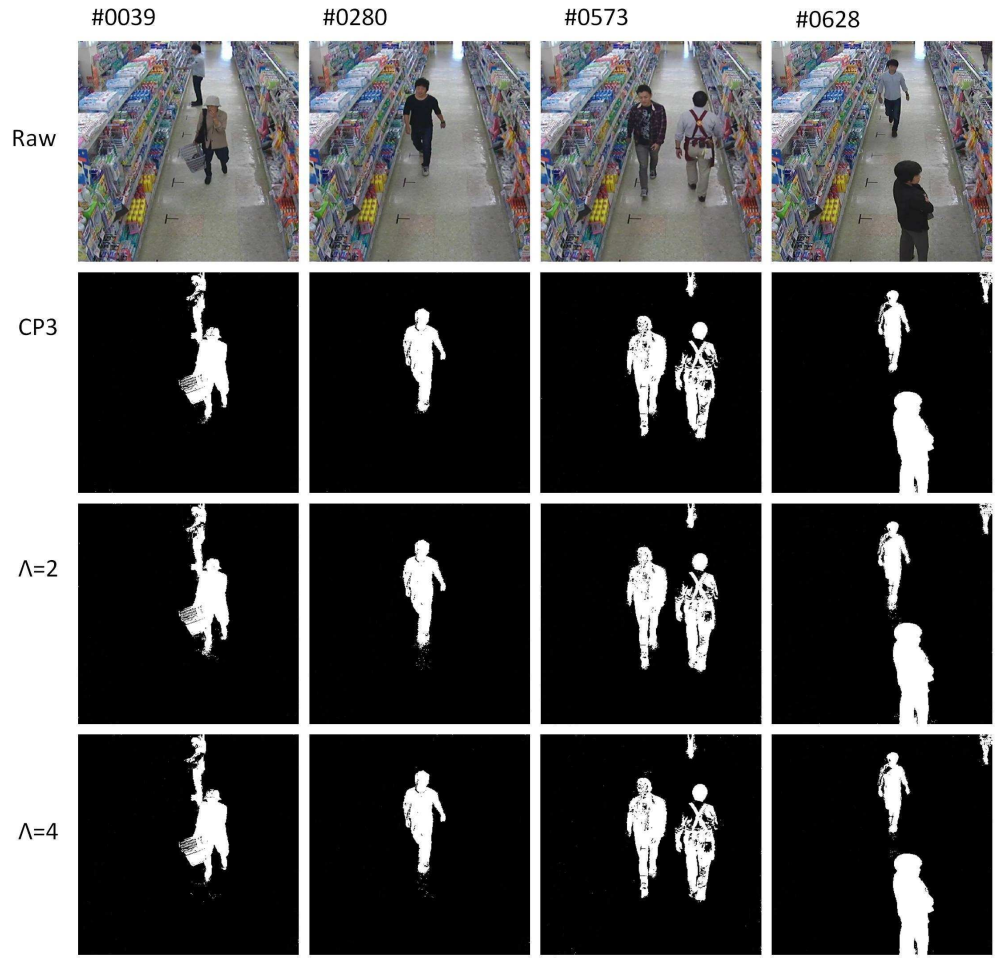


Figure 2.20: Detection using CP3 and its accelerated algorithm.

$\{\psi(Q_k^P)_0\} = \Phi$.

(II) Clustering:

Assign each $Q_n(u', v')$ to a cluster with the closest seed location:

$$\{Q_k^P\}_\eta = \{Q_n : d(Q_n, Q_k^P[\eta]) \leq d(Q_n, Q_k^P[\eta])\},$$

where η is iteration times, and $\forall\, k \neq k'$, until Q_n goes into exactly one $\{Q_k^P\}_\eta$.

(III) Updating centroid:

The updated centroid of the cluster is calculated by the mean locations:

$$Q_k^P[\eta + 1] = \#^{-1} \sum_{(u',v')} \{Q_k^P\}_\eta,$$

where # is the number of $Q_n \in \{Q_k^P\}_\eta$. It is important to note that the location of P without updating in step (III) also allows $Q_n \in \{P\}$ to calculate $d(P,\ Q_n)$ in step (II).

Chapter 3. CP3 for image matching

3.1 Overview

In this chapter, we introduce a novel framework of robust image matching based on Co-occurrence Probability based Pixel Pairs (CP3) [54]. CP3 was originally a statistical background model, proposed for pixel-level object detection under dynamic backgrounds, such as sudden illumination fluctuation and physical motion background. It has been proved to be very robust and accurate for the object detection task. Theoretically speaking, both the object detection and image matching can be seen as a model matching problem. The differences is that the object detection is to seek the regions of interest (ROI) which violate/mismatch the background model, while the image matching is to seek the ROI which can match the image model optimally. Therefore, we further extend the use of CP3 to the robust image matching task.

CP3 employs spatial pixel pairs to represent each other by using their intensity differences which is much more stable than the intensity of a single pixel, especially when the intensity of a single pixel changes dramatically over time. The overview of the proposed image matching framework is shown in Figure 3.1. Different from the traditional image matching method, CP3 firstly carries out a spatial-temporal learning stage, to build the intensity difference of each specific pixel pair as a single Gaussian model, which is effective to cope with complex scenes. Although an additional learning stage is necessary, the experiment results show that the proposed method is robust under several imaging cases and it also outperforms other two well-known robust similarity measure Normalized Cross Correlation (NCC) and Statistical Reach Feature (SRF).

It should be noted that, in this study, the proposed method aims to deal with the blur, partial illumination, highlight and occlusion imaging conditions in natural and complex environment, but it is not designed for dealing with the rotation, scale change, and affine change problems so far. In addition, since this method relies on a off-line supervised learning phase, it cannot be directly used for an on-line unsupervised learning based robot vision system, so far.

The remainder of this chapter is organized as follows. In the next section, SRF method is discussed. Section 3.3 details the learning phase. Section 3.4 presents the similarity measure procedure. Section 3.5 presents the experimental results, and Section 3.6 is the summary of this chapter.

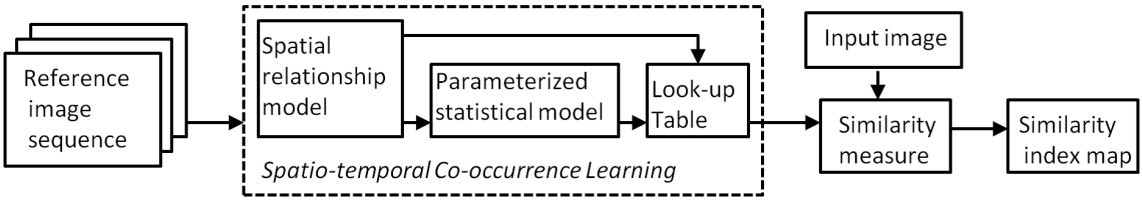


Figure 3.1: Overview for the proposed image matching framework.

3.2 SRF for image matching

As we mentioned in Section 1.2.4, SRF is robust under illumination fluctuation and it is also computational efficient. However, the sign coding of SRF is produced by a global threshold T_P, which would not be accurate enough to describe the intensity relationship of a specific pixel pair. In detail, T_P is a trade-off parameter, small T_P cannot maintain a reliable relationship between a pixel pair because the image noise always exists. On the other hand, large T_P would cause a large number of ineffective pixel pair, i.e. the sign code $srf(P \rightarrow Q; T_P) = \varnothing$. This will make SRF lose some details of the image, leading to an insensitive matching. This problem cannot be ignored when the image is a degenerated by defocusing or low illumination. As shown in Figure 3.4 to Figure 3.12 in the experiment sections, SRF resulted false matching.

Actually, the intensity difference of a pixel pair can be changed along with the illumination change, e.g. the intensity difference will be smaller under a low illumination case than it under a normal illumination, and vice versa. In this study, the proposed method based on CP3 also relies on the intensity relationship between pixel pairs. Different from SRF method, CP3 builds the intensity difference as a single Gaussian model for each specific pixel pair via a spatial-temporal learning stage, rather than a global threshold T_P in SRF, which is a more accurate description of a pixel pair. The proposed image matching framework also removes the definition of $\varnothing$ in SRF, allowing the pixel pair with similar intensity also build their Gaussian model, which maintain a constant number of supporting pixels Q for each target pixel P. On the basis of the illumination invariance property of CP3, we also propose a similarity measure function to achieve robust image matching.

3.3 Spatial-temporal learning

3.3.1 Spatial relationship model

In this section, we firstly define a spatial structure of the *supporting pixels* for each target pixel P, as shown in Figure 3.2. This structure is identical with the one of SRF that the supporting pixels are selected from the eight neighbourhood pixels, denoted as Q_k, where $k = 1, 2, ..., 8$. This spatial structure has some advantage: (1) it is very efficient for the following statistical calculation. For all target pixels, the 3-D lattice Γ can be operated by a matrix shift operation only once for one supporting pixel. For instance, the intensity difference between P and Q_1

can be calculate between $\Gamma(u, v, t)$ and $\Gamma(u+1, v+1, t)$. (2) the neighbouring pixels generally have much more stable intensity relationship than the pixels far away from each other, especially when the scene involves fluctuation.

