Name: Wang Yang

Degree: Ph.D.

Dept: Computer Science

Thesis Title: Segmenting and tracking objects in video sequences based on graphical probabilistic models

Abstract

Segmenting and tracking objects in video sequences is important in vision-based application areas, but the task could be difficult due to the potential variability such as object occlusions and illumination variations. In this thesis, three techniques of segmenting and tracking objects in image sequences are developed based on graphical probabilistic models (or graphical models), especially Bayesian networks and Markov random fields. First, this thesis presents a unified framework for video segmentation based on graphical models. Second, this work develops a dynamic hidden Markov random field (DHMRF) model for foreground object and moving shadow segmentation. Third, this thesis proposes a switching hypothesized measurements (SHM) model for multi-object tracking. By means of graphical models, the techniques deal with object segmentation and tracking from relatively comprehensive and general viewpoints, and thus can be universally employed in various application areas. Experimental results show that the proposed approaches robustly deal with the potential variability and accurately segment and track objects in video sequences.

Keywords: Bayesian network, foreground segmentation, graphical model, Markov random field, multi-object tracking, video segmentation.

SEGMENTING AND TRACKING OBJECTS IN VIDEO SEQUENCES BASED ON GRAPHICAL PROBABILISTIC MODELS

WANG YANG

(***B.Eng., Shanghai Jiao Tong University, China***)

(***M.Sc., Shanghai Jiao Tong University, China***)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2004

Acknowledgements

First of all, I would like to present sincere thanks to my supervisors, Dr. Kia-Fock Loe, Dr. Tele Tan, and Dr. Jian-Kang Wu, for their insightful guidance and constant encouragement throughout my Ph.D. study. I am grateful to Dr. Li-Yuan Li, Dr. Kar-Ann Toh, Dr. Feng Pan, Mr. Ling-Yu Duan, Mr. Rui-Jiang Luo, and Mr. Hai-Hong Zhang for their fruitful discussions and suggestions. I also would like to thank both National University of Singapore and Institute for Infocomm Research for their generous financial assistance during my postgraduate study. Moreover, I would like to acknowledge Dr. James Davis, Dr. Ismail Haritaoglu, and Dr. Andrea Prati et al. for providing test data on their websites. Last but not the least, I wish to express deep thanks to my parents for their endless love and support when I am studying abroad in Singapore.

Table of contents

List of figures

List of table

Summary

Object segmentation and tracking are employed in various application areas including visual surveillance, human-computer interaction, video coding, and performance analysis. However, to effectively and efficiently segment and track objects of interest in video sequences could be difficult due to the potential variability in complex scenes such as object occlusions, illumination variations, and cluttered environments. Fortunately, graphical probabilistic models provide a natural tool for handling uncertainty and complexity with a general formalism for compact representation of joint probability distribution. In this thesis, techniques of segmenting and tracking objects in image sequences are developed to deal with the potential variability in visual processes based on graphical models, especially Bayesian networks and Markov random fields.

Firstly, this thesis presents a unified framework for spatio-temporal segmentation of video sequences. Motion information among successive frames, boundary information from intensity segmentation, and spatial connectivity of object segmentation are unified in the video segmentation process using graphical models. A Bayesian network is presented to model interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. The notion of Markov Random field is used to encourage the formation of continuous regions. Given consecutive frames, the conditional joint probability density of the three fields is maximized in an iterative way. To effectively utilize boundary information from intensity segmentation, distance transformation is employed in local optimization. Moreover, the proposed video segmentation approach can be viewed as a compromise between previous motion based approach and region merging approach.

Secondly, this work develops a dynamic hidden Markov random field (DHMRF) model for foreground object and moving shadow segmentation in indoor video scenes monitored by fixed camera. Given an image sequence, temporal dependencies of consecutive segmentation fields and spatial dependencies within each segmentation field are unified in the novel dynamic probabilistic model that combines the hidden Markov model and the Markov random field. An efficient approximate filtering algorithm is derived for the DHMRF model to recursively estimate the segmentation field from the history of observed images. The foreground and shadow segmentation method integrates both intensity and edge information. In addition, models of background, shadow, and edge information are updated adaptively for nonstationary background processes. The proposed approach can robustly handle shadow and camouflage in nonstationary background scenes and accurately detect foreground and shadow even in monocular grayscale sequences.

Thirdly, this thesis proposes a switching hypothesized measurements (SHM) model supporting multimodal probability distributions and applies the model to deal with object occlusions and appearance changes when tracking multiple objects jointly. For a set of occlusion hypotheses, a frame is measured once under each hypothesis, resulting in a set of measurements at each time instant. The dynamic model switches among hypothesized measurements during the propagation. A computationally efficient SHM filter is derived for online joint object tracking. Both occlusion relationships and states of the objects are recursively estimated from the history of hypothesized measurements. The reference image is updated adaptively to deal with appearance changes of the objects. Moreover, the SHM model is generally applicable to various dynamic processes with multiple alternative measurement methods.

By means of graphical models, the proposed techniques handle object segmentation and tracking from relatively comprehensive and general viewpoints, and thus can be utilized in diverse application areas. Experimental results show that the proposed approaches robustly handle the potential variability such as object occlusions and illumination changes and accurately segment and track objects in video sequences.

Chapter 1

Introduction

1.1 Motivation

With the significant enhancement of machine computation power in recent years, in computer vision community there is a growing interest in segmenting and tracking objects in video sequences. The technique is useful in a wide spectrum of application areas including visual surveillance, human-computer interaction, video coding, and performance analysis.

In automatic visual surveillance systems, usually imaging sensors are mounted around a given site (e.g. airport, highway, supermarket, or park) for security or safety. Objects of interest in video scenes are tracked over time and monitored for specific purposes. A typical example is the car park monitoring, where the surveillance system detects car and people to estimate whether there is any crime such as car stealing to be committed in video scenes.

Vision based human-computer interaction builds convenient and natural interfaces for users through live video inputs. Users' actions or even their expressions in video data are captured and recognized by machines to provide controlling functionalities. The technique can be employed to develop game interfaces, control remote instruments, and construct virtual reality.

Modern video coding standards such as MPEG-4 focus on content-based manipulation of video data. In object-based compression schemes, video frames are decomposed into independently moving objects or coherent regions rather than into

fixed square blocks. The coherence of video segmentation helps improve the efficiency in video coding and allow object-oriented functionalities for further analysis. For example, in a videoconference, the system can detect and track faces in video scenes, then preserve more details for faces than for the background in coding.

Another application domain is performance analysis, which involves detailed tracking and analyzing human motion in video streams. The technique can be utilized to diagnose orthopedic patients in clinical studies and to help athletes enhance their performance in competitive sports.

In such applications, the ability to segment and track objects of interest is one of the key issues in the design and analysis of the vision system. However, usually real visual environments are very complex for machines to understand the structure in the scene. Effective and efficient object segmentation and tracking in image sequences could be difficult due to the potential variability such as partial or full occlusions of objects, appearance changes caused by illumination variations, as well as distractions from cluttered environments.

Fortunately, graphical probabilistic models (or graphical models) provide a natural tool for handling uncertainty and complexity through a general formalism for compact representation of joint probability distribution [33]. In particular, Bayesian networks and Markov random fields attract more and more attention in the design and analysis of machine intelligent systems [14], and they are playing an increasingly important role in many application areas including video analysis [12]. The introduction of Bayesian networks and Markov random fields can be found in [30] [37].

In this thesis, probabilistic approaches of object segmentation and tracking in video sequences based on graphical models are studied to deal with the potential variability in visual processes.

1.2 Organization

The rest chapters of the thesis are arranged as follows.

Chapter 2 gives a brief review of sate-of-the-art research on segmenting and tracking objects in video sequences. Section 2.1 surveys current work on video segmentation, Section 2.2 covers existing work on foreground segmentation by background subtraction, and Section 2.3 describes current research on multi-object tracking.

Chapter 3 develops a graphical model based approach for video segmentation. Section 3.1 introduces our technique and the related work. Section 3.2 presents the formulation of the approach. Section 3.3 proposes the optimization scheme. Section 3.4 discusses the experimental results.

Chapter 4 presents a dynamic hidden Markov random field (DHMRF) model for foreground object and moving shadow segmentation. Section 4.1 introduces our technique and the related work. Section 4.2 proposes the DHMRF model and derives its filtering algorithm. Section 4.3 presents the foreground and shadow detection method. Section 4.4 describes the implementation details. Section 4.5 discusses the experimental results.

Chapter 5 proposes a switching hypothesized measurements (SHM) model for joint multi-object tracking. Section 5.1 introduces our technique and the related work. Section 5.2 presents the formulation of the SHM model. Section 5.3 proposes the measurement process for joint region tracking. Section 5.4 derives the filtering

algorithm. Section 5.5 describes the implementation details. Section 5.6 discusses the experimental results.

Chapter 6 concludes our work. Section 6.1 summarizes the proposed techniques. Section 6.2 suggests the future research.

1.3 Contributions

As for the main contribution in this thesis, three novel techniques for segmenting and tracking objects in video sequences have been developed by means of graphical models to deal with the potential variability in visual environments.

Chapter 3 proposes a unified framework for spatio-temporal segmentation of video sequences based on graphical models [71]. Motion information among successive frames, boundary information from intensity segmentation, and spatial connectivity of object segmentation are unified in the video segmentation process using graphical models. A Bayesian network is presented to model interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. Markov random field and distance transformation are employed to encourage the formation of continuous regions. In addition, the proposed video segmentation approach can be viewed as a compromise between previous motion based approach and region merging approach.

Chapter 4 presents a dynamic hidden Markov random field (DHMRF) model for foreground object segmentation by background subtraction and shadow removal [67]. Given a video sequence, temporal dependencies of consecutive segmentation fields and spatial dependencies within each segmentation field are unified in the novel dynamic probabilistic model that combines the hidden Markov model and the

Markov random field. An efficient approximate filtering algorithm is derived for the DHMRF model to recursively estimate the segmentation field from the history of observed images. The proposed approach can robustly handle shadow and camouflage in nonstationary background scenes and accurately detect foreground and shadow even in monocular grayscale sequences.

Chapter 5 proposes a switching hypothesized measurements (SHM) model supporting multimodal probability distributions and applies the SHM model to deal with visual occlusions and appearance changes when tracking multiple objects [68]. An efficient approximate SHM filter is derived for online joint object tracking. Moreover, the SHM model is generally applicable to various dynamic processes with multiple alternative measurement methods.

By means of graphical models, the techniques are developed from relatively comprehensive and general viewpoints, and thus can be employed to deal with object segmentation and tracking in diverse application areas. Experimental results tested on public video sequences show that the proposed approaches robustly handle the potential variability such as partial or full occlusions and illumination or appearance changes as well as accurately segment and track objects in video sequences.

Chapter 2

Object Segmentation and Tracking: A Review

2.1 Video segmentation

Given a video sequence, it is important for a system to segment independently moving objects composing the scene in many applications including human-computer interaction and object-based video coding. One essential issue in the design of such systems is the strategy to extract and couple motion information and intensity information during the video segmentation process.

Motion information is one fundamental element used for segmentation of video sequences. A moving object is characterized by coherent motion over its support region. The scene can be segmented into a set of regions, such that pixel movements within each region are consistent with a motion model (or a parametric transformation) [66]. Examples of motion models are the translational model (two parameters), the affine model (six parameters), and the perspective model (eight parameters). Furthermore, spatial constraint could be imposed on the segmented region where the motion is assumed to be smooth or follow a parametric transformation. In the work of [9] [59] [65], the motion information and segmentation are simultaneously estimated. Moreover, layered approaches have been proposed to represent multiple moving objects in the scene with a collection of layers [31] [32] [62]. Typically, the expectation maximization (EM) algorithm is employed to learn the multiple layers in the image sequence.

On the other hand, intensity segmentation provides important hints of object boundaries. Methods that combine an initial intensity segmentation with motion information have been proposed [19] [41] [46] [64]. A set of regions with small intensity variation is given by intensity segmentation (or oversegmentation) of the current frame. Objects are then formed by merging together regions with coherent motion. The region merging approaches have two disadvantages. Firstly, the intensity segmentation remains unchanged so that motion information has no influence upon the segmentation during the entire process. Secondly, even an oversegmentation sometimes cannot keep all the object edges, and the boundary information lost in the initial intensity segmentation cannot be recovered later. Since motion information and intensity information should interact throughout the segmentation process, to utilize only motion estimation or fix intensity segmentation will degrade the performance of video segmentation. From this point of view, it is relatively comprehensive to simultaneously estimate the motion vector field, the intensity segmentation field, and the object segmentation field.

2.2 Foreground segmentation

When the video sequence is captured using a fixed camera, background subtraction is a commonly used technique to segment moving objects. The background model is constructed from observed images and foreground objects are identified if they differ significantly from the background. However, accurate foreground segmentation could be difficult due to the potential variability such as moving shadows cast by foreground objects, illumination or object changes in the background, and camouflage (i.e. similarity between appearances of foreground objects and the background) [6] [49] [72]. Besides local measurements such as depth and

chromaticity [22] [25] [28] [39], constraints in temporal and spatial information from the video scene are very important to deal with the potential variability during the segmentation process.

Temporal or dynamic information is a fundamental element to handle the evolution of the scene. The background model can be adaptively updated from the recent history of observed images to handle nonstationary background processes (e.g. illumination changes). In addition, once a foreground point is detected, it will probably continue being in the foreground for some time. Linear prediction of background changes from recent observations can be performed by Kalman filter [36] or Wiener filter [63] to deal with dynamics in background processes. In the W^4 system [24], a bimodal background model is built for each site from order statistics of recent observed values. In [15], the pixel intensity is modeled by a mixture of three Gaussians (for moving object, shadow, and background respectively), and an incremental EM algorithm is used to learn the pixel model. In [57], the recent history of a pixel is modeled by a mixture of (usually three to five) Gaussians for nonstationary background processes. In [13], nonparametric kernel density estimation is employed for adaptive and robust background modeling. Moreover, a hidden Markov model (HMM) is used to impose the temporal continuity constraint on foreground and shadow detection for traffic surveillance [52]. A dynamical framework of topology free HMM capable of dealing with sudden or gradient illumination changes is also proposed in [58].

Spatial information is another essential element to understand the structure of the scene. Spatial variation information such as gradient (or edge) feature helps improve the reliability of structure change detection. In addition, contiguous points are likely

to belong to the same background or foreground region. [29] classifies foreground versus background by adaptive fusion of color and edge information using confidence maps. [56] assumes that static edges in the background remain under shadow and that penumbras exist at the boundary of shadows. In [54], spatial cooccurrence of image variations at neighboring blocks is employed to improve the detection sensitivity of background subtraction. Moreover, spatial smooth constraint is imposed on moving object and shadow detection by propagating neighborhood information [40]. In [45], spatial interaction constraint is modeled by the Markov random field (MRF). In [34], a three dimensional MRF model called spatio-temporal MRF involving two successive video frames is also proposed for occlusion robust segmentation of traffic images.

To robustly deal with the potential variability including shadow and camouflage for foreground segmentation, it will be relatively comprehensive to unify various temporal and spatial constraints in video sequences during the segmentation process.

2.3 Multi-object tracking

Multi-object tracking is important in application areas such as visual surveillance and human-machine interaction. Given a sequence of video frames containing the objects that are represented with a parametric motion model, the model parameters are required to be estimated in successive frames. Visual tracking could be difficult due to the potential variability such as partial or full occlusions of objects, appearance changes caused by variation of object poses or illumination conditions, as well as distractions from background clutter.

The variability in visual environments usually results in a multimodal state space probability distribution. Thus, one principle challenge for visual tracking is to develop an accurate and effective model representation. The Kalman filter [7] [43], a classical choice in early tracking work, is limited to representing unimodal probability distributions. Joint probabilistic data association (JPDA) [3] and multiple hypothesis tracking (MHT) [11] techniques are able to represent multimodal distributions by constructing data association hypotheses. A measurement in the video frame may either belong to a target or be a false alarm. The multiple hypotheses arise when there are more than one target and many measurements in the scene. Dynamic Bayesian networks (DBN) [20], especially switching linear dynamic systems (SLDS) [47] [48] and their equivalents [21] [35] [42] [55] have been used to track dynamic processes. The state of a complex dynamic system is represented with a set of linear models controlled by a switching variable. Moreover, Monte Carlo methods such as the Condensation algorithm [27] [38] support multimodal probability densities with sample based representation. By saving only the peaks of the probability density, relatively fewer samples are required in the work of [8].