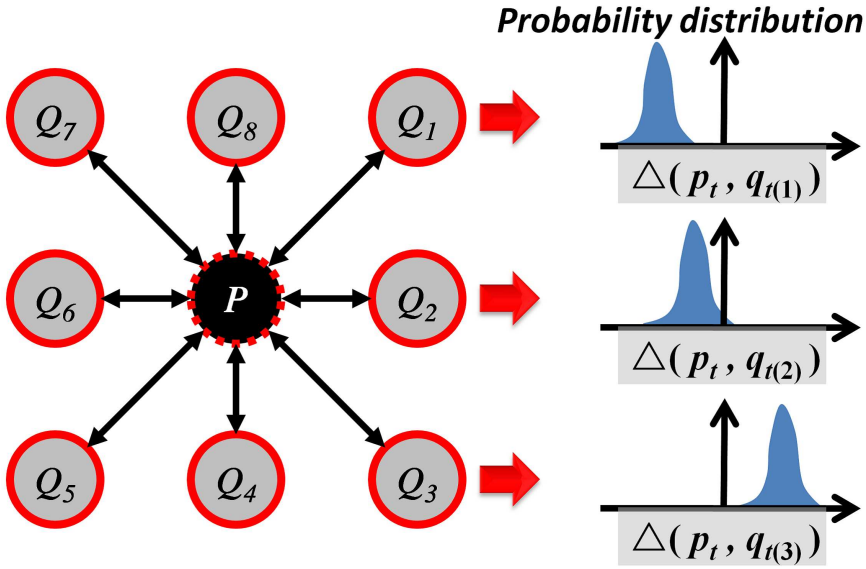


Figure 3.2: Spatial relationship model of a target pixel and its probability distribution.

3.3.2 Parametrized statistical model of a pixel pair

For each supporting pixel Q_k, we assume that the intensity of it keeps a bivariate difference b with the intensity of the target pixel P, that is $p_t \sim \mathcal{N}(q_{t(k)}+b, \sigma_\varepsilon^2)$, where $\mathcal{N}$ refers to a normalized distribution. We define $\Delta(p_t, q_{t(k)})$ as the difference of P and Q_k, then we have the Single Gaussian probability density function (SG-PDF) as follows,

$$f(\Delta(p_t, q_{t(k)}); \hat{b}, \hat{\sigma}_\varepsilon) = \frac{1}{\hat{\sigma}_\varepsilon \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{\Delta(p_t, q_{t(k)}) - \hat{b}}{\hat{\sigma}_\varepsilon}\right)^2\right), \tag{3.1}$$

where the estimation of difference is $\hat{b} = \mathcal{E}[p_t - q_{t(k)}]$ and the estimation of noise standard deviation $\hat{\sigma}_\varepsilon = \sigma_{p_t - q_{t(k)}}$. We use maximum likelihood method to estimate the two parameters,

$$\hat{b} = \mathcal{E}[p_t - q_{t(k)}] = \frac{1}{T}\sum_{t=1}^{T}(p_t - q_{t(k)}) \tag{3.2}$$

$$\hat{\sigma}_\varepsilon^2 = \sigma_{p_t - q_{t(k)}} = \frac{1}{T}\sum_{t=1}^{T}((p_t - q_{t(k)}) - \hat{b})^2 \tag{3.3}$$

The above two parameters $\hat{\sigma}_\varepsilon, \hat{b}$ are recorded in a look-up table (LUT) consisting of $\{Q_k\} \sim [\hat{\sigma}_\varepsilon, \hat{b}\,]$ for the following image matching procedure. So far, we have modelled the intensity difference of each pixel pair $\Delta(p_t, q_{t(k)})$ to be an unique statistical model SG-PDF $f(\Delta(p_t, q_{t(k)}); \hat{b}, \hat{\sigma}_\varepsilon)$

Algorithm 3: CP3 for spatial-temporal learning

Data: $B = \{I_1, I_2, ..., I_T\}$ with total T images.

Result: LUT of $\{Q_k\}_{k=1,2,...,8}$ consisting of $[\hat{\sigma}_\varepsilon, \hat{b}]$.

for *each* $P(u,v)$ $1 \le u \le U, 1 \le v \le V$ **do**

  for *each* Q_k **do**

    (1) Compute $\hat{b} = \frac{1}{T}\sum_{t=1}^{T}(p_t - q_{t(k)})$

    (2) Compute $\hat{\sigma}_\varepsilon^2 = \frac{1}{T}\sum_{t=1}^{T}((p_t - q_{t(k)}) - \hat{b})^2$

    (3) Record $[\hat{\sigma}_\varepsilon, \hat{b}]$ into LUT of $\{Q_k\}$.

with only two parameters $\hat{\sigma}_\varepsilon$ and $\hat{b}$. It has some characters compared with SRF method: (1) By learning a reference image sequence, $f(\Delta(p_t, q_{t(k)}); \hat{b}, \hat{\sigma}_\varepsilon)$ can tolerate the noise of the pixel pair and it is also adaptive to the intensity change of the pixel pair. (2) Without assuming any value of the two parameters, we estimate the values of the two parameters of each pixel pair respectively according to the real reference image sequence. It is obvious that this manner is much more accurate than setting a global value to describe the relation of a pixel pair.

In this work, we use the difference value of the pixel pair to build a single Gaussian distribution, this way is different from a traditional Gaussian distribution function for a single pixel; If we only observe the intensity of a single pixel, the distribution would be changed following the intensity variation, obviously, the single Gaussian method is not a proper model. When we observe the difference value of the pixel pair, most of the cases, the values are close to the mean value of the Gaussian model because the differential relationship is relatively stable than the intensity itself, even the illumination changed; when there are some more difficult situations, such as shades and shadows, most of the pixel pair could still maintain a stable difference because they are neighbouring pixels. Certainly, the difference value of a pixel pair cannot be always closed to the mean value. The calculation of the Gaussian model for each pixel pair can grantee a pair of unique parameters (mean and variance) for each pixel pair, which can tolerate some difficult situations, but maintain an accurate modelling according to the specific imaging condition in the reference images for training.

It should be noted that we use this SG-PDF to model the distribution of a pixel pair rather than a Gaussian mixture model (GMM) [11, 16, 19] because we found that a SG works better here than a GMM description, because most of the neighbouring pixel pairs have stable intensity relationship with each other and the probability distribution of a pixel pair is always with one major peak. In addition, SG is more efficient for computation than GMM approach, because in GMM method, an additional expectation maximization (EM) needs to be calculated to assign parameters to each component of GMM.

3.3.3 Prepare images for learning

Compared with the traditional single reference image based methods, CP3 based on a spatial-temporal learning stage, which need to prepare a certain number of reference images for learn-

ing. In principle, these reference images should better cover the possible changes of the scenario. For instance, we can use the camera to capture a scenario with a fixed time interval when manually providing different kinds of illumination (as the case of Experiment 1). Sometimes, we have to manually select a set of reference images to organize more complex situations (as the case of Experiment 2). In this case, the acquisition of the reference images is a manual process, which inevitably have some translation of position among the reference images. Therefore it requires a pre-processing to align the reference images for the subsequent spatial-temporal learning. When preparing the reference images for train at different locations, because we can roughly control the region we captured manually, there are just slight translation of the position among the different reference images, in which case using NCC is enough to correct the translation error because the correlation peak can be very concentrate. This is far more easier than using NCC to search a small target in a large and complex field of view, in which the similar structure and the local illumination would destroy an explicit correlation peak.

Obviously, the proposed method is a "Best Effort" method, we have to manually select a set of reference images according to the difficulty of specific application. Generally, to realize robust matching, it is necessary to collect a large number of reference images in order to randomly cover much more difficult cases. In practice, we suggest the user to use a way of incremental learning, that is firstly conduct a preliminary work using a relatively small number of samples, then if necessary, add more samples according to the performance evaluation.

3.4 Similarity measure

The similarity measure includes two stages: (1) identify the normal/abnormal state of each pixel pair (P, Q_k); (2) calculate a normalized similarity coefficient at each position of the input image of a scene, and finally produce a similarity index map.

3.4.1 Identifying the state of the pixel pair

The proposed CP3 model converts the image matching problem into a competitive binary classification problem by comparing the intensity pairs $(P, \{Q_k\}_{k=1,2,...,8})$ in turn. For each pixel pair (P, Q_k), the binary function $\beta(Q_k)$ used to discriminate the normal/abnormal state can be estimated as the following condition according to Eq. (3.1):

$$\beta(\{Q_k\}_{k=1,2,...,8}) = \begin{cases} 1 & \text{if } |(p - q_k) - \hat{b})| < C \cdot \hat{\sigma}_\varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

where p and q_k are the intensity value of P and Q_k, respectively, in the testing image J_{xy}. The testing image J_{xy} is essentially a sliding window with the same size of the reference image I_t, which scans the range of an input image S of a scene, where $(x, y) \in S$. Note that using the bivariate normal distribution of the pixel pair is different from a traditional SG-PDF based identification function; In a single Gaussian method, the ideal threshold should be changed

following the latest intensity variation. For example, the standard deviation should be larger when the fluctuations in illumination become more intense. In our proposed version, the stable difference of a pixel pair provides an adaptive observation so that $\hat{\sigma}_\varepsilon$ is only related to the noise acting on each pixel pair. Therefore, it is unnecessary to update an adjustable $\hat{\sigma}_\varepsilon$ to adapt to its changes caused by illumination changes. The constant C can be set from 1.0 to 3.0 to contain an area of 68-99% of its PDF. In the following experiment section we set $C = 3.0$. In addition, the records of $\hat{\sigma}_\varepsilon$ and $\hat{b}$ from the learning stage provide an accurate parametrized criterion to produce an explicit correlation peak, which will be confirmed in the experimental section.

3.4.2 Similarity coefficient

After identifying the normal/abnormal state of the pixel pairs, 8 bits of $\beta(Q_k)$ are produced for the following decisions of each P. In order to create the similarity index map, the normalized similarity coefficient $\xi(I, J(x, y))$ is defined as,

$$\xi(I, J_{xy})_{x,y\in S} = \frac{1}{8UV} \sum_{u=x+1}^{x+U} \sum_{v=y+1}^{y+V} \sum_{k=1}^{8} \beta(Q_k). \tag{3.5}$$

we can see that the similarity coefficient $\xi(I, J(x, y)) \in [0, 1]$ is a normalized summation to count the number of the pixel pair which can match the parametrized statistical model. This similarity degree of the regions in the scene and the parametrized statistical model is free from the effect of noise and illumination change. For the implementation of similarity measure, the procedure used to calculate $\xi(I, J(x, y))$ for every position (x, y) is performed by an LUT operation for calculating $\beta(Q_k)$ and a bit counting operation for calculating $\xi(I, J(x, y))$, both of which are quite efficient to implement on both conventional and embedded hardware. The pseudo-code of the similarity measure of CP3 is shown in Algorithm 4. By using the similarity coefficient $\xi(I, J_{xy})_{x,y\in S}$ to scan the whole scene, it finally produce a similarity index map containing the similarity coefficient value of each pixel perdition. The true matching position $(X, Y)_{match}$ can be got at the maximum value,

$$(X, Y)_{match} = argmax(\xi(I, J_{xy})_{x,y\in S}) \tag{3.6}$$

3.5 Experimental results

To evaluate the performance of the proposed method, we tested it on data including challenging environments for both qualitative and quantitative analysis.

For quantitative analysis, we utilize the output signal-to-noise ratio (SNR), as did in [49], which can evaluate the sharpness of the correlation peak in the similarity index map, as follows

$$SNR = \frac{|\xi_{truth} - M|}{D}, \tag{3.7}$$

Algorithm 4: Similarity measure of CP3

Data: Testing sliding window J_{xy} from an input image S of a scene.
Result: Similarity index map with values of $\xi(I, J_{xy})$.
for *Each sliding position* (x, y) **do**
  for *Each pixel P in J* **do**
    (I) Initialize:
    Load look-up table $\sim [\hat{\sigma}_\varepsilon, \hat{b}]$ of $\{Q_k\}_{k=1,2,...,8}$.
    (II) Pixel pair identification:
    for $k = 1, 2, ..., 8$ **do**
      if $|(p - q_k) - \hat{b})| < C \cdot \hat{\sigma}_\varepsilon$ **then**
        $\beta(Q_k) = 1$
      else
        $\beta(Q_k) = 0$
  (III) Similarity coefficient:
  Compute $\xi(I, J_{xy})_{x,y \in S} = \frac{1}{8UV} \sum_{u=x+1}^{x+U} \sum_{v=y+1}^{y+V} \sum_{k=1}^{8} \beta(Q_k)$.

where ξ_{truth} is the similarity coefficient value of the similarity index map corresponding to the true position, and M and D are the mean and standard deviation of $\xi(I, J_{xy})_{x,y \in S}$ respectively, in the region excluding the maximum. When the incorrect matching occur, we also define $\Delta(\xi_{fault}, \xi_{truth}) = |\xi_{fault} - \xi_{truth}|$ to measure the departure degree of the true similarity coefficient, where ξ_{fault} is the maximum similarity coefficient of the similarity index map corresponding to the false position.

We compared our algorithm with two methods: (1) NCC [38] method, a standardized method among single reference image based methods, and (2) SRF [46] because it is a predecessor and has a homologous methodology with CP3. For SRF, we set $T_p = 15$ which is suggested in [46].

3.5.1 Experiment 1

Figure 3.3 shows the experimental images, the input image of a scene is a circuit board captured under various imaging conditions, and a tiny chip is the target, which has similar appearance to some other parts of the circuit board. As depicted in the right hand of Figure 3.3, we used 67 pieces of reference images which keep gradual illumination changes as the reference image sequence for CP3 learning, and we use the average image of the sequence as the single reference image for NCC and SRF. Subsequently, we evaluate the performance of the three methods under normal illumination, defocusing, low-light with noise, and occlusion, respectively. As shown in Figure 3.4 to Figure 3.7, the similarity index maps of CP3 have sharp correlation peaks with overwhelming magnitudes, which provide clear maxima over the whole similarity index maps. As shown in Figure 3.4 to Figure 3.7, the correlation index maps of SRF are much narrow-ranged than CP3 and they also produce larger range noise than CP3 especially in the low-light and noise case shown in Figure 3.6, because its threshold T_P is sensitive to the

compression of the intensity's dynamic range under low illumination. The similarity index maps of NCC shown in Figure 3.4 to Figure 3.7, have so many disturbed local maxima because of the similar appearance between the tiny chip and other parts of circuit board, and NCC itself only considers a global illumination change but ignore the spatial structure of a local neighbourhood per pixel, leading to less discriminative similarity coefficient values among the similarity index map. Table 3.1 summarized the corresponding quantitative analysis results, where it is clear that even under severe imaging condition, CP3 can capture a correct correlation peak, and the SNR values of CP3 are significantly higher than both the one of SRF and NCC. In addition, we can observe in Figure 3.4 (d) to Figure 3.7 (d) that, there is a "wall" on the right side of the similarity index map of NCC method. That is because the boundary of the board has a line of connecting finger, whose appearance looks like the connecting finger around the boundary of the chip. This phenomenon can be also observed in Figure 3.12 (f) in Experiment 2, when the scene has some similar appearance with the target.

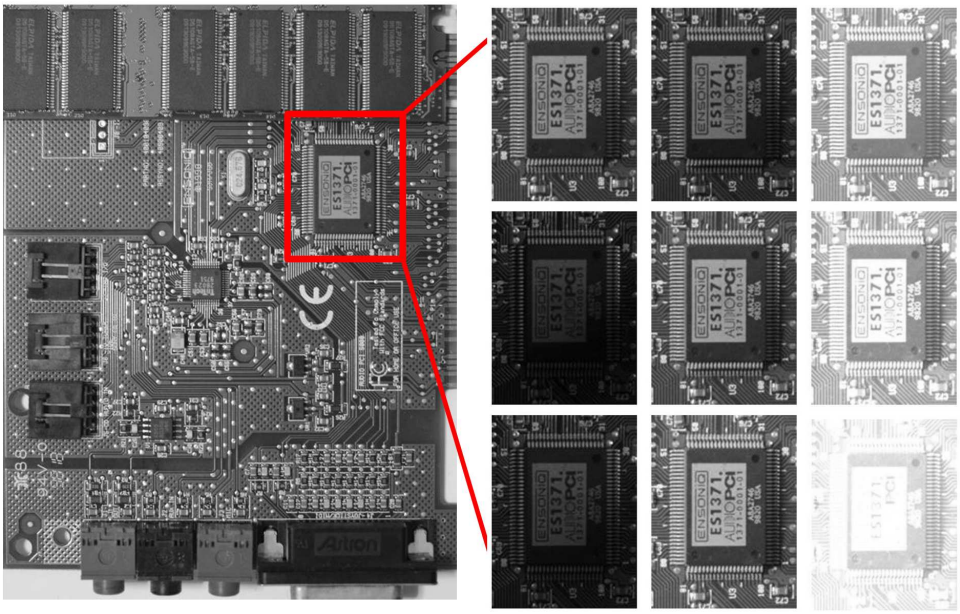
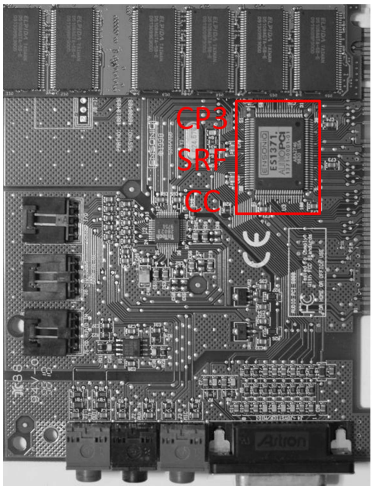
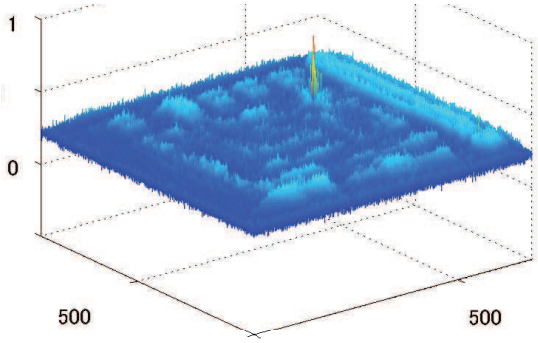


Figure 3.3: Left hand is the input image of the scene for Experimental 1 (600×800). Right hand is a subset of the reference image sequence for CP3 learning (magnified version).

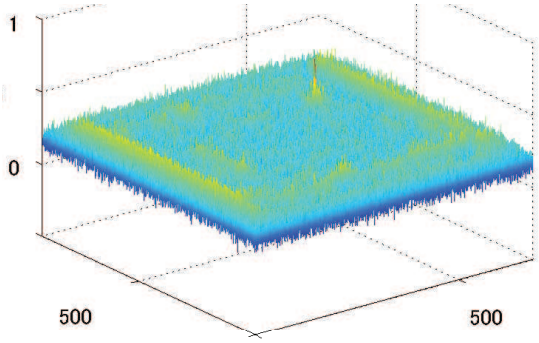
Figure 3.8 shows a demonstration of the two parameters $\hat{b}$ and $\hat{\sigma}_\varepsilon$ of $\Delta(p_t, q_{t(k)})$ at $P(40,23)$ with its supporting pixels. It is clear that each Q have a specific $\hat{\sigma}_\varepsilon$. When the $\hat{b}$ is far from 0, the corresponding $\hat{\sigma}_\varepsilon$ could have a larger value. In addition, we can see that the distribution of Q_1,Q_3 and Q_4 near 0, and Q_1 and Q_3 cover both a plus and a minus range. CP3 can describe this case sufficiently using the two statistical parameters. We also show the SRF threshold T_p and list the SRF sign of each Q. We can see that SRF has to set this sign of Q_1,Q_3 and Q_4 to $\varnothing$, which means they cannot work for the matching. To show the intuitive representation of the distribution of $\Delta(p_t, q_{t(k)})$ in Figure 3.8, we provide the corresponding histogram of the distribution in Figure 3.8 (c).