On the other hand, measurements are not readily available from video frames in visual tracking. Even an accurate tracking model may have a poor performance if the measurements are too noisy. Thus, the measurement process is another essential issue in visual tracking to deal with the potential variability. Parametric models can be used to describe appearance changes of target regions [23]. In the work of [16] and [17], adaptive or virtual snakes are used to resolve the occlusion. A joint measurement process for tracking multiple objects is described in [51]. Moreover, layered approach [32] [60] is an efficient way to represent multiple moving objects

during visual tracking, where each moving object is characterized by a coherent motion model over its support region.

To robustly handle the potential variability including occlusions during multi-object tracking, it will be relatively comprehensive to develop a multimodal model together with an occlusion adaptive measurement process.

Chapter 3

A Graphical Model Based Approach of Video Segmentation

3.1 Introduction

This chapter presents a probabilistic framework of video segmentation in which spatial (or motion) information and temporal (or intensity) information act on each other during the segmentation process. A Bayesian network is proposed to model the interactions among the motion vector field, the intensity segmentation field, and the video (or object) segmentation field. The notion of Markov random field (MRF) is employed to boost spatial connectivity of segmented regions. A three-frame approach is adopted to deal with occlusions. The segmentation criterion is the maximum a posteriori (MAP) estimate of the three fields given consecutive video frames. To perform the optimization, we propose a procedure that minimizes the corresponding objective functions in an iterative way. Distance transformation is employed in local optimization to effectively couple the boundary information from intensity segmentation. Experiments show that our technique is robust and generates spatio-temporally consistent segmentation results. Theoretically, the proposed video segmentation approach can be viewed as a compromise between motion based approach and region merging approach.

Our method is closely related to the work of Chang et al. [9] and Patras et al. [46]. Both approaches simultaneously estimate the motion vector field and the video segmentation field using a MAP-MRF algorithm. The method proposed by Chang et al. adopts a two-frame approach and does not use the constraint from the intensity

segmentation field during the video segmentation process. Although the algorithm has successfully identified multiple moving objects in the scene, the object boundaries are inaccurate in their experimental results. The method of Patras et al. employs an initial intensity segmentation and adopts a three-frame approach to deal with occlusions. However, the method retains the disadvantage of region merging approaches. The boundary information neglected by the initial intensity segmentation field could no longer be recovered by the motion vector field, and the temporal information could not act on the spatial information. In order to overcome the above problems, the proposed algorithm simultaneously estimates the three fields to form spatio-temporally coherent results. The interrelationships among the three fields and successive video frames are described by a Bayesian network model, in which spatial information and temporal information interact on each other. In our approach, regions in the intensity segmentation can either merge or split according to the motion information. Hence boundary information lost in the intensity segmentation field can be recovered by the motion vector field.

The rest of the chapter is arranged as follows: Section 3.2 presents the formulation of our approach. Section 3.3 proposes the optimization scheme. Section 3.4 discusses the experimental results.

3.2 Method

3.2.1 Model representation

For an image sequence, assume that the intensity remains constant along a motion trajectory. Ignoring both illumination variations and occlusions, it may be stated as

$$y_k(\mathbf{x}) = y_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})), \tag{3.1}$$

where $y_k(\mathbf{x})$ is the pixel intensity within the kth video frame at site $\mathbf{x}$, with $k \in \mathbf{N}$, $\mathbf{x} \in \mathbf{X}$, and $\mathbf{X}$ is the spatial domain of each video frame. $\mathbf{d}_k(\mathbf{x})$ is the motion vector from frame k–1 to frame k. The entire motion vector field is expressed compactly as $\mathbf{d}_k$.

Since the video data is contaminated with certain level of noise in the image acquisition process, an observation model is required for the sequence. Assume that independent and identically distributed (i. i. d.) Gaussian noise corrupts each pixel, thus the observation model for the kth frame becomes

$$g_k(\mathbf{x}) = y_k(\mathbf{x}) + n_k(\mathbf{x}), \tag{3.2}$$

where $g_k(\mathbf{x})$ is the observed image intensity at site $\mathbf{x}$, and $n_k(\mathbf{x})$ is the independent zero-mean additive noise with variance σ_n^2.

In our work, video segmentation refers to grouping pixels that belong to independently moving objects in the scene. To deal with occlusions, we assume that each site $\mathbf{x}$ in the current frame g_k cannot be occluded in both the previous frame g_{k-1} and the next frame g_{k+1}. Thus a three-frame method is adopted for object segmentation. Given consecutive frames of the observed video sequence, g_{k-1}, g_k, and g_{k+1}, we wish to estimate the joint conditional probability distribution of the motion vector field $\mathbf{d}_k$, the intensity segmentation field s_k, and the object (or video) segmentation field z_k. Using the Bayes' rule, we know

$$p(\mathbf{d}_k, s_k, z_k \mid g_k, g_{k-1}, g_{k+1})$$

$$= \frac{p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1})}{p(g_k, g_{k-1}, g_{k+1})}, \tag{3.3}$$

where $p(\mathbf{d}_k, s_k, z_k \mid g_k, g_{k-1}, g_{k+1})$ is the posterior probability density function (pdf) of the three fields, and the denominator on the right side is constant with respect to the unknowns.

The interrelationships among $\mathbf{d}_k$, s_k, z_k, g_k, g_{k-1}, g_{k+1} are modeled using the Bayesian network shown in Figure 3.1. Motion estimation establishes the pixel correspondence among the three consecutive frames. The intensity segmentation field provides a set of regions with relatively small intensity variation in the current frame. In order to identify independently moving objects in the scene, these regions are encouraged to group into segments with coherent motion. Meanwhile, if multiple motion models coexist within one region, the region may split into several segments. Thus according to the motion vector field, regions in the intensity segmentation field can either merge or split to form spatio-temporally coherent segments. Moreover, spatial connectivity should be encouraged during the video segmentation process.

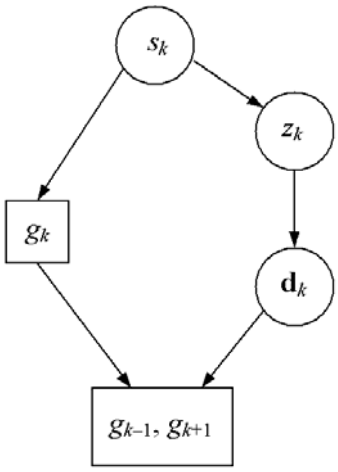


Figure 3.1 Bayesian network model for video segmentation.

The conditional independence relationships implied by the Bayesian network allow us to represent the joint distribution more compactly. Using the chain rule [30], the joint probability density can be factorized as

$$p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1})$$

$$= p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k) p(g_k \mid s_k) p(s_k) p(\mathbf{d}_k \mid z_k) p(z_k \mid s_k) . \quad (3.4)$$

Hence, the maximum a posteriori (MAP) estimate of the three fields becomes

$$(\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) = \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(\mathbf{d}_k, s_k, z_k \mid g_k, g_{k-1}, g_{k+1})$$

$$= \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1})$$

$$= \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k) p(g_k \mid s_k) p(s_k) p(\mathbf{d}_k \mid z_k) p(z_k \mid s_k) . \quad (3.5)$$

3.2.2 Spatio-temporal constraints

The conditional probability density $p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k)$ shows how well the motion estimation fits the given consecutive frames. Assuming that the probability is completely specified by the random field of displaced frame difference (DFD) [61], the video observation model can be employed to compute $p(g_{k-1}, g_{k+1} \mid \mathbf{d}_k, g_k)$. We can define the backward DFD $e_k^b(\mathbf{x})$ and forward DFD $e_k^f(\mathbf{x})$ at site $\mathbf{x}$ as

$$e_k^b(\mathbf{x}) = g_k(\mathbf{x}) - g_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x}))$$

$$= n_k(\mathbf{x}) - n_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})) , \quad (3.6a)$$

$$e_k^f(\mathbf{x}) = g_k(\mathbf{x}) - g_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x}))$$

$$= n_k(\mathbf{x}) - n_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x})) . \quad (3.6b)$$

The vector $(e_k^b(\mathbf{x}), e_k^f(\mathbf{x}))^T$ is denoted as $\mathbf{e}_k(\mathbf{x})$. With the i. i. d. Gaussian noise assumption, we know that $\mathbf{e}_k(\mathbf{x})$ is of zero mean bivariate normal distribution. The correlation coefficient of $e_k^b(\mathbf{x})$ and $e_k^f(\mathbf{x})$ is

$$\rho = \frac{\mathrm{Cov}[e_k^b(\mathbf{x}), e_k^f(\mathbf{x})]}{\sqrt{\mathrm{Var}[e_k^b(\mathbf{x})]\mathrm{Var}[e_k^f(\mathbf{x})]}} = \frac{\sigma_n^2}{2\sigma_n^2} = \frac{1}{2}. \tag{3.7}$$

Assuming conditional independence among spatially distinct observations, the probability density can be factorized as

$$p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k)$$

$$\approx \prod_{\mathbf{x}\in\mathbf{X}} p(g_{k-1}(\mathbf{x}-\mathbf{d}_k(\mathbf{x})), g_{k+1}(\mathbf{x}+\mathbf{d}_k(\mathbf{x})) \mid g_k(\mathbf{x}))$$

$$= \prod_{\mathbf{x}\in\mathbf{X}} p(e_k^b(\mathbf{x}), e_k^f(\mathbf{x}))$$

$$= \left(\frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_\mathbf{e}|}}\right)^{|\mathbf{X}|} \exp\left[-\sum_{\mathbf{x}\in\mathbf{X}} \frac{1}{2}\mathbf{e}_k^T(\mathbf{x})\boldsymbol{\Sigma}_\mathbf{e}^{-1}\mathbf{e}_k(\mathbf{x})\right]$$

$$\propto \exp\left[-\sum_{\mathbf{x}\in\mathbf{X}} U_\mathbf{x}^{g|\mathbf{d}}(\mathbf{d}_k(\mathbf{x}))\right], \tag{3.8a}$$

$$U_\mathbf{x}^{g|\mathbf{d}}(\mathbf{d}_k(\mathbf{x})) = (e_k^b(\mathbf{x}))^2 - 2\rho e_k^b(\mathbf{x}) e_k^f(\mathbf{x}) + (e_k^f(\mathbf{x}))^2, \tag{3.8b}$$

where $\boldsymbol{\Sigma}_\mathbf{e}$ is the covariance matrix for each site $\mathbf{x}$, and the correlation coefficient ρ has been computed in (3.7).

The term $p(g_k \mid s_k)$ shows how well the intensity segmentation fits the scene. Assuming Gaussian distribution for each segmented region in the current frame, the conditional probability density could be factorized as

$$p(g_k \mid s_k) = \prod_{\mathbf{x}\in\mathbf{X}} p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}))$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma_\eta}\right)^{|\mathbf{X}|} \exp\left[-\sum_{\mathbf{x}\in\mathbf{X}} \frac{1}{2\sigma_\eta^2}(g_k(\mathbf{x}) - \mu_{s_k(\mathbf{x})})^2\right]$$

$$\propto \exp[-\sum_{\mathbf{x}\in\mathbf{X}} U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x}))], \tag{3.9a}$$

$$U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x})) = (g_k(\mathbf{x}) - \mu_{s_k(\mathbf{x})})^2, \tag{3.9b}$$

where $s_k(\mathbf{x}) = l$ assigns site $\mathbf{x}$ to region l, μ_l is the intensity mean within region l, and σ_η^2 is the variance for each region.

The pdf $p(s_k)$ represents the prior probability of the intensity segmentation. To encourage the formation of continuous regions, we model the density $p(s_k)$ by a Markov random field [18]. That is, if $N_{\mathbf{x}}$ is the neighborhood of a pixel at $\mathbf{x}$, then the conditional distribution of a single variable at site $\mathbf{x}$ depends only on the variables within its neighborhood $N_{\mathbf{x}}$. According to the Hammersley-Clifford theorem, the density is given by a Gibbs distribution with the following form.

$$p(s_k) \propto \exp[-\sum_{c\in C} V_c^s(s_k(\mathbf{x}) \mid \mathbf{x} \in c)], \tag{3.10}$$

where C is the set of all cliques c, and V_c^s is the clique potential function. A clique is a set of pixels that are neighbors of each other, and the potential function V_c^s depends only on the points within clique c.

Spatial constraint can be imposed by the following two-pixel clique potential.

$$V_c^s(s_k(\mathbf{x}), s_k(\mathbf{y}))$$

$$\propto U_{\mathbf{x},\mathbf{y}}^s(s_k(\mathbf{x}), s_k(\mathbf{y}))$$

$$= \frac{1}{\| \mathbf{x} - \mathbf{y} \|^2}[1 - \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))], \tag{3.11}$$

where $\delta(x) = \begin{cases} 1, \text{if } x = 0 \\ 0, \text{otherwise} \end{cases}$ is the Kronecker delta function, and $\|\cdot\|$ denotes the Euclidean distance. Thus two neighboring pixels are more likely to belong to the same class than to different classes. The constraint becomes strong with the decrease of the distance between the neighboring sites.

The term $p(\mathbf{d}_k \mid z_k)$ is the conditional probability density of the motion vector field given the video segmentation field. To boost spatial connectivity, it is modeled by a Gibbs distribution with the following potential function.

$$V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) \mid z_k)$$

$$\propto U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y}))$$

$$= \frac{1}{\| \mathbf{x} - \mathbf{y} \|^2} \delta(z_k(\mathbf{x}) - z_k(\mathbf{y})) \| \mathbf{d}_k(\mathbf{x}) - \mathbf{d}_k(\mathbf{y}) \|^2 . \tag{3.12}$$

The pairwise smoothness constraint of the motion vectors is imposed only when the two neighboring pixels share the same video segmentation label. It encourages one region to split into several segments when different motion models coexist. Hence $U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}$ can viewed as the region splitting force.

The last term $p(z_k \mid s_k)$ represents the posterior probability density of the video segmentation field when the intensity segmentation field is given. The density is modeled by a Gibbs distribution with the following potential function.

$$V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) \mid s_k)$$

$$\propto U_{\mathbf{x},\mathbf{y}}^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y}))$$

$$= \frac{1}{\| \mathbf{x} - \mathbf{y} \|^2} [1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))] +$$

$$\frac{\alpha}{\| \mathbf{x} - \mathbf{y} \|^2} \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))[1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))]. \quad (3.13)$$

The first term on the right side encourages the spatial connectivity of video segmentation, while the second term encourages two neighboring pixels to share the same video segmentation label when they are within one region of the intensity segmentation field. Therefore $U_{\mathbf{x},\mathbf{y}}^{z|s}$ encourages intensity segmentation regions to group altogether and can be viewed as the region merging force. The parameter α controls the strength of the constraint imposed by intensity segmentation.

The interactions in the Bayesian network are modeled by the above spatio-temporal constraints. Combining these pdf terms, the MAP estimation criterion becomes

$$(\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) = \arg \min_{(\mathbf{d}_k, s_k, z_k)} [\sum_{\mathbf{x} \in \mathbf{X}} U_{\mathbf{x}}^{g|\mathbf{d}}(\mathbf{d}_k(\mathbf{x})) + \lambda_1 \sum_{\mathbf{x} \in \mathbf{X}} U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x})) +$$

$$\lambda_2 \sum_{\{\mathbf{x},\mathbf{y}\} \in C} U_{\mathbf{x},\mathbf{y}}^{s}(s_k(\mathbf{x}), s_k(\mathbf{y})) + \lambda_3 \sum_{\{\mathbf{x},\mathbf{y}\} \in C} U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y})) +$$

$$\lambda_4 \sum_{\{\mathbf{x},\mathbf{y}\} \in C} U_{\mathbf{x},\mathbf{y}}^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y}))], \quad (3.14)$$

where the parameters λ_1, λ_2, λ_3, and λ_4 control the contribution of individual terms.

3.2.3 Notes on the Bayesian network model

In our model, the video segmentation is affected by both spatial information and temporal information. It should be noted that the direction of the links in the Bayesian network model does not mean that the influence between the cause and consequence is just one-way.