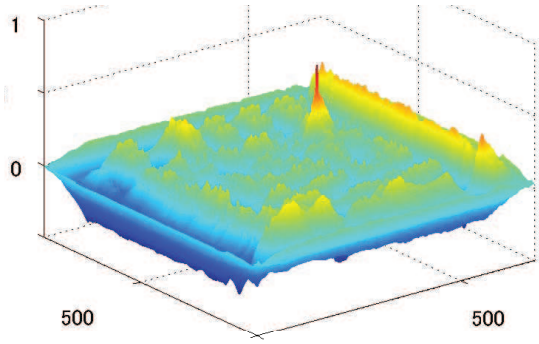
(a) Normal



(b) CP3 (Normal)

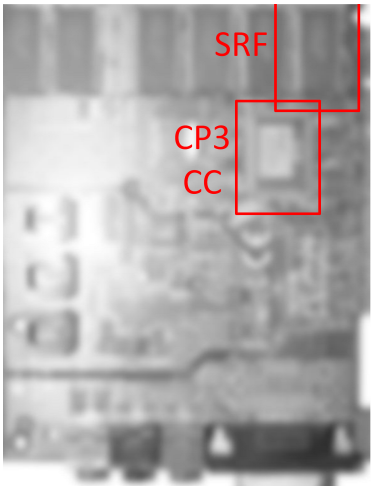


(c) SRF (Normal)

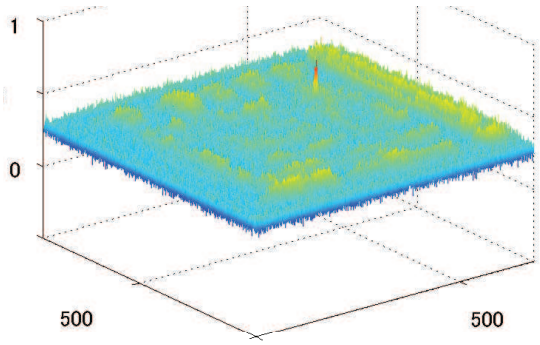


(d) NCC (Normal)

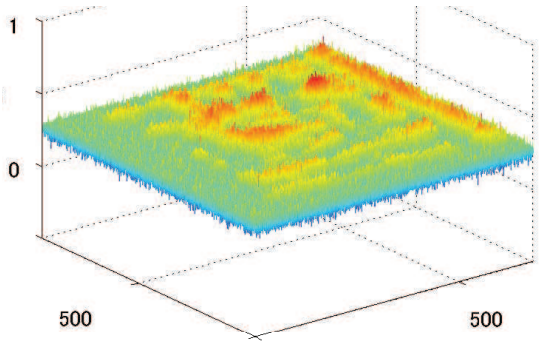
Figure 3.4: Experiment 1 (Normal). Input images of a scene and matching results (600×800), and the corresponding similarity index maps of CP3, SRF and NCC.



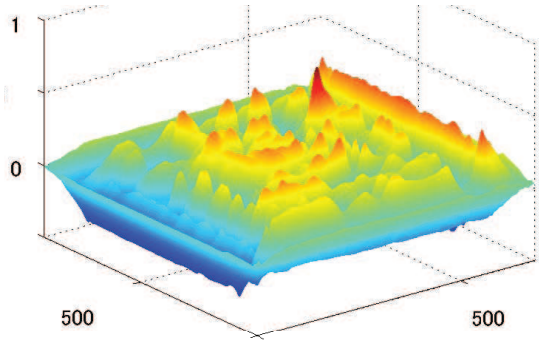
(a) Defocusing



(b) CP3 (Defocusing)

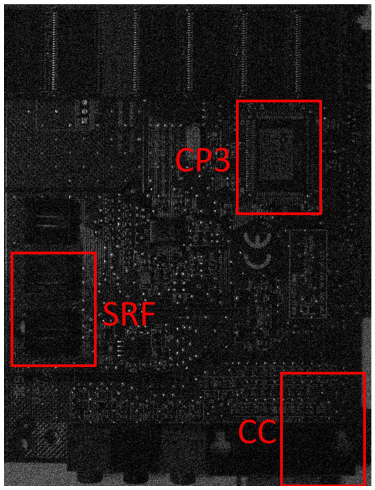


(c) SRF (Defocusing)

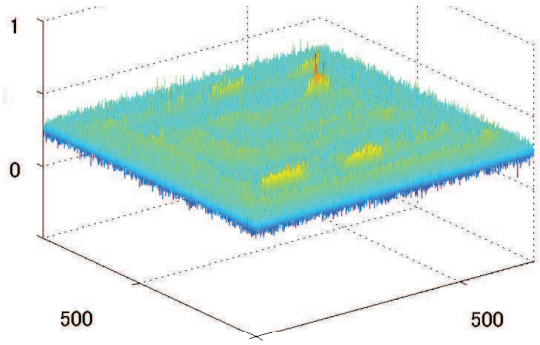


(d) NCC (Defocusing)

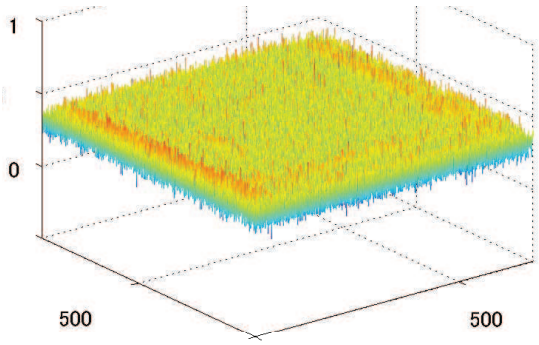
Figure 3.5: Experiment 1 (Defocusing). Input images of a scene and matching results (600 × 800), and the corresponding similarity index maps of CP3, SRF and NCC.



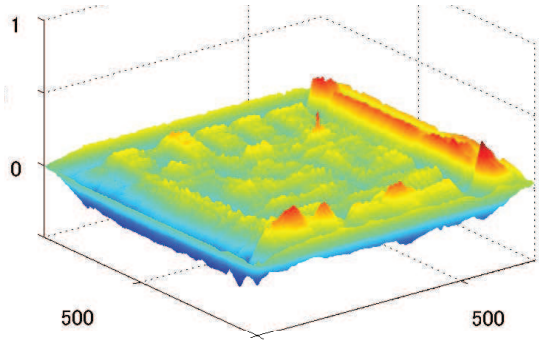
(a) Low-light



(b) CP3 (Low-light)

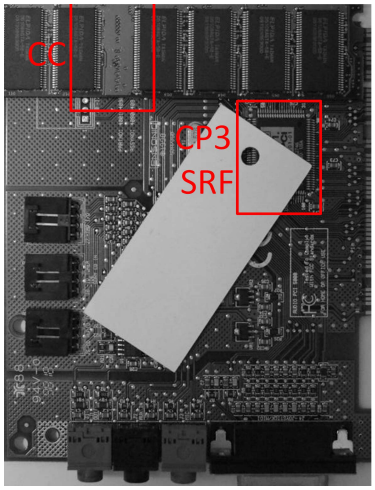


(c) SRF (Low-light)

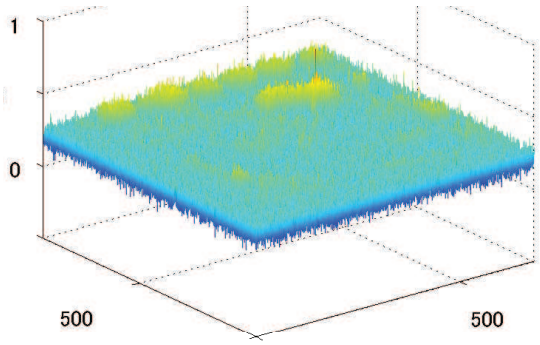


(d) NCC (Low-light)

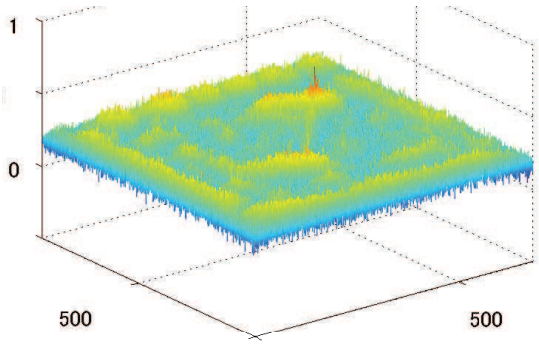
Figure 3.6: Experiment 1 (Low-light). Input images of a scene and matching results (600×800), and the corresponding similarity index maps of CP3, SRF and NCC.



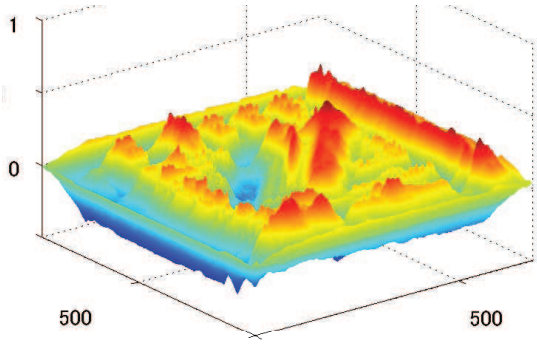
(a) Occlusion



(b) CP3 (Occlusion)



(c) SRF (Occlusion)



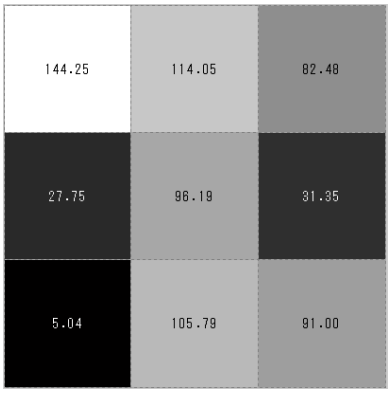
(d) NCC (Occlusion)

Figure 3.7: Experiment 1 (Occlusion). Input images of a scene and matching results (600×800), and the corresponding similarity index maps of CP3, SRF and NCC.

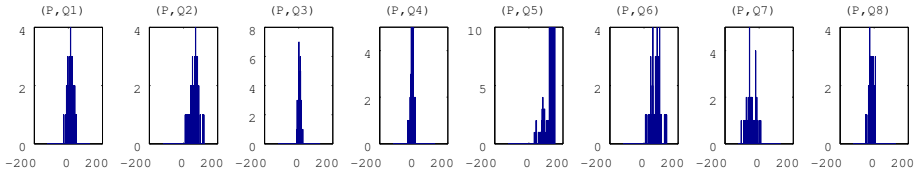
Table 3.1: SNR of CP3, SRF and NCC (Experiment 1).

Conditions	Method	Match	ξ_{truth}	ξ_{fault}	Δ	M	$\lvert\xi_{truth} - M\rvert$	D	SNR
Normal	CP3	○	0.649	-	-	0.217	0.432	0.105	4.114
	SRF	○	0.504	-	-	0.209	0.295	0.127	2.323
	NCC	○	0.351	-	-	0.075	0.276	0.140	1.971
Defocusing	CP3	○	0.570	-	-	0.259	0.311	0.137	2.270
	SRF	×	0.270	0.311	0.041	-	-	-	-
	NCC	○	0.338	-	-	0.087	0.251	0.154	1.630
Low-light	CP3	○	0.549	-	-	0.258	0.291	0.104	2.798
	SRF	×	0.394	0.417	0.023	-	-	-	-
	NCC	×	0.139	0.273	0.134	-	-	-	-
Occlusion	CP3	○	0.590	-	-	0.241	0.349	0.119	2.933
	SRF	○	0.435	-	-	0.110	0.325	0.124	2.621
	NCC	×	0.141	0.145	0.004	-	-	-	-

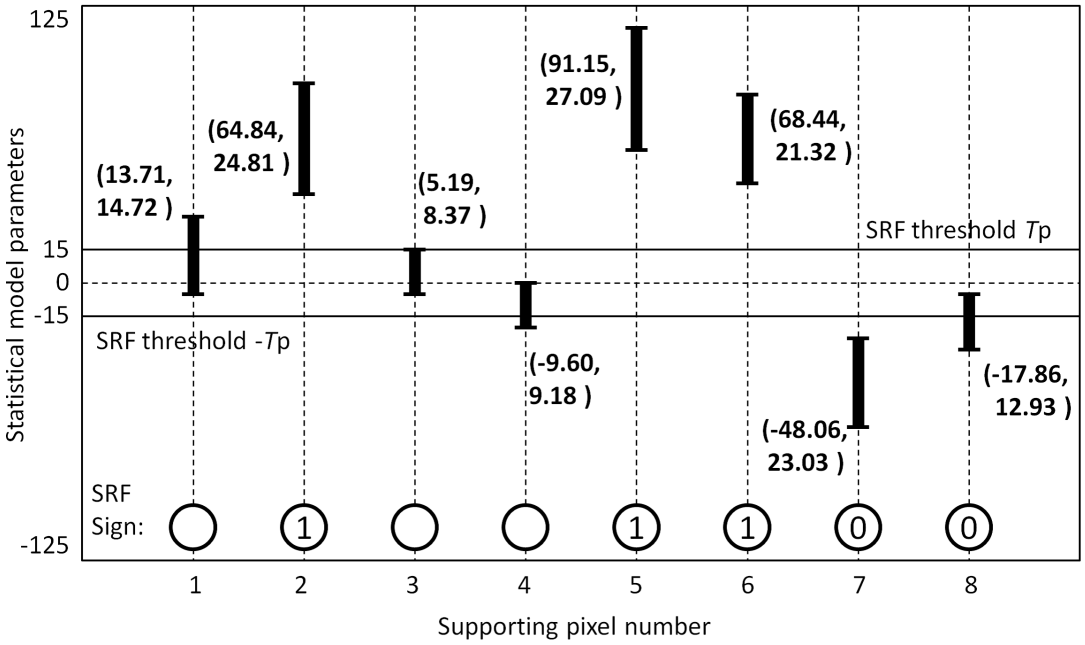
(a) A target pixel



(b) Magnified neighbourhood and their mean intensity values



(c) CP3 histogram



(d) CP3 and SRF sign. Pillars are the range of $\hat{b} \pm \hat{\sigma}_\varepsilon$ of CP3, values $\hat{b}$ and $\hat{\sigma}_\varepsilon$ are in brackets. SRF signs are in the circles at the bottom.

Figure 3.8: An example of CP3 and SRF.

3.5.2 Experiment 2

Next, we perform an experiment based on a real application, including comprehensive highlight and occlusion conditions. As shown in Figure 3.9, the input image of a scene is a group of glossy packages of food on a shelf of a supermarket under fluorescent lamp illumination. Moreover, the individual differences of the plastic covers of the packages lead to the variety of highlight and occlusions.

Using a robot vision device designed for locating the expiration date of food can improve the quality of service. However, it is very difficult to select an appropriate reference image when using traditional single reference image matching framework. In addition, in this provided example, there are eight potential targets in this scene, which upraise the difficulty for matching, where the eighth correlation peak should be higher than any of the disturbed local false maxima. We selected total 57 possible occurred reference images for CP3 learning, as shown in the left hand of Figure 3.9. For NCC and SRF, we used the average image of this sequence as a reference image.

In Figure 3.10, the similarity index map of CP3 has eight sharp spikes, which successfully locate the eight targets in the scene as shown in Figure 3.10. The similarity index maps of SRF and NCC are shown in Figure 3.11 and Figure 3.12, both of which lose some matching positions shown in Figure 3.11 and Figure 3.12.

Table 3.2 summarizes the corresponding quantitative analysis results. Since this scene contains multiple targets, to avoid confusion, we gave up using the departure degree of the true coefficient $\Delta(\xi_{fault}, \xi_{truth}) = |\xi_{fault} - \xi_{truth}|$ for comparison. The average SNR of CP3 is significantly higher than the one of both SRF and NCC.

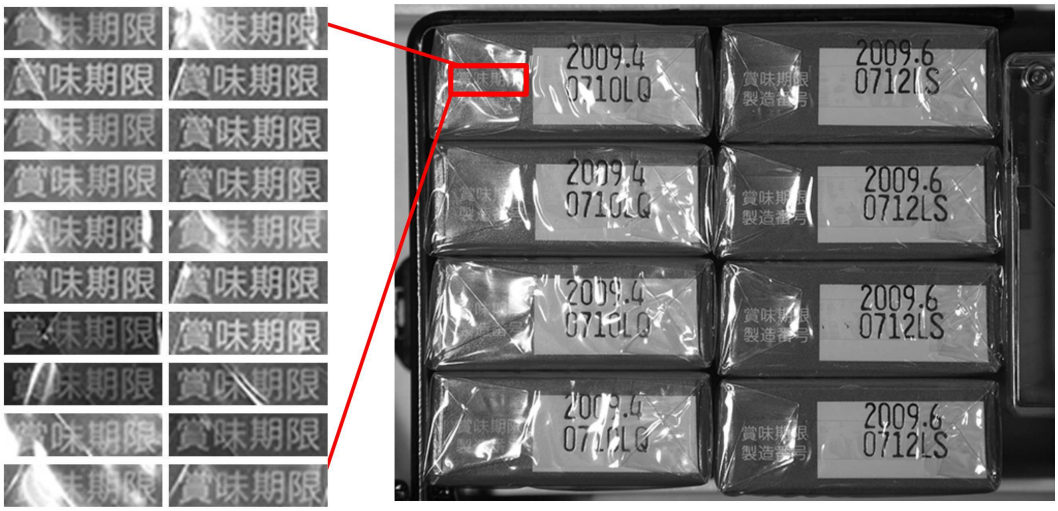
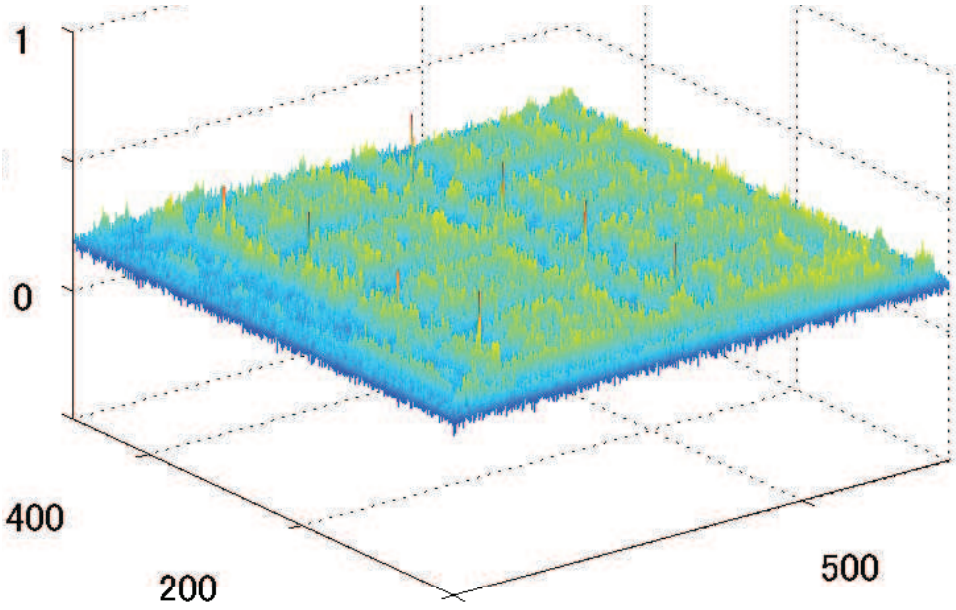


Figure 3.9: The right hand is the input image of the scene for Experiment 2 (640×480). The left hand is a subset of the prepared reference image sequence for CP3 learning (magnified version).