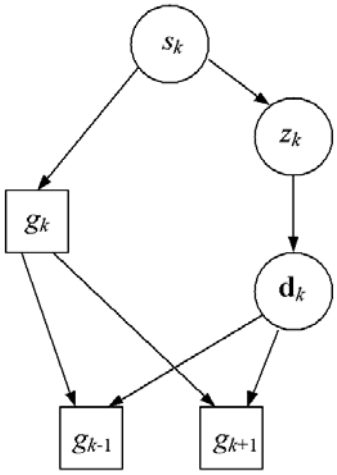


Figure 3.2 Simplified Bayesian network model for video segmentation.

The current video frame could be thought as the cause of the next frame. For an image sequence, both the original sequence and the one in the reverse sequence order are understandable from the viewpoint of segmentation. Thus, the current frame could also be viewed as the cause of the previous frame (in the reversed sequence). In our model, g_k is the cause of both the next frame g_{k+1} and the previous frame g_{k-1}.

The motion vector field establishes the correspondence between the current frame and its two neighboring frames. When frame g_{k+1} and frame g_{k-1} are separated (as shown in Figure 3.2), the interrelationship seems clearer at the first glance. However, from the structure of the Bayesian network, we know that in this case,

$$
\begin{aligned}
& p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k) \\
& = p(g_{k-1} \mid g_k, \mathbf{d}_k) p(g_{k+1} \mid g_k, \mathbf{d}_k) \\
& = \prod_{\mathbf{x} \in \mathbf{X}} p(e_k^b(\mathbf{x})) p(e_k^f(\mathbf{x})) \\
& \propto \exp[-\sum_{\mathbf{x} \in \mathbf{X}} (e_k^b(\mathbf{x}))^2 + (e_k^f(\mathbf{x}))^2]. \qquad (3.15)
\end{aligned}
$$

Comparing with (3.8), the correlation coefficient of $e_k^b(\mathbf{x})$ and $e_k^f(\mathbf{x})$ is zero in (3.15). The Bayesian network in Figure 3.2 neglects the interaction between the forward DFD and the backward DFD. Therefore, the Bayesian network model in Figure 3.2 is just a simplification of the original model.

In (3.13), when the parameter α becomes zero, the constraint from the intensity segmentation disappears so that our method degenerates into motion based approach. Meanwhile, when α becomes infinity, boundaries in the video segmentation field must come from the intensity segmentation field, and our technique turns into region merging approach. Therefore, the interactive approach can be viewed as a compromise between motion based approach and region merging approach.

3.3 MAP estimation

3.3.1 Iterative estimation

Obviously, there is no simple method of directly minimizing (3.14) with respect to all unknowns. We propose an optimization strategy iterating over the following two steps.

Firstly, we update $\mathbf{d}_k$ and s_k given the estimate of the video segmentation field z_k. From the structure of the proposed Bayesian network, we can see that $\mathbf{d}_k$ and s_k are conditionally independent when video segmentation field z_k and the three successive frames are given. The joint estimation can be factorized as

$$(\hat{\mathbf{d}}_k, \hat{s}_k) = \arg\max_{(\mathbf{d}_k, s_k)} p(\mathbf{d}_k, s_k \mid g_k, g_{k-1}, g_{k+1}, \hat{z}_k)$$

$$= (\arg\max_{\mathbf{d}_k} p(\mathbf{d}_k \mid g_k, g_{k-1}, g_{k+1}, \hat{z}_k), \arg\max_{s_k} p(s_k \mid g_k, \hat{z}_k)). \quad (3.16)$$

Using the chain rule, the MAP estimate becomes

$$\hat{\mathbf{d}}_k = \arg\max_{\mathbf{d}_k} p(\mathbf{d}_k \mid g_k, g_{k-1}, g_{k+1}, \hat{z}_k)$$

$$= \arg\max_{\mathbf{d}_k} p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k) p(\mathbf{d}_k \mid \hat{z}_k), \tag{3.17a}$$

$$\hat{s}_k = \arg\max_{s_k} p(s_k \mid g_k, \hat{z}_k)$$

$$= \arg\max_{s_k} p(g_k \mid s_k) p(s_k) p(\hat{z}_k \mid s_k). \tag{3.17b}$$

Secondly, we update z_k given the estimate of the motion field $\mathbf{d}_k$ and the intensity segmentation field s_k.

$$\hat{z}_k = \arg\max_{z_k} p(z_k \mid g_k, g_{k-1}, g_{k+1}, \hat{\mathbf{d}}_k, \hat{s}_k)$$

$$= \arg\max_{z_k} p(z_k \mid \hat{\mathbf{d}}_k, \hat{s}_k)$$

$$= \arg\max_{z_k} p(\hat{\mathbf{d}}_k \mid z_k) p(z_k \mid \hat{s}_k). \tag{3.18}$$

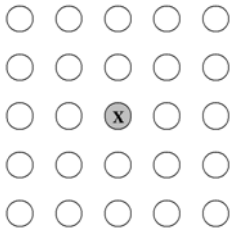


Figure 3.3 The 24-pixel neighborhood.

In our work, the 24-point neighborhood system (the fifth order neighbor system, see Figure 3.3) is used, and potentials are defined only on two-point cliques. Using the terms in (3.14), the Bayesian MAP estimates in (3.17) and (3.18) can be obtained by minimizing the following objective functions.

$$F^{\mathbf{d}}(\mathbf{d}_k) = \sum_{\mathbf{x}\in\mathbf{X}}[U_{\mathbf{x}}^{g|\mathbf{d}}(\mathbf{d}_k(\mathbf{x})) + \frac{1}{2}\lambda_3 \sum_{\mathbf{y}\in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), \hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y}))], \quad (3.19a)$$

$$F^{s}(s_k) = \sum_{\mathbf{x}\in\mathbf{X}}[\lambda_1 U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x})) + \frac{1}{2}\lambda_2 \sum_{\mathbf{y}\in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^{s}(s_k(\mathbf{x}), s_k(\mathbf{y})) +$$

$$\frac{1}{2}\lambda_4 \sum_{\mathbf{y}\in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^{z|s}(\hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y}))], \quad (3.19b)$$

$$F^{z}(z_k) = \sum_{\mathbf{x}\in\mathbf{X}}[\frac{1}{2}\lambda_3 \sum_{\mathbf{y}\in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}(\hat{\mathbf{d}}_k(\mathbf{x}), \hat{\mathbf{d}}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y})) +$$

$$\frac{1}{2}\lambda_4 \sum_{\mathbf{y}\in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}), \hat{s}_k(\mathbf{x}), \hat{s}_k(\mathbf{y}))], \quad (3.19c)$$

where $N_{\mathbf{x}}$ is the neighborhood of the pixel at $\mathbf{x}$.

3.3.2 Local optimization

In general, the objective functions are nonconvex and do not have a unique minimum. The iterated conditional modes (ICM) algorithm is used to arrive at a sub-optimal estimate of each objective function [4]. The ICM algorithm employs the greedy strategy in iterative minimization. Given the observed data and other estimated labels, the segmentation label is sequentially updated by locally minimizing the objective function at each site.

To effectively employ boundary hints supplied by spatial information in the local optimization, distance transformation [5] is performed on the intensity segmentation field. Each pixel $\mathbf{x}$ in the distance transformed image has a value $d_{\mathbf{x}}(s_k)$ representing the distance between the pixel and the nearest boundary pixel in s_k. Here a boundary

pixel **x** has at least one point **y** within its neighborhood where $s_k(\mathbf{y})$ is not the same as $s_k(\mathbf{x})$. The term $U_{\mathbf{x},\mathbf{y}}^{z|s}$ in (3.19c) is replaced by

$$U'^{z|s}_{\mathbf{x},\mathbf{y}}(z_k(\mathbf{x}), z_k(\mathbf{y}), \hat{s}_k(\mathbf{x}), \hat{s}_k(\mathbf{y})) = \frac{1}{\| \mathbf{x}-\mathbf{y} \|^2}[1-\delta(z_k(\mathbf{x})-z_k(\mathbf{y}))] +$$

$$\frac{\alpha\theta(d_{\mathbf{x}}(\hat{s}_k)-d_{\mathbf{y}}(\hat{s}_k))}{\| \mathbf{x}-\mathbf{y} \|^2}\delta(\hat{s}_k(\mathbf{x})-\hat{s}_k(\mathbf{y}))[1-\delta(z_k(\mathbf{x})-z_k(\mathbf{y}))], \qquad (3.20)$$

where $\theta(x) = \begin{cases} 2, \text{if } x < 0 \\ 1, \text{if } x = 0 \\ 0, \text{otherwise} \end{cases}$. The term θ helps to give a penalty on the pixel closer to the boundary in the intensity segmentation field if the two neighboring pixels within an intensity segmentation region do not share the same video segmentation label. It should be noted that $U'^{z|s}_{\mathbf{x},\mathbf{y}}$ does not destroy the symmetry of the two-pixel clique potential in MRF [69]. $U'^{z|s}_{\mathbf{x},\mathbf{y}}$ is associated with the objective function (3.19c) and the optimization algorithm. The optimization algorithm updates the label by locally minimizing the objective function at each site. A two-point potential is accounted on both sites. $U'^{z|s}_{\mathbf{x},\mathbf{y}}$ is equivalent to $U^{z|s}_{\mathbf{x},\mathbf{y}}$ for the objective function because the total penalty for the entire field is the same. $U^{z|s}_{\mathbf{x},\mathbf{y}}$ is symmetric and it complies with the definition of MRF. The difference between them occurs in the local minimization of the optimization process. We prefer the form of (3.20) since in our experiments, the boundary information are more accurately estimated by giving all the penalty to the site near the boundary instead of evenly allocating the penalty for both sites in local optimization (see Section 3.4).

Similarly in (3.19b), $U^{z|s}_{\mathbf{x},\mathbf{y}}$ could be replaced by

$$U''^{z|s}_{\mathbf{x},\mathbf{y}}(\hat{z}_k(\mathbf{x}),\hat{z}_k(\mathbf{y}),s_k(\mathbf{x}),s_k(\mathbf{y}))$$

$$=\frac{\alpha\theta(d_{\mathbf{x}}(\hat{z}_k)-d_{\mathbf{y}}(\hat{z}_k))}{\|\mathbf{x}-\mathbf{y}\|^2}\,\delta(s_k(\mathbf{x})-s_k(\mathbf{y}))[1-\delta(\hat{z}_k(\mathbf{x})-\hat{z}_k(\mathbf{y}))]. \tag{3.21}$$

Comparing with (3.13), the first term in (3.13) is ignored in (3.21) since it is constant when the video segmentation field is given.

Thus, we obtain the actual local objective functions that are sequentially optimized at each site.

$$F^{\mathbf{d}}_{\mathbf{x}}(\mathbf{d}_k)=U^{g|\mathbf{d}}_{\mathbf{x}}(\mathbf{d}_k(\mathbf{x}))+\frac{1}{2}\lambda_3\sum_{\mathbf{y}\in N_{\mathbf{x}}}U^{\mathbf{d}|z}_{\mathbf{x},\mathbf{y}}(\mathbf{d}_k(\mathbf{x}),\mathbf{d}_k(\mathbf{y}),\hat{z}_k(\mathbf{x}),\hat{z}_k(\mathbf{y}))\,, \tag{3.22a}$$

$$F^{s}_{\mathbf{x}}(s_k)=\lambda_1 U^{g|s}_{\mathbf{x}}(s_k(\mathbf{x}))+\frac{1}{2}\lambda_2\sum_{\mathbf{y}\in N_{\mathbf{x}}}U^{s}_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}),s_k(\mathbf{y}))+$$

$$\frac{1}{2}\lambda_4\sum_{\mathbf{y}\in N_{\mathbf{x}}}U''^{z|s}_{\mathbf{x},\mathbf{y}}(\hat{z}_k(\mathbf{x}),\hat{z}_k(\mathbf{y}),s_k(\mathbf{x}),s_k(\mathbf{y}))\,, \tag{3.22b}$$

$$F^{z}_{\mathbf{x}}(z_k)=\frac{1}{2}\lambda_3\sum_{\mathbf{y}\in N_{\mathbf{x}}}U^{\mathbf{d}|z}_{\mathbf{x},\mathbf{y}}(\hat{\mathbf{d}}_k(\mathbf{x}),\hat{\mathbf{d}}_k(\mathbf{y}),z_k(\mathbf{x}),z_k(\mathbf{y}))+$$

$$\frac{1}{2}\lambda_4\sum_{\mathbf{y}\in N_{\mathbf{x}}}U'^{z|s}_{\mathbf{x},\mathbf{y}}(z_k(\mathbf{x}),z_k(\mathbf{y}),\hat{s}_k(\mathbf{x}),\hat{s}_k(\mathbf{y}))\,. \tag{22c}$$

3.3.3 Initialization and parameters

The intensity segmentation field is initialized using a generalized K-means clustering algorithm to include the spatial constraint. Each cluster is characterized by a constant intensity, and the spatial constraints are performed by the two-point clique potential in (3.11). The initialization algorithm is actually a simplification of the adaptive clustering algorithm proposed by Papps [44]. The initial motion vector field is

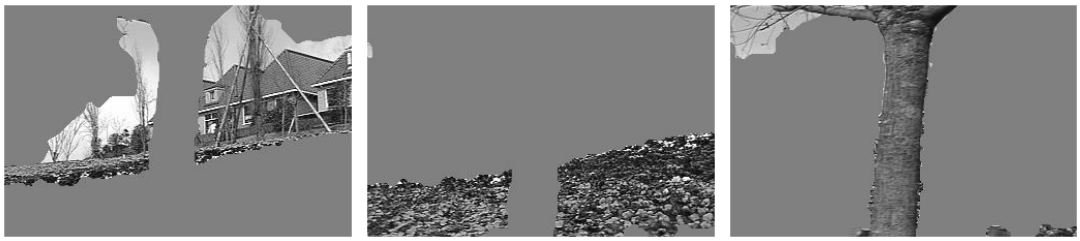
obtained by the MAP estimation with pairwise smoothness constraint [61]. Wang and Adelson [66] have proposed a procedure for initialization of the video segmentation field. Given the initial motion estimates, the current frame is divided into small blocks and an affine transformation is computed for each block. A set of motion models is known by adaptively clustering the affine parameters. Then video segmentation labels are assigned in a way that minimizes the motion distortion. In our work, the video segmentation field is initialized by combining this procedure with the spatial constraint on the assignment of regions. The parameter α is manually determined to control the constraint imposed by intensity segmentation. Given the initial estimates of the three fields, we employ the idea of parameter selection proposed by Chang et al. [9]. The parameters (λ_1, λ_2, λ_3, and λ_4) are determined by equalizing the contributions of the terms in (3.14). Details can be found in the references.

3.4 Results and discussion

The results tested on the "flower garden" sequence and the "table tennis" sequence are shown in Figure 3.4-5. We assume that there are four objects in the video segmentation field.

(a) (b) (c)

(d) (e) (f)



(g) (h) (i)

(j) (k) (l)

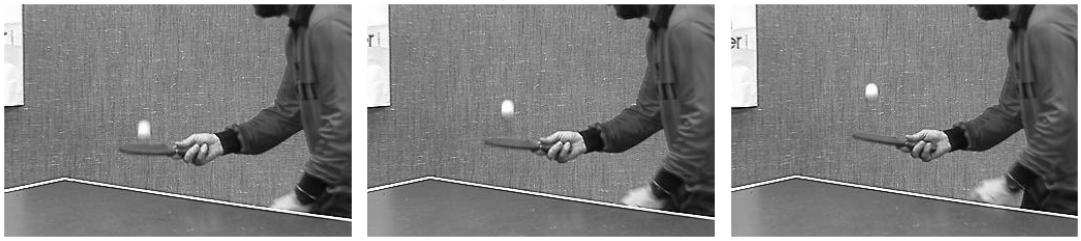
(m) (n) (o)

Figure 3.4 Segmentation results of the "flower garden" sequence. (a)-(c) Three consecutive frames of the sequence. (d) The motion vector field. (e) The four-level intensity segmentation field, (f) the corresponding distance transformed image and

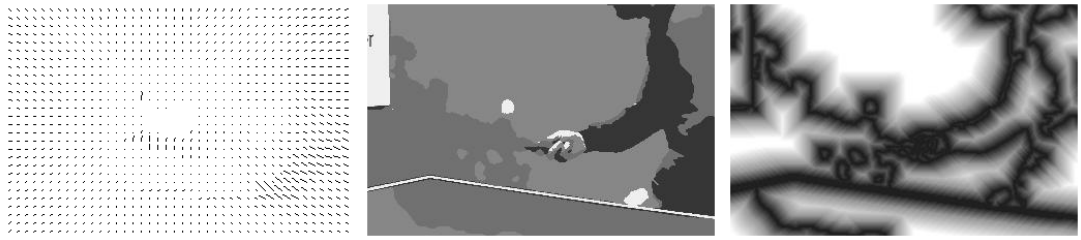
(g)-(j) video segmentation results. (k) The three-level intensity segmentation field and (l)-(o) the corresponding video segmentation results.

The motion vector field, intensity segmentation field, and the video segmentation field are recovered using the proposed technique for both sequences. The spatial connectivity is clearly exhibited in the estimation results. From the motion vector fields shown in Figure 3.4d and 3.5d, we can see that motion occlusions are successfully overcome. The results of the four-level intensity segmentation are depicted in Figure 3.4e and 3.5e, where an area with constant intensity represents an intensity segment. Figure 3.4f and 3.5f are the corresponding distance transformed images. Darker gray levels are used to represent the pixels with smaller distance values. In Figure 3.4g-j and 3.5g-j, we represent the video segmentation results obtained by our approach. In the "flower garden" sequence, the edge information is preserved well in intensity segmentation field (see Figure 3.4e). The algorithm is capable of distinguishing the different objects in the scene by successfully grouping the small regions that are spatio-temporally coherent. While in the "table tennis" sequence, the boundary information lost in Figure 3.5e (boundary information may be lost even in an oversegmentation, e.g., the boundary between the body and the left arm) is recovered according to the information from the motion vector field. However, boundaries are detected more accurately when both spatial and temporal features are matched (e.g., the tree in Figure 3.4i and the body in Figure 3.5g). The segmentation algorithm is robust even at the largely homogeneous areas (e.g., the sky in Figure 3.4j and table in Figure 3.5j), where there is little motion information. Figure 3.4l-o and 3.5l-o show the video segmentation results with three-level and six-level intensity segmentation for the "flower garden" sequence and the "table tennis" sequence respectively. Comparing with the video segmentation results shown

in Figure 3.4g-j and 3.5g-j, it can be seen that our method is robust to achieve spatio-temporally coherent results without strong requirement of intensity segmentation.

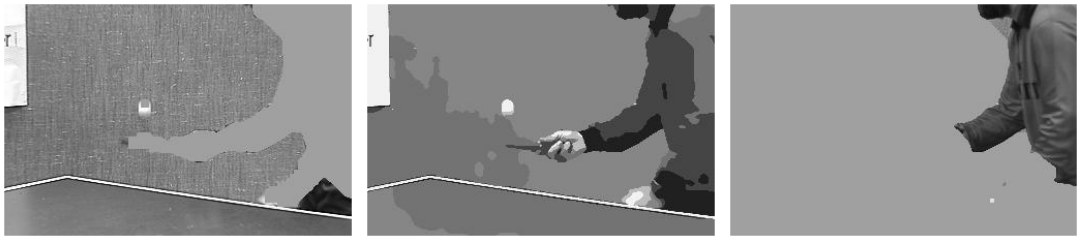


(a) (b) (c)

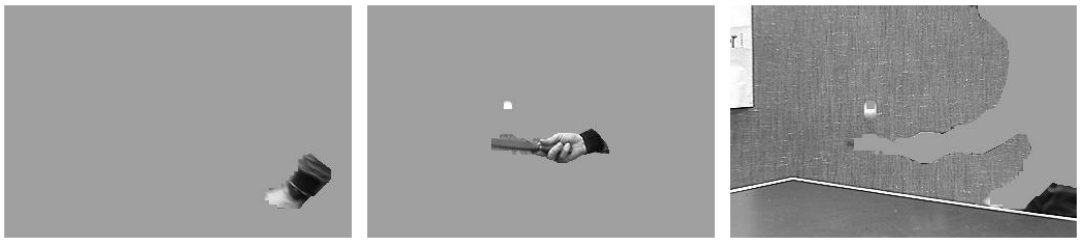


(d) (e) (f)

(g) (h) (i)



(j) (k) (l)



(m) (n) (o)

Figure 3.5 Segmentation results of the "table tennis" sequence. (a)-(c) Three consecutive frames of the sequence. (d) The motion vector field. (e) The four-level intensity segmentation field, (f) the corresponding distance transformed image and (g)-(j) video segmentation results. (k) The six-level intensity segmentation field and (l)-(o) the corresponding video segmentation results.

Figure 3.6 shows part of the video segmentation results using (3.13) in local objective functions instead of (3.21) and (3.22) for the two sequences. Comparing with the segmented results in Figure 3.4 and 3.5, it can be seen that the utilization of distance transformation in local optimization has greatly improves the boundary accuracy of video segmentation.

(a) (b) (c)

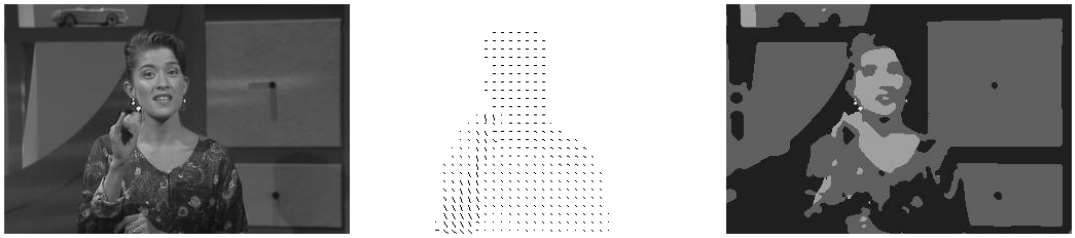
Figure 3.6 The video segmentation results without using distance transformation in local optimization for (a) the "flower garden" sequence and (b) (c) the "table tennis" sequence.

To test the robustness of the algorithm, Figure 3.7-8 show the video segmentation results by the proposed method for the “coastguard” sequence and the “sign” sequence, respectively. In Figure 3.7-8, it is assumed that there are three objects in the scene. The motion vector field and the intensity segmentation field for the “sign” sequence are also shown in Figure 3.8. The experimental results exhibit satisfactory spatio-temporal coherence.

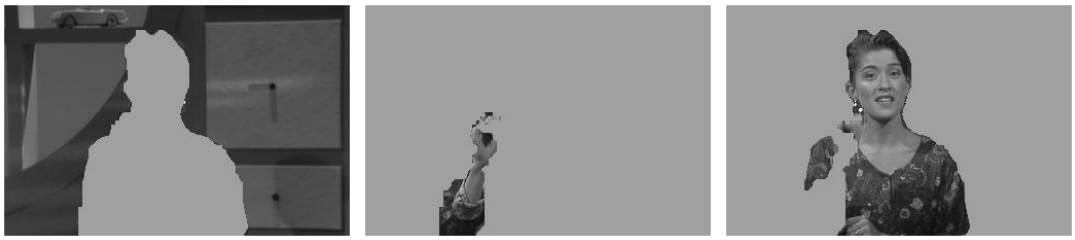
(a) (b)

(c) (d)

Figure 3.7 Segmentation results of the “coastguard” sequence. (a) One frame of the sequence. (b)-(d) The video segmentation results.



(a) (b) (c)



(d) (e) (f)

Figure 3.8 Segmentation results of the "sign" sequence. (a) One frame of the sequence. (b) The motion vector field. (c) The four-level intensity segmentation field. (d)-(f) The video segmentation results.

The intensity segmentation constraint helps generate accurate boundaries in spatio-temporally coherent areas. Since sometimes one area of similar intensity may belong to different objects, the intensity segmentation constraint is weakened when the motion information within one intensity segmentation region is incoherent. This is why boundaries lost in the intensity segmentation can be recovered by the motion information in our work. As a compromise, the boundary is not anticipated to be accurate in the incoherent area because the intensity segmentation constraint is weak there. Moreover, the incoherence of spatio-temporal information may be caused by the boundary information loss in the intensity segmentation field or the estimation error in the motion vector field. In the worst case, our algorithm will fail in the area where the boundary is lost in intensity segmentation, and the motion is erroneously estimated at the same time (e.g. the segmentation error for the part of the right hand in Figure 3.8e). Hence our approach may not consistently produce accurate edges in the entire video segmentation field. However, the approach has an advantage in application areas where it is important to discover areas with different motions (such

as in human machine interaction and video indexing). Therefore, the new approach is complementary to region merging methods in this aspect.

In this chapter, we have proposed a unified framework for video segmentation based on graphical models. The spatio-temporal consistency of segmentation is expressed in terms of interactions among the motion field, the intensity segmentation field, and the video segmentation field. The solution is obtained by the MAP estimate, and an optimization procedure that iteratively maximizes the conditional probability density of the three fields is proposed. There are three main contributions within the chapter. The first is building a Bayesian network based framework that combines both the spatial and temporal information in the video segmentation process. The second is formulating the spatio-temporal constraints by utilizing Markov random fields, distance transformation, and multivariate normal distribution. The third is theoretically making a compromise between motion based approach and region merging approach. The approach deals with video segmentation from a relatively comprehensive and general viewpoint, and thus can be universally applied. Our method exhibits good robustness and spatio-temporal coherence.

Chapter 4

A Dynamic Hidden Markov Random Field Model for Foreground Segmentation

4.1 Introduction

A probabilistic model of spatial and temporal constraints in video sequences, the dynamic hidden Markov random field (DHMRF) model, is proposed in this chapter for segmenting indoor foreground objects by background subtraction and shadow removal. Spatial and temporal dependencies in the segmentation process are unified in the dynamic probabilistic model (DHMRF) that combines the Markov random field (MRF) and the hidden Markov model (HMM). A computationally efficient approximate filtering algorithm is derived for the DHMRF model to recursively estimate the segmentation field. Each pixel in the scene is classified as foreground, shadow, or background from the history of video images. The foreground segmentation method integrates both intensity and edge features, and it adaptively updates the models of background, shadow, and edge information. Experimental results show that the proposed approach robustly handles shadow and camouflage in nonstationary background scenes and improves the accuracy of foreground detection in monocular video sequences.

As to the related work, Paragios and Ramesh use the MRF model to combine different types of features and incorporate spatial constraints for subway monitoring [45]. The method detects changes between the background and the current image, and it does not utilize the previous images. In our approach, the foreground is

estimated from the history of all observed images. Rittscher et al. use both HMM and MRF for foreground and shadow segmentation [52]. In their work, each site (or block) is modeled by a single HMM independent of the neighboring sites (or blocks). The HMM and the MRF are employed in two different processes to impose temporal and spatial contextual constraints respectively. In our work, the state of a single site is influenced by its neighboring sites, meanwhile spatial and temporal constraints are unified in a dynamic model. Mikic et al. model the pixel color change under shadow by a diagonal matrix for traffic scenes [40]. In their approach the variance under shadow is assumed to be smaller than the variance in the background for the same site, which sometimes is not valid for indoor environments. For the background updating process, the Gaussian mixture method by Stauffer and Grimson [57] is slightly modified in our work to employ the estimation by the DHMRF filtering algorithm.

The rest of the chapter is arranged as follows: Section 4.2 proposes the DHMRF model and derives its filtering algorithm. Section 4.3 presents the foreground and shadow detection method. Section 4.4 describes the implementation details. Section 4.5 discusses the experimental results.

4.2 Dynamic hidden Markov random field

Given an image sequence $\{g_k\}$, the segmentation label for a point $\mathbf{x}$ within the kth image is denoted by $s_k(\mathbf{x})$. Label $s_k(\mathbf{x}) \in \{1, 2, \ldots, L\}$ assigns the point $\mathbf{x}$ to one of L (L equals 3 in this work, see Section 4.3) classes at time k. Here $k \in \mathbf{N}$, $\mathbf{x} \in \mathbf{X}$, and $\mathbf{X}$ is the spatial domain of the video scene. The entire label field is expressed compactly as s_k. Spatial and temporal constraints in the segmentation process can be imposed through a dynamic model of statistical dependencies of neighboring sites.

4.2.1 DHMRF model

Given the observed data up to time k, the posterior probability distribution of the segmentation field s_k is modeled by a Markov random field [18] to formulate spatial dependencies. In the MRF model, if $N_{\mathbf{x}}$ is the neighborhood of a site $\mathbf{x}$, then the conditional distribution of a single label at $\mathbf{x}$ depends only on the labels within its neighborhood $N_{\mathbf{x}}$. According to the Hammersley-Clifford theorem, the probability is given by a Gibbs distribution that has the following form.

$$p(s_k \mid g_{1:k}) \propto \exp[-\sum_{c \in C} V_c(s_k(c) \mid g_{1:k})], \tag{4.1}$$

where $g_{1:k}$ denotes $\{g_1, g_2, \ldots, g_k\}$, C is the set of all cliques c, V_c is the clique potential function, and $s_k(c)$ denotes $\{s_k(\mathbf{x}) \mid \mathbf{x} \in c\}$. A clique is a set of pixels that are neighbors of each other, and the potential function V_c depends only on the points within clique c.

Only one-pixel and two-pixel cliques are used in our work. The one-pixel potential $V_{\mathbf{x}}(s_k(\mathbf{x}) \mid g_{1:k})$ reflects the information (or constraint) from the observation for a single site, and the two-pixel potential imposes the spatial constraint to form contiguous regions. To simplify the computation, the pairwise constraint is assumed to be independent of the observed images. Hence the two-point potential is written as $V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}), s_k(\mathbf{y}))$. The posterior distribution at time k becomes

$$p(s_k \mid g_{1:k}) \propto \exp\{-\sum_{\mathbf{x} \in \mathbf{X}} [V_{\mathbf{x}}(s_k(\mathbf{x}) \mid g_{1:k}) + \frac{1}{2} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}), s_k(\mathbf{y}))]\}. \tag{4.2}$$

Spatial connectivity constraint can be imposed by the following two-pixel potential.

$$V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}) = i, s_k(\mathbf{y}) = j) \propto \frac{1}{\| \mathbf{x} - \mathbf{y} \|^2} (1 - \delta(i - j)), \tag{4.3}$$

where $1 \le i, j \le L$, $\|\cdot\|$ denotes the Euclidian distance, and $\delta(\cdot)$ is the Kronecker delta function. Thus two neighboring pixels are more likely to belong to the same class than to different classes. The spatial constraint becomes strong with decreasing distance between the neighboring sites.

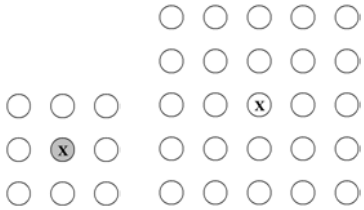
The dynamic or temporal dependencies of consecutive segmentation fields are formulated by a hidden Markov model [50]. In the HMM, image g_k is the kth observation, and segmentation field s_k is the hidden state at time k. Therefore the state transition model $p(s_{k+1} \mid s_k)$ and the observation (or likelihood) model $p(g_k \mid s_k)$ for the HMM should be built for the entire scene.

The label field state transition probability $p(s_{k+1} \mid s_k)$ is modeled by a Markov random field defined on one-pixel and two-pixel cliques as well.

$$p(s_{k+1} \mid s_k) \propto \exp[-\sum_{c \in C} V_c(s_{k+1}(c) \mid s_k)]$$

$$= \exp\{-\sum_{\mathbf{x} \in \mathbf{X}} [V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}})) + \frac{1}{2} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))]\}, \qquad (4.4)$$

where $M_{\mathbf{x}}$ designates the set of sites in the kth image that impact on site $\mathbf{x}$ in the (k+1)th image. The one-pixel potential $V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}}))$ models the label state transition for a single site, and the two-pixel potential $V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))$ imposes the pairwise spatial constraint. It should be noted that $M_{\mathbf{x}}$ is not equivalent to the neighborhood $N_{\mathbf{x}}$. $M_{\mathbf{x}}$ and $N_{\mathbf{x}}$ may have different sizes. $\mathbf{x} \notin N_{\mathbf{x}}$ while $\mathbf{x} \in M_{\mathbf{x}}$ (e.g. see Figure 4.1). To distinguish them, $N_{\mathbf{x}}$ is called the spatial neighborhood, and $M_{\mathbf{x}}$ the temporal neighborhood.