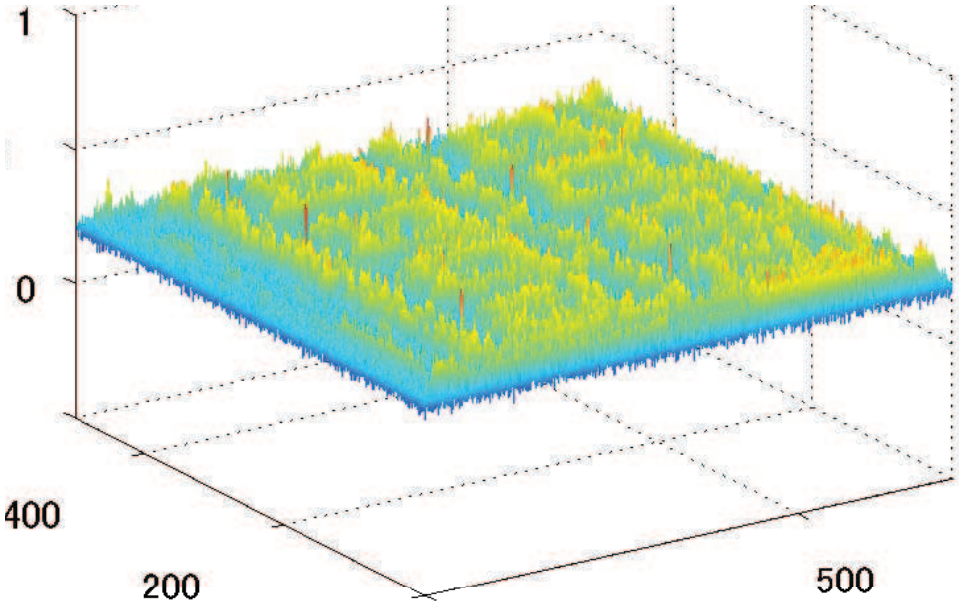
(a) CP3 matching



(b) CP3 similarity index map

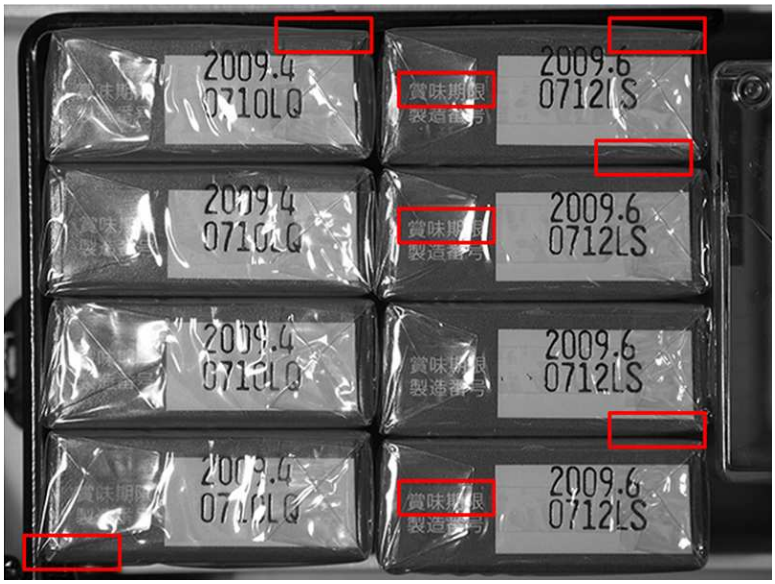
Figure 3.10: Experiment 2. Input images of a scene and matching results (640 × 480), and similarity index maps of CP3.

(a) SRF matching

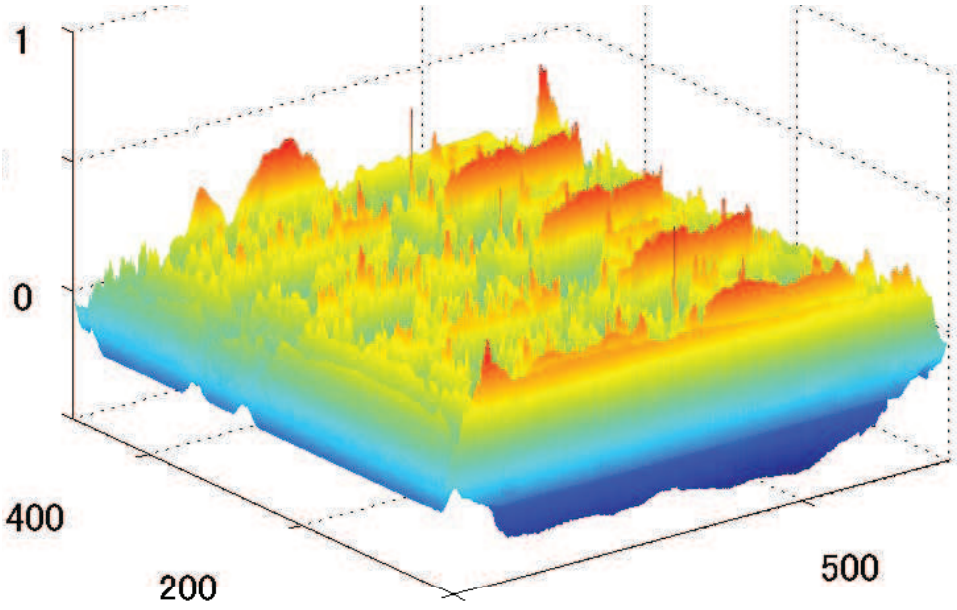


(b) SRF similarity index map

Figure 3.11: Experiment 2. Input images of a scene and matching results (640 × 480), and similarity index maps of SRF.



(a) NCC matching



(b) NCC similarity index map

Figure 3.12: Experiment 2. Input images of a scene and matching results (640 × 480), and similarity index maps of NCC.

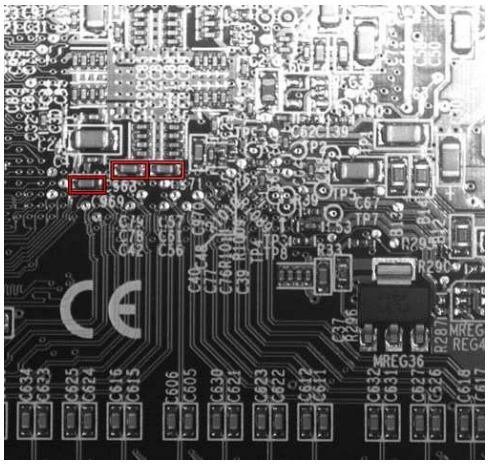
Table 3.2: Average SNR of CP3, SRF and NCC (Experiment 2).

Methods	ξ_{truth}	M	D	SNR
CP3	0.57	0.16	0.19	2.16
SRF	0.49	0.22	0.21	1.29
NCC	0.37	-0.05	0.39	1.08

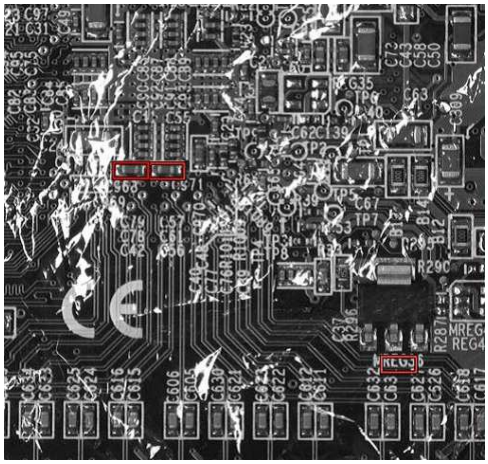
3.5.3 Experiment 3

Next, we performed an experiment based on the input image of another circuit board captured under partial illumination, highlight and occlusion imaging conditions, and three capacitors with a very tiny size and an identical model, are the targets, which is the most difficult case among the three experiments. As shown in Figure 3.13, we selected total 70 occurred reference images for CP3 learning. In Figure 3.14 (a), CP3 successfully locates the three targets in the scene under partial illumination condition. However, in Figure 3.14 (b) and (c), we could find that both of them lose a correct matching position, under much more severe occlusion and highlight situations.

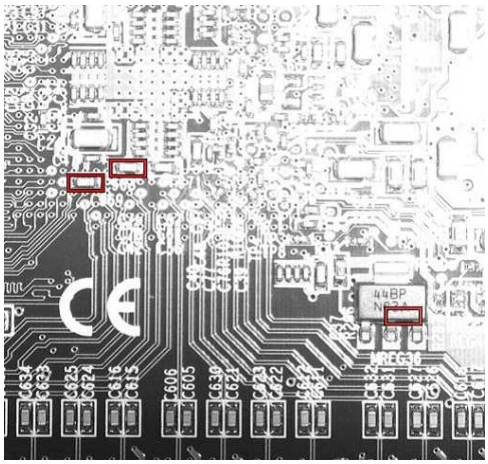
Figure 3.13: A subset of the prepared reference image sequence for CP3 learning for Experiment 3 (magnified version).



(a) Scene 1 (mathed all the 3 targets correctly)



(b) Scene 2 (mathed 2 targets correctly, mismathed 1 target under serve occlusion)



(c) Scene 3 (mathed 2 targets correctly, mismathed 1 target under serve highlight)

Figure 3.14: Experiment 3. (a)-(c) Input images of a scene and matching results using CP3.