(a) (b)

Figure 4.1 Illustration of spatial neighborhood and temporal neighborhood. (a) The 8-pixel spatial neighborhood. (b) The 25-pixel temporal neighborhood.

Assuming conditional independence between spatially distinct observations, the observation model $p(g_k \mid s_k)$ is factorized as

$$p(g_k \mid s_k) = \prod_{\mathbf{x} \in \mathbf{X}} p(\mathbf{o}_k(\mathbf{x}) \mid s_k(\mathbf{x})), \tag{4.5}$$

where $\mathbf{o}_k(\mathbf{x})$ is the observation for site $\mathbf{x}$ that consists of locally measured information such as intensity and gradient features (see Section 4.3.2).

By (4.2), (4.4), and (4.5), spatial and temporal dependencies in the segmentation process are unified in a dynamic model that combines the MRF and the HMM. Therefore it is called the dynamic hidden Markov random field (DHMRF) model.

4.2.2 DHMRF filter

From a Bayesian perspective, the filtering algorithm is to recursively update the posterior distribution of the segmentation field. Given the potentials of the distribution $p(s_k \mid g_{1:k})$, the posterior $p(s_{k+1} \mid g_{1:k+1})$ at time k+1 can be efficiently approximated by a Markov random field with the following potential functions (see Appendix A).

$$V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) = i \mid g_{1:k+1})$$

$$= -\ln\{\sum_{j}\exp[-\alpha_{i|j} - \lambda_k \sum_{\mathbf{y}\in M_{\mathbf{x}}} \frac{1}{|M_{\mathbf{y}}|} V_{\mathbf{y}}(s_k(\mathbf{y}) = j \mid g_{1:k})]\} -$$

$$\ln p(\mathbf{o}_{k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x}) = i), \tag{4.6a}$$

$$V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}) = i, s_{k+1}(\mathbf{y}) = j) = \frac{\beta}{\|\mathbf{x} - \mathbf{y}\|^2}(1 - \delta(i - j)), \tag{4.6b}$$

where $1 \le i, j \le L$, $|\cdot|$ denotes the size (number of points) of the set, $\alpha_{i|j}$ is the potential of state transition (from j to i) that imposes the temporal continuity constraint on segmentation label, λ_k and β weight the constraint from previous observations and the constraint of spatial connectivity respectively. The parameters are initialized and determined in Section 4.4.2. In the one-pixel potential (4.6a), the first term reflects the information from previously observed images for a single site $\mathbf{x}$, which is affected by its temporal neighborhood $M_{\mathbf{x}}$. The second term in (4.6a) reflects the information from the current observation. The two-pixel potential (4.6b) imposes the constraint from the spatial neighborhood.

4.3 Foreground and shadow segmentation

Given the video sequence, each pixel in the scene is to be classified as background, shadow, or foreground. For a site $\mathbf{x}$ in the kth frame, the segmentation label $s_k(\mathbf{x})$ equals 1 for a background pixel, 2 for shadow, and 3 for foreground. Here static shadows are considered to be part of the background.

4.3.1 Local observation

In order to segment the foreground, the system should first model the background and shadow information. Edge information also helps improve the reliability of detection.

Since indoor environments are relatively stable compared to outdoor scenes, we assume that each pixel in the background is of Gaussian distribution. At time *k*,

$$b_k(\mathbf{x}) = \mu_{b,k}(\mathbf{x}) + n_{b,k}(\mathbf{x}), \tag{4.7}$$

where random variable $b_k(\mathbf{x})$ is the intensity of a pixel **x** within the background, $\mu_{b,k}(\mathbf{x})$ is the intensity mean, and $n_{b,k}(\mathbf{x})$ is independent zero-mean Gaussian noise with variance $\sigma^2_{b,k}(\mathbf{x})$ at time *k*. Intensity means and variances in the background can be estimated from previous images (see Section 4.4.1).

Given the intensity of a background point, we use a linear model to describe the change of intensity for the same point when shadowed in the video frame. At time *k*,

$$g_k(\mathbf{x}) = ab_k(\mathbf{x}) + n_{s,k}(\mathbf{x}) \text{, if } s_k(\mathbf{x}) = 2, \tag{4.8}$$

where the coefficient $a \in [0,1]$, and $n_{s,k}(\mathbf{x})$ is independent zero-mean Gaussian noise with variance $\sigma^2_{s,k}(\mathbf{x})$ at time *k*. The shadow noise $n_{s,k}(\mathbf{x})$ models the deviation from the simple linear approximation in real visual environments, especially when the entire background scene is not flat. Since it is difficult to compute $\sigma^2_{s,k}(\mathbf{x})$ individually for every site **x** in the scene, we assume that $\sigma^2_{s,k}(\mathbf{x})$ equals $\rho^2\sigma^2_{b,k}(\mathbf{x})$, and that the shadow noise is independent of the background noise. Thus the intensity of a shadowed point is of Gaussian distribution with the following mean and variance.

$$E[g_k(\mathbf{x})] = a\mu_{b,k}(\mathbf{x}),$$

$$\text{Var}[g_k(\mathbf{x})] = (a^2 + \rho^2)\sigma^2_{b,k}(\mathbf{x}) \text{, if } s_k(\mathbf{x}) = 2. \tag{4.9}$$

Parameters a and ρ are manually determined. Their values depend on the visual environment, usually $0.5 \le a < 1$ and $0.5 \le \rho \le 1.5$ in indoor scenes.

The edge model is built by applying an edge operator to the scene. For a site $\mathbf{x}$, denote $\mathbf{x}_l$ and $\mathbf{x}_r$ as its two horizontally neighboring (left and right) points, $\mathbf{x}_u$ and $\mathbf{x}_d$ its two vertically neighboring (up and down) points. At time k, the image edge vector $\mathbf{e}_{g,k}(\mathbf{x})$ is denoted by $(e^h_{g,k}(\mathbf{x}), e^v_{g,k}(\mathbf{x}))^T$, where $e^h_{g,k}(\mathbf{x}) = g_k(\mathbf{x}_l) - g_k(\mathbf{x}_r)$ is the horizontal difference, and $e^v_{g,k}(\mathbf{x}) = g_k(\mathbf{x}_u) - g_k(\mathbf{x}_d)$ is the vertical difference. The entire image edge field is expressed as $\mathbf{e}_{g,k}$.

Similarly, we can model the edge information for the background. At time k, the background edge vector $\mathbf{e}_{b,k}(\mathbf{x})$ for a site $\mathbf{x}$ is denoted by $(e^h_{b,k}(\mathbf{x}), e^v_{b,k}(\mathbf{x}))^T$, where $e^h_{b,k}(\mathbf{x}) = b_k(\mathbf{x}_l) - b_k(\mathbf{x}_r)$ and $e^v_{b,k}(\mathbf{x}) = b_k(\mathbf{x}_u) - b_k(\mathbf{x}_d)$ are the horizontal difference and the vertical difference respectively. It can be known from the background model that $\mathbf{e}_{b,k}(\mathbf{x})$ is of bivariate normal distribution. According to the independent background noise assumption, the corresponding mean $\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})$ of the distribution can be calculated from the intensity means and variances of the four neighboring points.

$$\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x}) = (\mu_{b,k}(\mathbf{x}_l) - \mu_{b,k}(\mathbf{x}_r), \mu_{b,k}(\mathbf{x}_u) - \mu_{b,k}(\mathbf{x}_d))^T,$$

$$\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x}) = \begin{pmatrix} \sigma^2_{b,k}(\mathbf{x}_l) + \sigma^2_{b,k}(\mathbf{x}_r), 0 \\ 0, \sigma^2_{b,k}(\mathbf{x}_u) + \sigma^2_{b,k}(\mathbf{x}_d) \end{pmatrix}. \tag{4.10}$$

The edge model can be used to detect structure changes in the scene as edge features appear, vanish, or rotate. Although other edge operators such as the Sobel operator

can be applied as well, we use the above operator with a diagonal covariance matrix to simplify the computation.

4.3.2 Likelihood model

Since the image edge field $\mathbf{e}_{g,k}$ is totally determined by the image g_k, the observation (or likelihood) model $p(g_k \mid s_k)$ can be written as $p(g_k, \mathbf{e}_{g,k} \mid s_k)$. Then the factorization of the likelihood in (4.5) becomes

$$p(g_k \mid s_k) = p(g_k, \mathbf{e}_{g,k} \mid s_k)$$

$$= \prod_{\mathbf{x} \in \mathbf{X}} p(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x})), \tag{4.11}$$

where $\mathbf{o}_k(\mathbf{x})$ in (4.5) is replaced by $(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}))$ to integrate both intensity and edge features. Given the segmentation label, we assume that the image intensity and image edge are conditionally independent on each other at each site. Hence the local likelihood can be factorized as the product of intensity likelihood and edge likelihood.

$$p(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x}))$$

$$= p(g_k(\mathbf{x}) \mid s_k(\mathbf{x})) p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x})). \tag{4.12}$$

When site $\mathbf{x}$ is in the background, the intensity likelihood can be calculated using the background model.

$$p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 1) = N(g_k(\mathbf{x}); \mu_{b,k}(\mathbf{x}), \sigma^2_{b,k}(\mathbf{x})), \tag{4.13}$$

where $N(\mathbf{z}; \mathbf{m}, \mathbf{\Sigma})$ is a Gaussian distribution with argument $\mathbf{z}$, mean $\mathbf{m}$, and covariance $\mathbf{\Sigma}$.

When site **x** is shadowed, the probability density can be calculated by the shadow model.

$$p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 2)$$

$$= N(g_k(\mathbf{x}); a\mu_{b,k}(\mathbf{x}), (a^2 + \rho^2)\sigma_{b,k}^2(\mathbf{x})). \quad (4.14)$$

When site **x** is in the foreground, the background has no influence on the pixel intensity information. Uniform distribution is assumed for the foreground pixel intensity. The conditional probability density becomes

$$p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 3) = \frac{1}{y_{\max}}. \quad (4.15)$$

Here $[0, y_{\max}]$ is the intensity range for a point in the scene.

For each point **x**, denote the set of its four nearest neighboring points by $N'_{\mathbf{x}} = \{\mathbf{x}_l, \mathbf{x}_r, \mathbf{x}_u, \mathbf{x}_d\}$. Considering the spatial connectivity of the scene, we assume that the four neighboring points have the same segmentation label as **x**. Thus the edge likelihood is approximated by

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x}) = j) \approx p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = j). \quad (4.16)$$

Similarly, when the area $N'_{\mathbf{x}}$ is in the background, the probability density can be computed by the edge model.

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 1)$$

$$= N(\mathbf{e}_{g,k}(\mathbf{x}); \boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x}), \boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})). \quad (4.17)$$

When the area $N'_{\mathbf{x}}$ is shadowed, the edge likelihood can be computed using the models in Section 4.3.1.

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 2)$$

$$= N(\mathbf{e}_{g,k}(\mathbf{x}); a\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x}), (a^2 + \rho^2)\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})) \,. \tag{4.18}$$

When the area $N'_{\mathbf{x}}$ belongs to the foreground, we assume that the point intensity within the foreground is independent and identically distributed (i. i. d.). From (4.15), it can be known that

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 3)$$

$$= p(e^h_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 3)\ p(e^v_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 3)$$

$$= \left(\frac{1}{y_{\max}} - \frac{|\, e^h_{g,k}(\mathbf{x})\,|}{y^2_{\max}}\right)\left(\frac{1}{y_{\max}} - \frac{|\, e^v_{g,k}(\mathbf{x})\,|}{y^2_{\max}}\right). \tag{4.19}$$

4.3.3 Segmentation algorithm

Substitute ($g_{k+1}(\mathbf{x})$, $\mathbf{e}_{g,k+1}(\mathbf{x})$) for $\mathbf{o}_{k+1}(\mathbf{x})$ in (4.6a) and combine the likelihood model in Section 4.3.2, then the one-pixel potential function for the segmentation field at time k+1 can be updated by the DHMRF filter.

$$V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) = i \mid g_{1:k+1})$$

$$= -\ln\{\sum_j \exp[-\alpha_{i|j} - \lambda_k \sum_{\mathbf{y}\in M_{\mathbf{x}}} \frac{1}{|\, M_{\mathbf{y}}\,|} V_{\mathbf{y}}(s_k(\mathbf{y}) = j \mid g_{1:k})]\} -$$

$$\ln(g_{k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x}) = i) - \ln(\mathbf{e}_{g,k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x}) = i)\,, \tag{4.20}$$

where $1 \le i, j \le 3$. Meanwhile the two-pixel potential $V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))$ can be calculated using (4.6b).

At time k+1, the MAP (maximum a posteriori) estimate of the segmentation field is computed as

$$\hat{s}_{k+1} = \arg\max_{s_{k+1}} p(s_{k+1} \mid g_{1:k+1})$$

$$= \arg\min_{s_{k+1}} \{ \sum_{\mathbf{x} \in \mathbf{X}} [V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid g_{1:k+1}) + \frac{1}{2} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))] \} . \quad (4.21)$$

4.4 Implementation

4.4.1 Background updating

For stationary background scenes, the intensity mean and variance of each background point can be estimated from a sequence of background images recorded at the beginning.

For nonstationary background scenes, the background updating process is based on the idea of Stauffer and Grimson [57]. The recent history of each pixel is modeled by a mixture of Gaussians. As parameters of the mixture model change, the Gaussian distribution that has the highest ratio of weight over variance is chosen as the background model. After the segmentation of an image, each pixel is checked to match the existing Gaussian distributions. For a matched Gaussian, its weight increases and the corresponding mean and variance are updated utilizing the pixel value. For unmatched distributions, the means and variances remain the same, while the weights should be renormalized. If none of the distributions match the pixel value, the distribution of the lowest weight is replaced with a Gaussian with the pixel value as its mean, initially low weight and high variance.

The main difference between the Gaussian mixture method and our approach in background updating is the definition of match. In [57], a Gaussian is matched if the pixel value is within 2.5 standard deviations of the distribution. In our work, if the

point is classified as background by the segmentation algorithm (DHMRF filtering), then the Gaussian corresponding to the background model is matched, otherwise a Gaussian is matched if the value is within 2.5 standard deviations of the distribution. Thus the estimation by the DHMRF filter is employed in the updating process. Each time after background updating, the models of shadow and edge information can be updated by (4.9) and (4.10).

4.4.2 Parameters and optimization

In the one-pixel potential function (4.20), the potential of state transition is expressed as $\alpha_{i|j} \propto (1 - \delta(i - j))$, so that segmentation labels for the same site are likely to remain the same at consecutive time instants. To balance the influence of the terms in (4.20), we assume that

$$\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}[\sum_{j}\lambda_k \sum_{\mathbf{y}\in M_{\mathbf{x}}}\frac{1}{|M_{\mathbf{y}}|}V_{\mathbf{y}}(s_k(\mathbf{y})=j \mid g_{1:k})] = \sum_{j}\alpha_{i|j} = \gamma\,. \tag{4.22}$$

Hence λ_k and $\alpha_{i|j}$ are estimated as

$$\lambda_k = \frac{\gamma}{\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}\sum_{j}V_{\mathbf{x}}(s_k(\mathbf{x})=j \mid g_{1:k})},$$

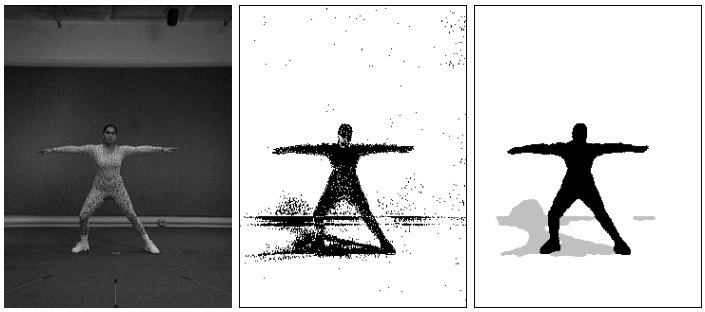
$$\alpha_{i|j} = \frac{1}{2}\gamma(1-\delta(i-j))\,,\ 1 \le i,j \le 3. \tag{4.23}$$

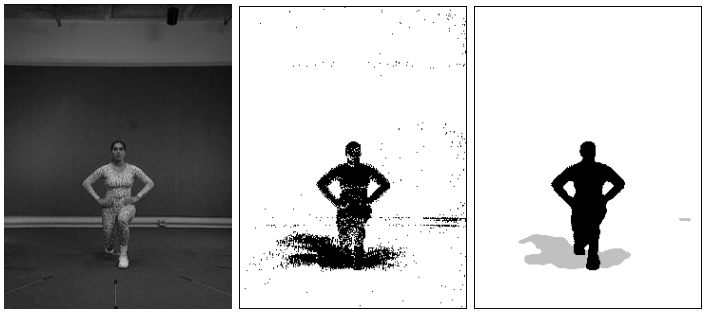
The parameters γ and β in (4.6b) are manually determined to reflect the importance of observed information and spatial connectivity respectively. Initially, the one-pixel potential $V_{\mathbf{x}}(s_0(\mathbf{x})=j) = \frac{\gamma}{3}$ for all $\mathbf{x}$ and j, and $\lambda_0 = 1$.

At each time, the MAP estimate is obtained by minimizing the objective function in (4.21). The objective function is nonconvex and does not have a unique minimum. Obviously, there is no simple method of performing the optimization. To arrive at a sub-optimal estimate, we use a local technique known as iterated conditional modes (ICM) [4]. The ICM algorithm employs the greedy strategy in iterative minimization. Initially, segmentation labels are set by maximizing the likelihood. Given the observed data and estimated labels of the latest iterative step, segmentation labels are sequentially updated by locally minimizing the objective function at each site.

4.5 Results and discussion

The proposed approach has been tested on monocular grayscale video sequences captured in different indoor environments. (For color images, they are first converted into grayscale ones.) Figure 4.2-3 show the segmentation results of two sequences with stationary background scenes, and Figure 4.4-5 show the segmentation results of two sequences with nonstationary background scenes. In Figure 4.4-5, our technique is compared to the Gaussian mixture (GM) method [57] and the method used in the W^4 system [24]. Unless otherwise stated, the segmentation results by our method are obtained using the 24-pixel spatial neighborhood and the 81-pixel temporal neighborhood.

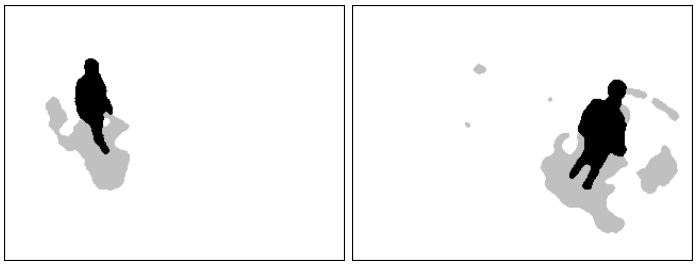




(a) (b) (c)

Figure 4.2 Segmentation results of the "aerobic" sequence. (a) Two frames of the sequence. (b) Segmentation results by simple background subtraction. (c) Segmentation results by the proposed method.

(a.1) (a.2)



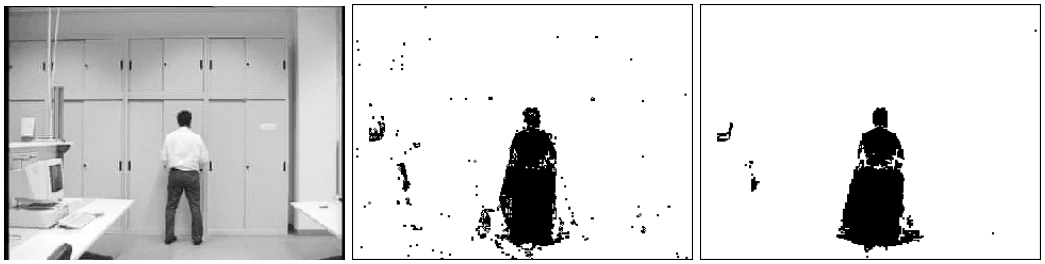
(b.1) (b.2)

(c.1) (c.2) (c.3)

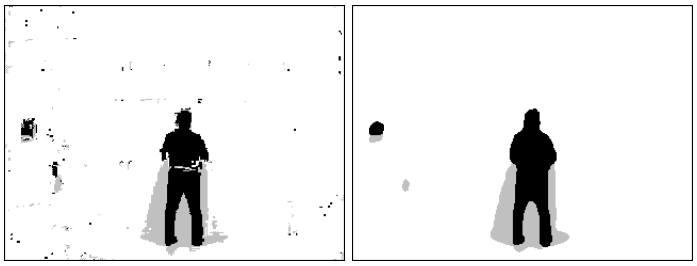
Figure 4.3 Segmentation results of the "room" sequence. (a) Two frames of the sequence. (b) Segmentation results by the proposed method. (c) Information of background, shadow, and foreground from previous frames for (a.2).

Figure 4.2 shows the segmentation results for two frames of the "aerobic" sequence using simple background subtraction and the proposed method. The gray regions in Figure 4.2c represent moving cast shadows. Compared to simple background subtraction, the proposed approach greatly improves the accuracy of foreground detection. The moving cast shadows attached to the woman in Figure 4.2b are exactly removed from the foreground in Figure 4.2c. The flickering pixels in the background and camouflage regions at the woman's neck and legs are erroneously detected in Figure 4.2b, while these problems are overcome by the proposed method.

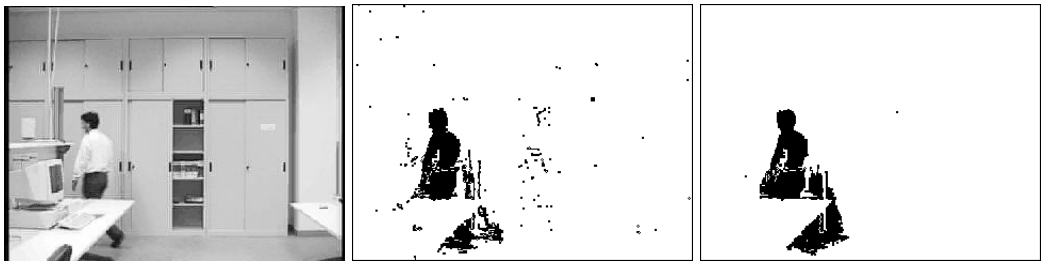
Figure 4.3 shows the segmentation results for two frames of the "room" sequence by the proposed method. Moving shadows cast at different locations of the wall and the floor are discriminated from the man in Figure 4.3b. When shadows are cast on multiple planes in the background scene, the noise term in the shadow model (4.8) ameliorates the linear approximation of intensity change under shadow. Figure 4.3c shows the information from previous video frames (the first term in the one-pixel potential function (4.20)) for the second image. The bright gray levels indicate high prior probability for the corresponding class (background, shadow, and foreground). It can be seen that previous observations enhance the confidence of foreground and shadow segmentation.



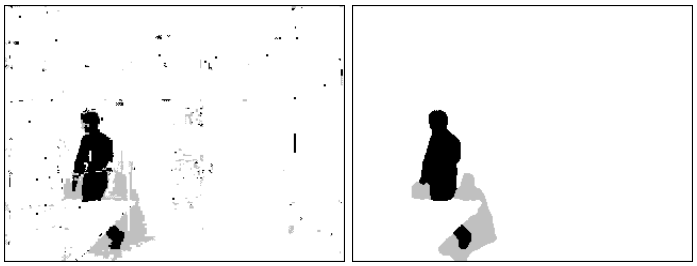
(a.1) (b.1) (c.1)



(d.1) (e.1)



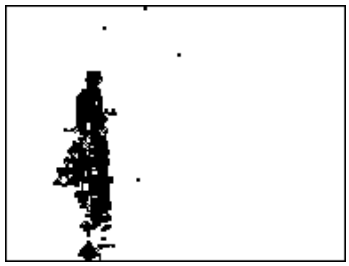
(a.2) (b.2) (c.2)



(d.2) (e.2)

Figure 4.4 Segmentation results of the "laboratory" sequence. (a) Two frames of the sequence. (b) Segmentation results by GM. (c) Segmentation results by W^4. (d) Segmentation results by the proposed method using the 4-pixel spatial neighborhood

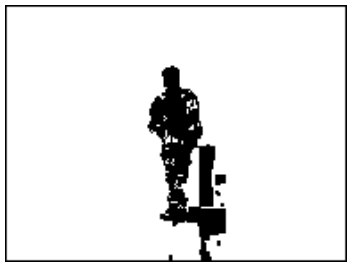
and the 9-pixel temporal neighborhood. (e) Segmentation results by the proposed method using the 24-pixel spatial neighborhood and the 81-pixel temporal neighborhood.

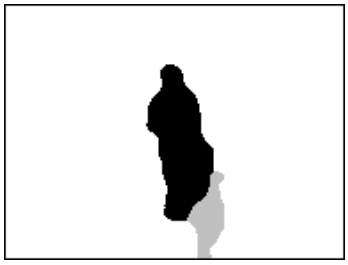

(a.1) (b.1) (c.1)


(d.1) (e.1)





(a.2) (b.2) (c.2)




(d.2) (e.2)

Figure 4.5 Segmentation results of another "laboratory" sequence. (a) Two frames of the sequence. (b) Segmentation results by GM. (c) Segmentation results by W^4. (d) Segmentation results by the proposed method without using edge information. (e) Segmentation results by the proposed method.

Figure 4.4 shows the segmentation results using GM, W^4, and the proposed method for two frames of the "laboratory" sequence with background object change. The open cabinet in the second image is classified as background in Figure 4.4b-e by all the methods after a period of background updating. Figure 4.4d and 4.4e show the influence of neighborhood size. The camouflage regions and flickering areas in Figure 4.4d are corrected in Figure 4.4e by increasing spatio-temporal contextual constraints when the noise in the scene is heavy.

Figure 4.5 shows the segmentation results by GM, W^4, and the proposed method for two frames of another "laboratory" sequence with background illumination change. The illumination change in the second image caused by switching off part of the light is updated for the background in Figure 4.5b-e by all the methods. Figure 4.5d and 4.5e show that the integration of edge information helps locate structure changes of the scene and improves the reliability of foreground detection.

Table 4.1 Quantitative evaluation of foreground segmentation results.

	false negative	false positive
GM	3.0%	3.5%
W^4	5.2%	2.7%
proposed	3.3%	0.9%

The results are also evaluated quantitatively in terms of false negative rate (the portion of foreground pixels that are misclassified as non-foreground) and false positive rate (the portion of non-foreground pixels that are misclassified as foreground) by comparing to the manually segmented ground-truth foreground images. Before quantitative comparison, the segmentation results by the two other methods are smoothed to remove small erroneously detected areas. The average error rates for ten frames of the two laboratory sequences (five frames with different foreground object positions for each sequence) are summarized in Table 4.1. The moving shadows cast on the floor, wall, and table result in an increase of falsely detected foreground pixels (false positive) in Figure 4.4b-c and 4.5b-c. With an explicit shadow model, it is relatively easy for our approach to know which part of the pixel intensity distribution is likely to be produced by shadows. Moreover, both spatial and temporal constraints are employed in our approach. Hence the false positive rate is reduced by the proposed method with a tradeoff in relatively high computation load. On the other hand, in indoor scenes the intensity variance of a point under shadow is usually greater than the variance of the same site in the background. Since the pixel intensity distribution of the foreground is assumed to be uniform, foreground regions darker than the background tend to be misclassified when the intensity variances under shadow are excessively large. This effect makes part of the man's arms erroneously detected as shadow in the first image of Figure 4.4e, and the false negative rate of our approach higher than that of the Gaussian mixture method.

There are two main contributions in this chapter. First, we have proposed a dynamic hidden Markov random field (DHMRF) model that combines the HMM and the MRF for video sequences. Second, we have derived an efficient approximate

DHMRF filtering algorithm and applied it to moving object and cast shadow detection in indoor scenes. The DHMRF model unifies the constraints of spatial connectivity and temporal continuity in the segmentation process. Objects and shadows usually form contiguous regions, and a point is likely to have the same segmentation label in consecutive frames. Two other kinds of spatial and temporal information are employed in our approach as well. The spatial gradient (or edge) information is integrated to help detect structure changes in the scene, and the recent history of observed images is used to adaptively update the models of background, shadow, and edge information. The proposed approach does not require range or color data and performs robust foreground segmentation. Experimental results show that our method accurately distinguishes moving objects from their cast shadows in nonstationary background scenes.

Chapter 5

Multi-object Tracking with Switching Hypothesized Measurements

5.1 Introduction

The idea of hypothesized measurements, which results in a switching hypothesized measurements (SHM) model that differs from previous state space models, is proposed in this chapter. The ability to support multimodality makes the model suitable for handling the potential variability in visual tracking. At each time instant, the approach acquires a set of hypothesized measurements for different occlusion hypotheses rather than uses a uniform measurement process. A computationally efficient filtering algorithm is derived for tracking multiple objects jointly. Both occlusion relationships and states of the objects are estimated from the history of hypothesized measurements. The proposed method helps prevent distractions from background clutter. When there is a high confidence in nonocclusion, the reference regions can be adaptively updated to deal with object appearance changes. Moreover, the SHM model is generally applicable to dynamic processes with multiple alternative measurement methods.

As to the related work, Ghahramani and Hinton introduced a dynamic Bayesian network framework for learning and inference in switching state space models [21]. Pavlovic et al. proposed a switching linear dynamic system (SLDS) approach for human motion analysis [47]. A switching model framework for the Condensation algorithm is also proposed by Isard and Blake [26]. In their work, the switching

variable determines which dynamic model is in effect at each time instant. Rather than switches among a set of models, the SHM approach switches among a set of known hypothesized measurements. The joint probabilistic data association algorithm [3] can be cast in the framework of SLDS as well. Moreover, in our model each measurement component corresponds to one and only one given target region (see section 5.3). Rasmussen and Hager describe a joint measurement process enumerating all possible occlusion relationships [51]. The measurement with respect to the most possible occlusion relationship is determined using the information from the current frame. The corresponding measurement is then plugged into a Kalman tracker. In our approach, the estimation is based on the history of all the (hypothesized) measurements. In the work of Galvin et al. [17], two virtual snakes, a background and a foreground snake for each object, are generated to resolve the occlusion when two objects intersect. Their manner parallels to the case of acquiring measurements under a set of two hypotheses in our method.

The remainder of the chapter is arranged as follows: Section 5.2 presents the formulation of the SHM model. Section 5.3 proposes the measurement process for joint region tracking. Section 5.4 derives the filtering algorithm. Section 5.5 describes the implementation details. Section 5.6 discusses the experimental results.

5.2 Model

5.2.1 Generative SHM model

To model a dynamic system with state space representation, consider the evolution of a hidden state sequence $\{\mathbf{z}_k\}$ ($k \in \mathbf{N}$), given by

$$\mathbf{z}_{k+1} = \mathbf{f}_k(\mathbf{z}_k, \mathbf{n}_k), \tag{5.1}$$

where $\mathbf{f}_k : \mathbf{R}^{n_z} \times \mathbf{R}^{n_n} \rightarrow \mathbf{R}^{n_z}$ is a state transition function, and $\{\mathbf{n}_k\}$ is a process noise sequence. The objective of online tracking is to recursively estimate $\mathbf{z}_k$ from a measurement sequence. In a complex system with dynamic mode control, there exists a mode or switching state sequence $\{s_k\}$, with $s_k \in \{1, 2, \ldots, L\}$ ($L \in \mathbf{N}$). The switching state s_k determines which mode is in effect at time k. Usually the sequence $\{s_k\}$ is modeled as an unobserved discrete first order Markov process.

Specifically, the mode switching is associated with the measurement process in our work. The notion of a uniform measurement is extended to a set of L hypothesized measurements $\mathbf{y}_k = (\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, \ldots, \mathbf{y}_{k,L})$ [70]. Each $\mathbf{y}_{k,j}$ ($1 \le j \le L$) is called a hypothesized measurement since it is obtained by assuming that the switching state s_k is j at time k. For the measurement under the jth hypothesis,

$$\mathbf{y}_{k,j} = \mathbf{h}_{k,j}(s_k, \mathbf{z}_k, \mathbf{v}_{k,j}), \tag{5.2}$$

where $\mathbf{h}_{k,j} : \mathbf{N} \times \mathbf{R}^{n_z} \times \mathbf{R}^{n_v} \rightarrow \mathbf{R}^{n_y}$ is the measurement function, and $\mathbf{v}_{k,j}$ is the measurement noise under the jth hypothesis. To make the model computationally efficient, we assume that the hypothesized measurements are conditionally independent on each other when both the hidden state $\mathbf{z}_k$ and the switching state s_k are given. This switching hypothesized measurements (SHM) model can be represented by a dynamic Bayesian network shown in Figure 5.1.