3.5.4 Computation cost

Table 3.3 shows computation time per pixel for CP3, SRF and NCC, using the dataset in Experiment 2. In the learning stage, the processing time of CP3 for per pixel is about three times than that of the similarity measure stage. This is an acceptable time cost because the reference images are in much smaller size than the size of the input image of a scene. In similarity measure stage, the computational cost of CP3 is much smaller than the one of NCC because it does not do any pixel-wise multiply operation, but do a quite efficient LUT operation and bits counting operation. The computational cost of CP3 is almost the same level as the one of SRF because both of them contain an LUT operation with only two parameters, even if CP3 includes a more

accurate statistical model. The runtime is measured on a computer with a Intel Xeon 3.0 GHz processor and implemented by C language.

Table 3.3: Comparing computational cost of Experiment 2 ($\mu s/pixel$).

Methods	Learning stage	Similarity measure
CP3	1.093	0.381
SRF	-	0.342
NCC	-	2.277

3.6 Summary

In this chapter, an image matching method based on co-occurrence probability based pixel pairs, which has the nature of robustness for severe imaging conditions, has been proposed. By learning a set of reference images, the spatial and statistical models effectively improved the quality of similarity measure, and the pixel pair structure effectively reduces the disturbances of the illumination fluctuation. The corresponding quantitative analysis results clearly shows that in most severe imaging conditions, CP3 can capture a correct correlation peak, and the SNR values of CP3 are significantly higher than both the one of SRF and NCC.

There are still some works worth to be explored in this study. Firstly, the acquisition of the reference images needs to be improved so that it can be more flexibly applied to a robot system. Secondly, this method could be potentially integrated with some mathematical manipulation to deal with geometric variance, e.g. rotation, scale change, and affine change. Some candidate ways could be considered to integrate with CP3 to realize rotated and scaled image matching, for example in [62], a hierarchical frame work are employed for rotated image matching. A more general way is [49], which used Fourier-Mellin invariant (FMI) as a descriptor to deal with translated, rotated, and scaled reference images. We will consider how to combine the merits of our proposed method and the previous methods in our future work.

Chapter 4. Conclusion and future work

4.1 Conclusion

This thesis proposed a novel background model for object detection based on co-occurrence pixel pairs. This background model breaks the constraint of the assumption that the pixels' correlation degree depends on the spatial distance between them that each observed pixel has higher correlation with its neighbouring pixels.

The background model performs robust detection under outdoor and indoor extreme environments. CP3 determines stable co-occurrence pixel pairs, instead of building the parametrized/non-parametrized model for a single pixel. These pixel pairs maintain a reliable background model, which can be used to capture structural background motion and cope with local and global illumination changes. As a spatial-dependence method, CP3 does not predefine any local operator, subspace or block, and it provides an accurate detection criterion even though the gray-scale dynamic range is compressed under weak illumination.

Compared with GAP, the proposed method employs a co-occurrence histogram to describe the relationship of a pixel pair, which is free from any intensity differences, and calculates normalized correlation coefficients for measuring the degree of co-occurrence which can deal with a dynamic background. It also introduces a spatial clustering operation to select optimal supporting pixels and then provides a more accurate parametrized detection criterion instead of a fixed double-sided threshold.

The speed-up algorithm based on the pixel sampling operation allows us to achieve high speed background modelling when dealing with high resolution surveillance videos. when using the accelerated version of CP3 efficiently reduce the time cost of background modelling and maintain a fairly good detection performance. In addition, an optimized implementation for object detection can meet the requirement of real-time detection utilization.

In this study, an image matching method based on co-occurrence probability based pixel pairs, which has the nature of robustness for severe imaging conditions, has been also proposed. By learning a set of reference images, the spatial and statistical models effectively improved the quality of similarity measure, and the pixel pair structure effectively reduces the disturbances of the illumination fluctuation. Although an additional learning stage is necessary, the experiment results show that the proposed method is robust under several imaging cases and it also outperforms SRF method and traditional NCC. CP3 can capture a correct correlation peak, and the SNR values of CP3 are significantly higher than both the one of SRF and NCC. In similarity measure stage, the computational cost of CP3 is much smaller than the one of NCC and almost the same level as the one of SRF.

4.2 Future work

Shadow detection

In this study, we do not consider the problem of shadow. Moving shadows do, however, cause serious problems while extracting moving objects, due to the misclassification of shadow points as foreground. Shadows can cause merging of objects, object shape distortion and even object losses (due to the shadow cast over another object). The difficulties associated with shadow detection arise since shadow and objects share two important visual features. Shadows are dark and typically differ significantly from the background and they have the same motion as the objects casting them. We will introduce colour space analysis techniques and integrate CP3 into a shadow detection framework.

Image matching

There are still some works worth to be explored in this image matching study. Firstly, the acquisition of the reference images needs to be improved so that it can be more flexibly applied to a robot system. Secondly, this method could be potentially integrate with some mathematical manipulation to deal with geometric variance, e.g. rotation, scale change, and affine change. In future, the learning-based image matching approaches will play an essential and irreplaceable role. The authors will continue to develop such approaches.

Pedestrian tracking

Pedestrian tracking in real-world is a challenging task due to the presence of noise, occlusion, clutter, dynamic background elements and confusing background colours. The general way to track the pedestrian is to model the appearance of the pedestrian using an observation model. The commonly used observation models built for tracking are edge-based, colour-based, and contour-based features. However, algorithms relying on only one feature are less robust and suffer from various limitations in complex scenarios. The colour feature is robust to noise and partial occlusion but suffers from illumination changes. The edge and contour features are more robust to illumination variation compared to the colour feature but are much sensitive to background clutter. Due to the merits of CP3 in background modelling for object detection and appearance modelling for image matching, we will consider some specific detail to utilize CP3 as an appearance modelling for the elastic body of the pedestrian, and will integrate CP3 as a robust feature into an pedestrian tracking framework, such as Kalman filtering, Mean-shift or Sequential Monte Carlo, in our future work.

Acknowledgements

Thank you, my supervisor, Professor Shun'ichi Kaneko, for being not just an academic supervisor. Time and time again, your wise guidance help me to build clear thinking and deep understanding in our study, your warm encouragement, your plenty of enthusiasm, your patient comments and advices help me to build the soul of this study. Moreover, you taught me generic solutions to analyze problems, you taught me philosophical thought, and you taught me effective communication with partners. You let me know how to go farther and broader to fulfil both the academic mission and the social responsibility in my future work. All of these will have a remarkable influence on my whole life.

Thank you, Professor Masahiko Onosato, for reviewing my thesis and providing me so many key comments and suggestions.

Thank you, Associate Professor Takayuki Tanaka. You participated in many fruitful discussions concerning my research. You give me many valuable suggestions when I encounter difficulties. You provide me a good chance to enrich my theoretical basis and to have the experience in engineering field.

Thank you, Professor Manabu Hashimoto from Chukyo University, Japan, Dr. Yutaka Satoh and Dr. Kenji Iwata from National Institute of Advanced Industrial Science and Technology (AIST), Japan, Assistant Professor Xinyue Zhao from Zhejiang University, China, for co-supervising and supporting me. During three years study in pattern recognition and computer vision, all of you contribute lots of original ideas to the projects. Without your valuable advice and energetic assistance, this study would not be completed.

Thank you, all staffs in Human-Centric Engineering Laboratory, Hokkaido University: Assistant Professor Akihiko Matsushita, Ms. Miyuki Hashimoto, Ms. Miki Ki'kawa, Dr. Takashi Kusaka, Dr. Yumeko Imamura. Thank you for all your help during my research.

All my lab buddies at Human-Centric Engineering Laboratory, Hokkaido University, made it a convivial place to work. I would like to thank: Dr. Deshan Chen, Ms. Rong Zhou, Mr. Shoichi Kondo, Mr. Masaya Ito, Mr. Tenko Oh, Mr. Nobuyuki Tamai, Mr. Yoshio Tsuchiya, Mr. Keigo Yoneda, Mr. Zhantao Lai, Mr. Motoaki Wakasugi, Mr. Shinnichiro Aoyama, Mr. Naoki Nakano, Mr. Soma Shirakabe, Ms. Saori Miyajima, Mr. Kyohei Wada, and all the previous students in this lab. All your frank conversation with me builds a trustful personal relationship. Thank you for sharing me so many wonderful and unforgettable days and nights together in the same research room.

Thank you, my wife Liyv Liu. During the past 42 months in Japan, we share depression and happiness with each other, we realize our dreams one by one and step by step. I am so lucky to have you in my life because you give me the root to always know where home is, and give me

the wings that let me fly and explore the new world.

Thank you, my parents, Chuanwen Liang and Fangfang Qv. Thank you for always being close to me, despite the physical distance. Thank you for rising me up and looking over my shoulder, everywhere, anywhere, always. Without your encouragement and understanding, it would be impossible to finish my Ph.D. study.

Finally, I appreciate China Scholarship Council and Hokkaido University, who provide me abundant financial support.

References

[1] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Computing Surveys 38 (4) (2006) 1–43.

[2] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–575.

[3] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[4] I. Haritaoglu, D. Harwood, L. S. Davis, W4: real-time surveillance of people and their activities, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 809–830.

[5] A. Datta, M. Shah, N. Da Vitoria Lobo, Person-on-person violence detection in video data, in: 16th International Conference on Pattern Recognition, Vol. 1, IEEE, 2002, pp. 433–438.

[6] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34 (3) (2004) 334–352.

[7] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding 104 (2) (2006) 90–126.

[8] A. Iosifidis, S. G. Mouroutsos, A. Gasteratos, A hybrid static/active video surveillance system, International Journal of Optomechatronics 5 (1) (2011) 80–95.