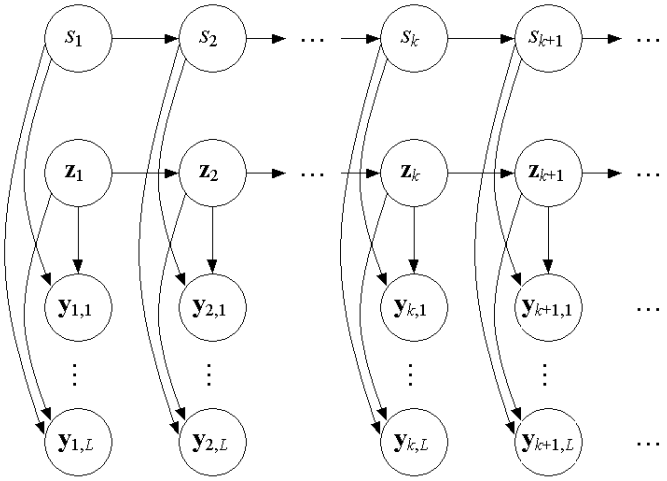


Figure 5.1 Bayesian network representation of the SHM model.

5.2.2 Example of hypothesized measurements

To illustrate the idea of hypothesized measurements in the SHM model, a simple example of the measurement process for jointly tracking a rectangle and a circle is studied in this section. The two objects translationally move in an image sequence $\{g_k\}$.

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

Figure 5.2 Illustration of hypothesized measurements. (a) (b) Two frames of the "rectangle and circle" sequence under different occlusion relationships. (c) Masked image under the first occlusion hypothesis. (d) Masked image under the second occlusion hypothesis.

When measuring the centroids of these two objects from the kth frame g_k, two occlusion relationship hypotheses (hypotheses corresponding to the rectangle being

in front of the circle and the circle being in front of the rectangle, see Figure 5.2a and 5.2b) should be considered. The switching state s_k is introduced to describe the depth ordering at time k. s_k equals 1 if the rectangle is in front of the circle, and 2 if the circle is in front of the rectangle. The hypothesized measurement $\mathbf{y}_{k,j}$ ($1 \leq j \leq 2$) is written as $(\mathbf{y}_{k,j}^{(1)}, \mathbf{y}_{k,j}^{(2)})^T$, where $\mathbf{y}_{k,j}^{(1)}$ is the measurement of the rectangle centroid, and $\mathbf{y}_{k,j}^{(2)}$ is the measurement of the circle centroid under the jth hypothesis.

Under the hypothesis of $s_k = 1$, i.e. the circle is occluded by the rectangle at time k, the rectangle should be measured first to acquire $\mathbf{y}_{k,1}^{(1)}$. Then the observed rectangle is masked in the image (see Figure 5.2c). The occluded area of the circle is ignored and only the visible region is matched normally to get $\mathbf{y}_{k,1}^{(2)}$. Thus, the occlusion will not affect the measurement result. Similarly, under the hypothesis of $s_k = 2$, i.e. the rectangle is occluded by the circle, the circle should be matched first to get $\mathbf{y}_{k,2}^{(2)}$, then the masked image (see Figure 5.2d) is used to measure $\mathbf{y}_{k,2}^{(1)}$.

Given the occlusion relationship s_k at time k, the hypothesized measurement $\mathbf{y}_{k,j}$ for $j \neq s_k$ may bias the true value since the measurement is obtained under a false hypothesis. Unfortunately, whether the rectangle occludes the circle or the circle occludes the rectangle is not given before hand. So it is not known whether $\mathbf{y}_{k,1}$ or $\mathbf{y}_{k,2}$ is the proper measurement for frame g_k. To handle this uncertainty, the occlusion relationship could be estimated from the history of all the hypothesized measurements.

Moreover, it is obvious that both hypothesized measurements support the condition of nonocclusion since different depth orderings of nonoverlapping objects are

visually equivalent. The values of $p(s_k = 1)$ and $p(s_k = 2)$ should be equal in the case of nonocclusion.

5.2.3 Linear SHM model for joint tracking

For joint tracking of M ($M \in \mathbf{N}$) objects in the scene, the switching state s_k represents the occlusion relationship (or depth ordering) at time k, $s_k \in \{1, \ldots, L\}$. The number of all occlusion relationship hypotheses is $L = M!$. The switching state transition probability is given as

$$p(s_{k+1} = i \mid s_k = j) = \alpha_{i,j}, \text{ with} \sum_i \alpha_{i,j} = 1. \tag{5.3}$$

The hidden state $\mathbf{z}_k$ is denoted as $(\mathbf{z}_k^{(1)}, \mathbf{z}_k^{(2)}, \ldots, \mathbf{z}_k^{(M)})^T$, with $\mathbf{z}_k^{(m)}$ $(1 \le m \le M)$ being the state of the mth object (e.g. position and velocity) at time k. For a linear process with Gaussian noise, the hidden state transition function is

$$\mathbf{z}_{k+1} = \mathbf{F}\mathbf{z}_k + \mathbf{n},$$

$$p(\mathbf{z}_{k+1} \mid \mathbf{z}_k) = N(\mathbf{z}_{k+1}; \mathbf{F}\mathbf{z}_k, \mathbf{Q}), \tag{5.4}$$

where $\mathbf{F}$ is the state transition matrix, $\mathbf{n}$ is a zero-mean Gaussian noise with covariance matrix $\mathbf{Q}$, and $N(\mathbf{z}; \mathbf{m}, \mathbf{\Sigma})$ is a Gaussian density with argument $\mathbf{z}$, mean $\mathbf{m}$, and covariance $\mathbf{\Sigma}$.

Given the switching state s_k at time k, the corresponding hypothesized measurement $\mathbf{y}_{k,s_k}$ could be considered as a proper measurement centering on the true value, while every other $\mathbf{y}_{k,j}$ for $j \neq s_k$ is an improper measurement generated under a wrong assumption. The improper measurement should be weakly influenced by the hidden state $\mathbf{z}_k$ and have a large variance. To simplify the computation, we assume a normal

distribution for a proper measurement and a uniform distribution for an improper measurement. The measurement function is simplified as

$$\mathbf{y}_{k,j} = \begin{cases} \mathbf{H}\mathbf{z}_k + \mathbf{v}_{k,j}, \text{if } j = s_k, \\ \mathbf{w}, \text{otherwise}, \end{cases}$$

$$p(\mathbf{y}_{k,j} \mid s_k, \mathbf{z}_k) = \begin{cases} N(\mathbf{y}_{k,j}; \mathbf{H}\mathbf{z}_k, \mathbf{R}_{k,j}), \text{if } j = s_k, \\ \text{a constant, otherwise}, \end{cases} \tag{5.5}$$

where $\mathbf{H}$ is the measurement matrix and $\mathbf{v}_{k,j}$ is a zero-mean Gaussian noise with covariance matrix $\mathbf{R}_{k,j}$. $\mathbf{w}$ is a uniformly distributed noise, whose density is a small positive constant. For the measurement of M objects (e.g. translation), $\mathbf{y}_{k,j}$ is denoted as $(\mathbf{y}_{k,j}^{(1)}, \mathbf{y}_{k,j}^{(2)}, \ldots, \mathbf{y}_{k,j}^{(M)})^T$, and $\mathbf{v}_{k,j}$ is written as $(\mathbf{v}_{k,j}^{(1)}, \mathbf{v}_{k,j}^{(2)}, \ldots, \mathbf{v}_{k,j}^{(M)})^T$.

Combining with the conditional independence among the hypothesized measurements, we know that

$$p(\mathbf{y}_k \mid s_k = j, \mathbf{z}_k) = p(\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, \ldots, \mathbf{y}_{k,L} \mid s_k = j, \mathbf{z}_k)$$

$$= \prod_l p(\mathbf{y}_{k,l} \mid s_k = j, \mathbf{z}_k)$$

$$= p(\mathbf{y}_{k,j} \mid s_k = j, \mathbf{z}_k) \prod_{l \neq j} p(\mathbf{y}_{k,l} \mid s_k = j, \mathbf{z}_k)$$

$$\propto N(\mathbf{y}_{k,j}; \mathbf{H}\mathbf{z}_k, \mathbf{R}_{k,j}). \tag{5.6}$$

5.3 Measurement

Multiple, occluding objects are modeled using layer representation. Layers are indexed by $m = 1, 2, \ldots, M$, with layer 1 being the layer that is closest to the camera and layer m being behind layer 1, 2, …, m–1. There is one object in each layer. Each

depth ordering permutation is tagged with an index j ($1 \le j \le L$). For the example in section 5.2.2, it is known that $M = 2$ and $L = 2$.

Under each occlusion relationship hypothesis, the object in the front layer 1 should be measured first from the image g_k at time k. Then the object in layer 2 can be matched from the masked image, and so on. At last the object in layer M can be measured. Thus occluded points are not matched when measuring the objects. Measurement results of nonoverlapping objects should be equivalent for different depth ordering permutations. During the measurement process, the motion of a point $\mathbf{x}$ within the target region is described by a parametric model $\mathbf{d}(\boldsymbol{\theta}, \mathbf{x})$, with $\mathbf{d}(\mathbf{0}, \mathbf{x}) = \mathbf{x}$. $\boldsymbol{\theta} = (\theta^{1)}, \theta^{2)}, \ldots, \theta^{(n_\theta)})$ is a set of motion parameters. The dimension of the motion vector $\boldsymbol{\theta}$, i.e. $n_{\boldsymbol{\theta}}$, changes under different motion models (e.g. two for the translational model, six for the affine model, and eight for the perspective model). Under the jth hypothesis, the measurement for the mth object $\mathbf{y}_{k,j}^{(m)}$ is denoted as $(y_{k,j}^{(m,1)}, y_{k,j}^{(m,2)}, \ldots, y_{k,j}^{(m,n_\theta)})$. Given the reference image g_r ($r < k$), the measurement is based on minimizing the mean of squared intensity differences between the current image and the reference region. The mth object is located at area D_m in the reference image. For each measured $\mathbf{y}_{k,j}^{(m)}$, $e_{k,j}^{(m)}$ is the corresponding minimum squared difference mean.

The measurement noise for the mth object $\mathbf{v}_{k,j}^{(m)}$ is denoted as $(v_{k,j}^{(m,1)}, v_{k,j}^{(m,2)}, \ldots, v_{k,j}^{(m,n_\theta)})$ under the jth hypothesis. From appendix B, it can be known that

$$E[(v_{k,j}^{(m,i)})^2] \approx \frac{|D_m|}{\sum_{\mathbf{x}\in D_m}[g_r(\mathbf{d}(\mathbf{e}_i,\mathbf{x})) - g_r(\mathbf{x})]^2} e_{k,j}^{(m)}, \tag{5.7}$$

where $\mathbf{e}_i$ is the unit vector of dimension $n_{\boldsymbol{\theta}}$ with a non-zero element in the ith position. To reduce the computation, it is assumed that the components of the measurement noise are uncorrelated to each other. Thus the diagonal matrix $\mathbf{R}_{k,j}$ can be easily computed from (5.7). Moreover, it should be noted that other measurement approaches (e.g. the snake methods in [16] and [17]) are also applicable for the SHM model.

5.4 Filtering

From a Bayesian perspective, the online tracking problem is to recursively calculate the posterior state space distribution. Given the measurement data $\mathbf{y}_{1:k} = \{\mathbf{y}_i\}_{1\le i\le k}$ up to time k, the probability density function (pdf) $p(s_k, \mathbf{z}_k \mid \mathbf{y}_{1:k})$ is expressed as

$$p(s_k = j, \mathbf{z}_k \mid \mathbf{y}_{1:k}) = p(s_k = j \mid \mathbf{y}_{1:k})\, p(\mathbf{z}_k \mid s_k = j, \mathbf{y}_{1:k})$$

$$= \beta_{k,j}\, N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j}), \tag{5.8}$$

where $p(s_k = j \mid \mathbf{y}_{1:k})$ is denoted as $\beta_{k,j}$, with $\sum_j \beta_{k,j} = 1$, and the pdf $p(\mathbf{z}_k \mid s_k = j, \mathbf{y}_{1:k})$ is modeled as a normal distribution $N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j})$ under each switching state hypothesis. Hence $p(\mathbf{z}_k \mid \mathbf{y}_{1:k})$ is a mixture of L Gaussians.

At time k+1, the set of hypothesized measurements $\mathbf{y}_{k+1}$ becomes available, and it is used to update $\{\beta_{k,j}, \mathbf{m}_{k,j}, \mathbf{P}_{k,j}\}_{1\le j\le L}$ to $\{\beta_{k+1,i}, \mathbf{m}_{k+1,i}, \mathbf{P}_{k+1,i}\}_{1\le i\le L}$. From appendix C, the filtering algorithm is

$$\beta_{k+1,i} = p(s_{k+1} = i \mid \mathbf{y}_{1:k+1})$$

$$= \frac{\sum_j \alpha_{i,j}\beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}{\sum_i \sum_j \alpha_{i,j}\beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}, \tag{5.9}$$

$$p(\mathbf{z}_{k+1} \mid s_{k+1}= i, \mathbf{y}_{1:k+1}) \approx N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i}, \mathbf{P}_{k+1,i}), \tag{5.10}$$

where

$$\mathbf{m}_{k+1|k,j} = \mathbf{F}\mathbf{m}_{k,j},$$

$$\mathbf{P}_{k+1|k,j} = \mathbf{F}\mathbf{P}_{k,j}\mathbf{F}^T + \mathbf{Q},$$

$$\mathbf{S}_{k+1,i|j} = \mathbf{H}\mathbf{P}_{k+1|k,j}\mathbf{H}^T + \mathbf{R}_{k+1,i},$$

$$\mathbf{K}_{k+1,i|j} = \mathbf{P}_{k+1|k,j}\mathbf{H}^T\mathbf{S}_{k+1,i|j}^{-1},$$

$$\mathbf{m}_{k+1,i|j} = \mathbf{m}_{k+1|k,j} + \mathbf{K}_{k+1,i|j}(\mathbf{y}_{k+1,i} - \mathbf{H}\mathbf{m}_{k+1|k,j}),$$

$$\mathbf{P}_{k+1,i|j} = \mathbf{P}_{k+1|k,j} - \mathbf{K}_{k+1,i|j}\mathbf{H}\mathbf{P}_{k+1|k,j},$$

$$\beta_{k+1,i|j} = \frac{\alpha_{i,j}\beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}{\sum_j \alpha_{i,j}\beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})},$$

$$\mathbf{m}_{k+1,i} = \sum_j \beta_{k+1,i|j}\mathbf{m}_{k+1,i|j},$$

$$\mathbf{P}_{k+1,i} = \sum_j \beta_{k+1,i|j}[\mathbf{P}_{k+1,i|j} + (\mathbf{m}_{k+1,i|j} - \mathbf{m}_{k+1,i})(\mathbf{m}_{k+1,i|j} - \mathbf{m}_{k+1,i})^T]. \tag{5.11}$$

The state at time k+1 is estimated as

$$\hat{s}_{k+1} = \arg\max_i p(s_{k+1} = i \mid \mathbf{y}_{1:k+1}) = \arg\max_i \beta_{k+1,i},$$

$$\hat{\mathbf{z}}_{k+1} = \arg\max_{\mathbf{z}_{k+1}} p(\mathbf{z}_{k+1} \mid s_{k+1} = \hat{s}_{k+1}, s_k = \hat{s}_k, \mathbf{y}_{1:k+1}) = \mathbf{m}_{k+1,\hat{s}_{k+1}|\hat{s}_k}. \tag{5.12}$$

It can be seen that the computation of the SHM filter is slightly more complex than the computation of multiple Kalman filters (or Gaussian sum filters [1]).

5.5 Implementation

When an object is totally (or mostly) occluded by the other objects at time k, no (or few) points of the target region will be matched. The corresponding squared difference mean is computed as $\lambda_1 e_{k-1,j}^{(m)}$ for the mth object under the jth hypothesis, where λ_1 ($\lambda_1 > 1$) is a penalty term. The estimation is based on the result of time k–1 when no visible region of the object is expected at time k. The penalty λ_1 helps prevent interpreting an object as being completely occluded when there is image evidence for its visibility.