[9] L. Nalpantidis, G. C. Sirakoulis, A. Gasteratos, Review of stereo vision algorithms: From software to hardware, International Journal of Optomechatronics 2 (4) (2008) 435–462.

[10] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, Pfinder: Real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 780–785.

[11] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, IEEE, 1999.

[12] A. Elgammal, R. Duraiswami, D. Harwood, L. S. Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, Proceedings of the IEEE 90 (7) (2002) 1151–1163.

[13] K. Kim, T. H. Chalidabhongse, D. Harwood, L. Davis, Real-time foreground–background segmentation using codebook model, Real-time imaging 11 (3) (2005) 172–185.

[14] Y. Sheikh, M. Shah, Bayesian modeling of dynamic scenes for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1778–1792.

[15] M. Sonka, V. Hlavac, R. Boyle, Image processing, analysis, and machine vision (Second Edition), PWS Publishing, 1999.

[16] C. Stauffer, W. E. L. Grimson, Learning patterns of activity using real-time tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

[17] J. Rittscher, J. Kato, S. Joga, A. Blake, A probabilistic background model for tracking, Computer Vision-ECCV 2000 (2000) 336–350.

[18] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, J. M. Buhmann, Topology free hidden markov models: Application to background modeling, in: Eighth IEEE International Conference on Computer Vision, Vol. 1, IEEE, 2001, pp. 294–301.

[19] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., Vol. 2, IEEE, 1999.

[20] J.-M. Guo, Y.-F. Liu, C.-H. Hsia, M.-H. Shih, C.-S. Hsu, Hierarchical method for foreground detection using codebook model, IEEE Transactions on Circuits and Systems for Video Technology 21 (6) (2011) 804–815.

[21] T. Matsuyama, T. Ohya, H. Habe, Background subtraction for non-stationary scenes, in: Proceedings of Asian Conference on Computer Vision, 2000, pp. 662–667.

[22] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: Principles and practice of background maintenance, in: The Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 1, IEEE, 1999, pp. 255–261.

[23] N. M. Oliver, B. Rosario, A. P. Pentland, A bayesian computer vision system for modeling human interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 831–843.

[24] A. Monnet, A. Mittal, N. Paragios, V. Ramesh, Background modeling and subtraction of dynamic scenes, in: Ninth IEEE International Conference on Computer Vision, IEEE, 2003, pp. 1305–1312.

[25] M. Heikkilä, M. Pietikäinen, A texture-based method for modeling the background and detecting moving objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 657–662.

[26] W. Hu, X. Li, X. Zhang, X. Shi, S. Maybank, Z. Zhang, Incremental tensor subspace learning and its applications to foreground segmentation and tracking, International journal of computer vision 91 (3) (2011) 303–327.

[27] M. Seki, T. Wada, H. Fujiwara, K. Sumi, Background subtraction based on cooccurrence of image variations, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, IEEE, 2003, pp. II–65.

[28] J. Zhong, S. Sclaroff, Segmenting foreground objects from a dynamic textured background via a robust kalman filter, in: Ninth IEEE International Conference on Computer Vision, IEEE, 2003, pp. 44–50.

[29] X. Zhao, Y. Satoh, H. Takauji, S. Kaneko, K. Iwata, R. Ozaki, Object detection based on a robust and accurate statistical multi-point-pair model, Pattern Recognition 44 (6) (2011) 1296–1311.

[30] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, X. Zhao, Y. Satoh, Co-occurrence-based adaptive background model for robust object detection, in: Proceedings of 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2013, pp. 401–406.

[31] X. ZHAO, Y. Satoh, H. Takauji, S. Kaneko, Robust tracking using particle filter with a hybrid feature, IEICE TRANSACTIONS on Information and Systems 95 (2) (2012) 646–657.

[32] X. Zhao, Z. He, S. Zhang, S. Kaneko, Y. Satoh, Robust face recognition using the gap feature, Pattern Recognition 46 (10) (2013) 2647–2657.

[33] X. Zhao, Z. He, S. Zhang, Defect detection of castings in radiography images using a robust statistical feature, JOSA A 31 (1) (2014) 196–205.

[34] B. Zitova, J. Flusser, Image registration methods: a survey, Image and vision computing 21 (11) (2003) 977–1000.

[35] J. Maintz, M. A. Viergever, A survey of medical image registration, Medical image analysis 2 (1) (1998) 1–36.

[36] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[37] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 404–417.

[38] J. Aggarwal, L. S. Davis, W. Martin, Correspondence processes in dynamic scene analysis, Proceedings of the IEEE 69 (5) (1981) 562–572.

[39] D. I. Barnea, H. F. Silverman, A class of algorithms for fast digital image registration, IEEE Transactions on Computers 100 (2) (1972) 179–186.

[40] S. Kaneko, Y. Satoh, S. Igarashi, Using selective correlation coefficient for robust image registration, Pattern Recognition 36 (5) (2003) 1165–1173.

[41] S. Kaneko, I. Murase, S. Igarashi, Robust image registration by increment sign correlation, Pattern Recognition 35 (10) (2002) 2223–2234.

[42] D. H. Ballard, Generalizing the hough transform to detect arbitrary shapes, Pattern recognition 13 (2) (1981) 111–122.

[43] Y. Lamdan, H. J. Wolfson, Geometric hashing: A general and efficient model-based recognition scheme, in: IEEE International Conference on Computer Vision, Vol. 88, 1988, pp. 238–249.

[44] D. N. Bhat, S. K. Nayar, Ordinal measures for image correspondence, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (4) (1998) 415–423.

[45] R. Ozaki, Y. Satoh, K. Iwata, K. Sakaue, Statistical reach feature method and its application to robust image registration, in: TENCON 2009 - 2009 IEEE Region 10 Conference, 2009, pp. 1–6.

[46] R. Ozaki, Y. Satoh, K. Iwata, K. Sakaue, Template matching by the statistical reach feature method, Electronics and Communications in Japan 96 (11) (2013) 54–69.

[47] F. Ullah, S. Kaneko, Using orientation codes for rotation-invariant template matching, Pattern recognition 37 (2) (2004) 201–209.

[48] D. Liang, Z. Wang, Y. Ma, A novel approach for plant embryonic cells serial section images registration, in: 3rd International Conference on Bioinformatics and Biomedical Engineering, IEEE, 2009, pp. 1–4.

[49] Q.-s. Chen, M. Defrise, F. Deconinck, Symmetric phase-only matched filtering of fourier-mellin transforms for image registration and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (12) (1994) 1156–1168.

[50] J.-M. Morel, G. Yu, Asift: A new framework for fully affine invariant image comparison, SIAM Journal on Imaging Sciences 2 (2) (2009) 438–469.

[51] J. Kybic, M. Unser, Fast parametric elastic image registration, IEEE Transactions on Image Processing 12 (11) (2003) 1427–1442.

[52] C. A. Glasbey, K. V. Mardia, A review of image-warping methods, Journal of applied statistics 25 (2) (1998) 155–171.

[53] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, X. Zhao, Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes, Pattern Recognition 48 (4) (2015) 1370–1386.

[54] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, X. Zhao, Y. Satoh, Robust object detection in severe imaging conditions using co-occurrence background model, International Journal of Optomechatronics 8 (1) (2014) 14–29.

[55] R. M. Haralick, K. Shanmugam, I. H. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics (6) (1973) 610–621.

[56] M. Hashimoto, M. Saito, High-speed and robust image matching using spatially distinctive and temporally stable pixels, in: 2011 International Symposium on Optomechatronic Technologies, IEEE, 2011, pp. 1–8.

[57] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, California, USA, 1967.

[58] S. Y. Elhabian, K. M. El-Sayed, S. H. Ahmed, Moving object detection in spatial domain using background removal techniques-state-of-art, Recent patents on computer science 1 (1) (2008) 32–54.

[59] T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (8) (2006) 861–874.

[60] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel, Performance measures for information extraction, in: In Proceedings of DARPA Broadcast News Workshop, 1999, pp. 249–252.

[61] A. Shimada, H. Nagahara, R.-i. Taniguchi, Background modeling based on bidirectional analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 1979–1986.

[62] F. Ullah, S. Kaneko, Using orientation codes for rotation-invariant template matching, Pattern recognition 37 (2) (2004) 201–209.

Appendix A Publication lists

A.1 Journal Papers

1. Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, Co-occurrence Probability-based Pixel Pairs Background Model for Robust Object Detection in Dynamic Scenes, Pattern Recognition, 48 (2015) 1370-1386.

2. Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, Yutaka Satoh, Robust Object Detection in Severe Imaging Conditions using Co-occurrence Background Model, International Journal of Optomechatronics, Volume 8, Issue 1, pp.14-29, 2014.

3. Dong Liang, Shun'ichi Kaneko, Yutaka Satoh, A Robust Appearance Model and Similarity Measure for Image Matching, Journal of Robotics and Mechatronics, Vol.27 No.2 (will appear on Apr. 20, 2015).

A.2 International Conferences

1. Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, Yutaka Satoh, Co-occurrence-based Adaptive Background Model for Robust Object Detection, 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp.401-406, 27-30 Aug. 2013.

2. Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, Yutaka Satoh, Robust Object Detection in Severe Imaging Conditions using Co-occurrence Background Model, 2013 International Symposium on Optomechatronic Technologies, 14 pages, 29-31 Oct. 2013.

3. Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, Statistical Spatial Multi-pixel-pair Model for Object Detection, 2012 International Symposium on Optomechatronic Technologies, 6 pages, 29-31 Oct. 2012.

A.3 Domestic Conference

1. Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, Yutaka Satoh, Co-occurrence-based Adaptive Background Model for Robust Object Detection,

Annual Conference of Institute of Electrical Engineers-Electronics, Information and Systems Division, pp.307-312, 2013.