Due to the variation of the object poses and illumination conditions, the reference image should be updated throughout the tracking process to deal with the object appearance changes. Frame g_k can be used as the reference image when the following is satisfied.

$$\min_j \beta_{k,j} > \frac{\lambda_2}{L}, \quad \max_j \beta_{k,j} < \frac{1}{L\lambda_2}. \tag{5.13}$$

The value of λ_2 is a little bit smaller than one. From (5.13), it is known that the update is with a high confidence in nonocclusion.

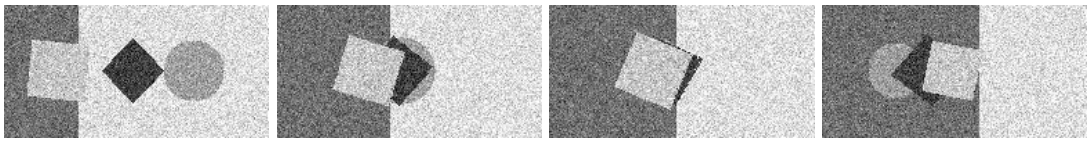
The switching state transition probability is set as

$$\alpha_{i,j} = \begin{cases} 1-\lambda_3, \text{if } i = j, \\ \dfrac{\lambda_3}{L-1}, \text{otherwise}, \end{cases} \tag{5.14}$$

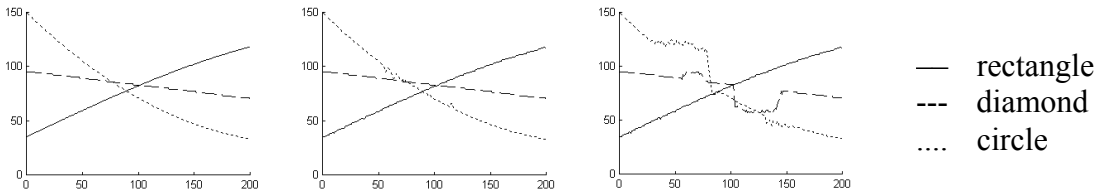
where λ_3 is a small positive value so that two successive switching states are more likely to be of the same label. The transition matrix **F**, covariance matrix **Q**, and measurement matrix **H** are defined in the same way as in a classical Kalman tracker with second order model [53]. The objects are assumed to be separated from each other in the initial image g_0. At the beginning, the reference image is set as $g_r = g_0$. The target regions are detected from the initial image using an adaptive foreground detection method [57]. The initial $\beta_{0,j}$, $\mathbf{m}_{0,j}$, and $\mathbf{P}_{0,j}$ should be equal for different j because of nonocclusion. $\beta_{0,j} = p(s_0 = j) = \frac{1}{L}$. According to the definition of the motion model **d**, the initial mean $\mathbf{m}_{0,j}$ is set as a zero vector. The initial covariance matrix $\mathbf{P}_{0,j}$ is set as diagonal with small variances since the initialization is assumed to be accurate.

5.6 Results and discussion

The proposed approach is tested on both synthetic data and realistic data. The parameter values are set as $\lambda_1 = 1.1$, $\lambda_2 = 0.98$, and $\lambda_3 = 0.1$.



(a.1) (a.2) (a.3) (a.4)



(b) (c) (d)

Figure 5.3 Tracking results of the "three objects" sequence. (a) Four frames of the sequence. (b) True horizontal trajectories of the objects. (c) Tracking result of the SHM filter. (d) Tracking result of the Kalman filter.

Figure 5.3 shows quantitatively the results of jointly tracking a rectangle, a diamond, and a circle under noisy background in a synthetic image sequence of 200 frames. The state of the tracker is the position, diameter, orientation, and the velocities of these parameters. Each measurement is a translation, scaling, and rotation. Figure 5.3a shows the 10th, 70th, 90th, and 130th frame of the sequence. It could be seen that the circle is totally occluded in Figure 5.3a.3. Figure 5.3b shows the true horizontal trajectories of the three objects. Figure 5.3c and 5.3d demonstrate the tracking results of the SHM filter and the Kalman filter. Comparing with the Kalman filter, the tracking performance is greatly improved by our algorithm when heavy occlusions take place among the three objects. The objects are correctly tracked even when total occlusion occurs.

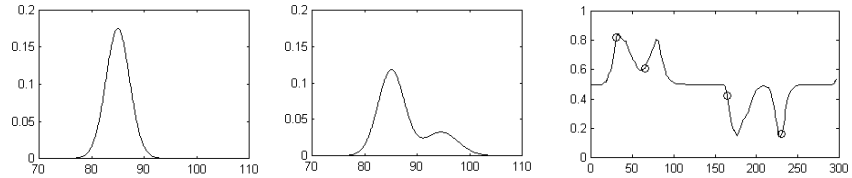
Figure 5.4 shows the tracking of two hands as they cross twelve times in a realistic image sequence of 800 frames. The state of the tracker is the position and orientation, and the velocities of these parameters. Each measurement is a translation and rotation. Figure 5.4a shows the 30th, 65th, 165th, and 230th frame of the sequence. Appearance variation due to hand pose changes is obvious (see Figure 5.4a.4). Figure 5.4b and 5.4c demonstrate the tracking efficacy of the SHM filter versus the Kalman filter. The SHM filter successfully tracks both hands under different occlusion relationships (the left hand being in the front or the right hand being in the front). In Figure 5.4b, one hand is drawn in black contour when the detected depth order indicates that it is in front of the other hand. The Kalman filter

has a similar performance when occlusions are not severe, but poor under heavy occlusions. In Figure 5.4c.4, the distraction from background clutter causes the Kalman tracker to fail. The posterior distributions for the vertical position of the occluded hand in Figure 5.4a.3 and 5.4a.4 are shown in Figure 5.4d and 5.4e. When the occlusion is not severe, measurements under the two hypotheses are similar, and the distribution is unimodal (see Figure 5.4d). Under heavy occlusions, the distribution becomes multimodal (see Figure 5.4e) because the two hypothesized measurements turn to be different. The measurement under true hypothesis matches the hand correctly, while the measurement under false hypothesis is distracted by background clutter. Figure 5.4f shows the probabilities of the first occlusion hypothesis (the left hand being in the front) over the first 300 frames. The probabilities for the four frames shown in Figure 5.4a are circled in Figure 5.4f. The probabilities of the two hypotheses are equal in the nonoverlapping cases, while the probability of the true hypothesis becomes dominant under occlusions. As a byproduct of the SHM filter, the quantitative information helps update reference regions correctly to deal with the object appearance changes.

(a.1) (a.2) (a.3) (a.4)

(b.1) (b.2) (b.3) (b.4)

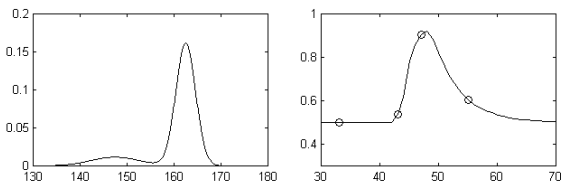
(c.1) (c.2) (c.3) (c.4)



(d) (e) (f)

Figure 5.4 Tracking results of the "crossing hands" sequence. (a) Four frames of the sequence. (b) Tracking results of the SHM filter. (c) Tracking results of the Kalman Filter. (d) (e) Posterior distributions of the left hand's vertical position in (a.3) and (a.4). (f) Probabilities of the left hand being in the front over time.

(a.1) (a.2) (a.3) (a.4)



(b) (c)

Figure 5.5 Tracking results of the "two pedestrians" sequence. (a) Results of tracking the four shanks of two persons. (b) Posterior distribution of the occluded

body's horizontal position in (a.3). (c) Probabilities of the woman's body being in the front over time.

Figure 5.5 shows the results of jointly tracking the four shanks of a man and a woman as they cross in a sequence of 80 frames. There should be totally 4! = 24 hypotheses if we directly apply the SHM filter. Two reasonable assumptions are made to prune less plausible hypotheses. Firstly, one's legs cannot simultaneously occlude and be occluded by the other's legs. Secondly, the occlusion relationship between the man and woman can be determined from their bodies. Thus, the whole tracking procedure is divided into three trackers. The first one tracks the two bodies of the walkers. According to the detected occlusion relationship, the two shanks of the person in the front are then tracked. At last, the shanks of the other person are tracked in the masked image. Figure 5.5a shows the tracking results for the 32nd, 42nd, 46th, and 54th frame of the sequence (circles are marked on the man's body and shanks, and rectangles are marked on the woman). The man's right shank has been totally occluded when they cross. Figure 5.5b shows the posterior distribution for the horizontal position of the occluded body in Figure 5.5a.3. Figure 5.5c shows the probabilities of the woman's body being in the front. The probabilities for the four frames in Figure 5.5a are circled. The number of occlusion relationship hypotheses grows nonlinearly with the increase of objects. To reduce the computation, less plausible hypotheses should be (progressively) pruned when the number of the objects for joint tracking is large.

Under realistic environments, it is understandable that comparing with the other hypothesized measurements, the measurement under the true occlusion hypothesis usually shows more regularity and has a smaller variance. Thus, the true information

(the switching state and the hidden state) could be enhanced through the propagation. In addition, comparing with a uniform measurement process, the acquirement of multiple hypothesized measurements helps decrease the information loss (e.g. caused by background clutter) in complex visual environments before filtering.

This chapter makes two main contributions. First, we propose a switching hypothesized measurements model for multimodal state space representation of dynamic systems. Second, we describe a measurement process and derive an efficient filtering algorithm for joint region tracking in image sequences. Our approach reasons about the occlusion relationships explicitly. The occlusion relationships are quantitatively estimated throughout the propagation. The information can be used for reference update and further analysis. Moreover, experimental results show that our method helps handle appearance changes and distractions. The SHM model discusses the measurement switching in dynamic systems, which is complementary to the idea of model switching in [21] [26] [48]. Furthermore, from section 5.2.1 it can be known that the SHM model is generally applicable to describe various dynamic processes in which there are multiple alternative measurement methods.

Chapter 6

Conclusion

6.1 Summary

In this thesis, probabilistic approaches of segmenting and tracking objects in image sequences based on graphical probabilistic models, especially Bayesian networks and Markov random fields, are studied to deal with the potential variability in visual scenes such as object occlusions, appearance changes, illumination variations, and cluttered environments.

Firstly, this work proposes a unified framework for spatio-temporal segmentation of video sequences. Motion information among successive frames, boundary information from intensity segmentation, and spatial connectivity of object segmentation are unified in the video segmentation process using graphical models. A Bayesian network is presented to model interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. The notion of Markov Random field is used to encourage the formation of continuous regions. Given consecutive frames, the conditional joint probability density of the three fields is maximized in an iterative way. To effectively utilize boundary information from intensity segmentation, distance transformation is employed in local objective functions. The proposed approach is robust and generates spatio-temporally coherent segmentation results. In addition, the proposed video segmentation approach can be viewed as a compromise between previous motion based approach and region merging approach.

Secondly, this thesis proposes a dynamic hidden Markov random field (DHMRF) model for foreground object and moving shadow segmentation in indoor video scenes monitored by a fixed camera. Given an image sequence, temporal dependencies of consecutive segmentation fields and spatial dependencies within each segmentation field are unified in the novel dynamic probabilistic model that combines the hidden Markov model and the Markov random field. An efficient approximate filtering algorithm is derived for the DHMRF model to recursively estimate the segmentation field from the history of observed images. The foreground and shadow segmentation method integrates both intensity and edge information. Moreover, models of background, shadow, and edge information are updated adaptively for nonstationary background processes. The proposed approach can accurately detect moving objects and their cast shadows even in monocular grayscale video sequences.

Thirdly, this work proposes a switching hypothesized measurements (SHM) model supporting multimodal probability distributions and presents the application of the model in handling potential variability in visual environments when tracking multiple objects jointly. For a set of occlusion hypotheses, a frame is measured once under each hypothesis, resulting in a set of measurements at each time instant. The dynamic model switches among hypothesized measurements during the propagation. A computationally efficient SHM filter is derived for online joint object tracking. Both occlusion relationships and states of the objects are recursively estimated from the history of hypothesized measurements. The reference image is updated adaptively to deal with appearance changes of the objects. Moreover, the SHM model is generally applicable to various dynamic processes with multiple alternative measurement methods.

The proposed approaches deal with object segmentation and tracking from relatively comprehensive and general viewpoints, and each of them can be used individually in video analysis. Experimental results show that the approaches accurately segment and track objects, and the techniques robustly deal with the potential variability under real visual environments.

6.2 Future work

In the video segmentation method, the localization properties in image sequences are not considered to simplify the computation. More advanced segmentation techniques that account for both local information and spatio-temporal information could be adopted, but that requires computation load reduction through efficient optimization schemes [10] [37]. This could be our future research. Moreover, adaptive methods for automatic determination of the number of objects and strength of the spatio-temporal constraints would be beneficial [2].

The SHM model studies the measurement switching in dynamic systems. It is complementary to the idea of model switching in [21]. Our future work is the effective combination of these two ideas, which may result in a more powerful approach for visual tracking.

To make the proposed techniques practical in applications, it will be beneficial for our future study to explore more accurate and efficient approximate filtering algorithms of both the DHMRF model and the SHM model, as well as automatically determine all the parameters for the segmentation and tracking approaches.

Appendix A The DHMRF filtering algorithm

At time k+1, image g_{k+1} is used to update the posterior distribution of the segmentation field via Bayes' rule.

$$p(s_{k+1} \mid g_{1:k+1}) = \frac{p(s_{k+1} \mid g_{1:k}) p(g_{k+1} \mid s_{k+1})}{p(g_{k+1} \mid g_{1:k})}$$

$$\propto p(s_{k+1} \mid g_{1:k}) p(g_{k+1} \mid s_{k+1}). \tag{A.1}$$

In the hidden Markov model, the conditional probability $p(s_{k+1} \mid g_{1:k})$ is computed as

$$p(s_{k+1} \mid g_{1:k}) = \sum_{s_k} p(s_{k+1}, s_k \mid g_{1:k})$$

$$= \sum_{s_k} p(s_{k+1} \mid s_k) p(s_k \mid g_{1:k}). \tag{A.2}$$

The posterior $p(s_k \mid g_{1:k})$ in (4.2) can be reformulated as

$$p(s_k \mid g_{1:k}) \propto \exp[-\sum_{\mathbf{x} \in \mathbf{X}} W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) \mid g_{1:k})],$$

$$W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) \mid g_{1:k}) = \sum_{\mathbf{y} \in M_{\mathbf{x}}} \frac{1}{|M_{\mathbf{y}}|} V_{\mathbf{y}}(s_k(\mathbf{y}) \mid g_{1:k}) + \frac{1}{2} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}), s_k(\mathbf{y})), \tag{A.3}$$

where $M'_{\mathbf{x}} = M_{\mathbf{x}} \cup N_{\mathbf{x}}$. It should be noted that $|M_{\mathbf{y}}|$ is not a constant for $\mathbf{y}$ since boundary points in the scene have relatively smaller neighborhood sizes. Combining (4.4), (A.2), and (A.3), the probability $p(s_{k+1} \mid g_{1:k})$ becomes

$$p(s_{k+1} \mid g_{1:k}) \propto \exp[-\frac{1}{2} \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))] \cdot$$

$$\sum_{s_k} \prod_{\mathbf{x} \in \mathbf{X}} \exp[-V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}})) - W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) \mid g_{1:k})]. \tag{A.4}$$

Accurate computation of (A.4) is intractable because all the possible assignments of field s_k should be considered. Since the segmentation field tends to form contiguous regions, the potentials in (A.4) are approximated as

$$V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}})) \propto V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}}) = s_k(\mathbf{x}))$$

$$W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) \mid g_{1:k}) \propto W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) = s_k(\mathbf{x}) \mid g_{1:k}). \tag{A.5}$$

Here for a set M, $s_k(M) = j$ means that $s_k(\mathbf{y}) = j$ for every point $\mathbf{y}$ in the set M. Thus the term in (A.4) becomes

$$\sum_{s_k} \prod_{\mathbf{x} \in \mathbf{X}} \exp[-V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}})) - W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) \mid g_{1:k})]$$

$$\approx \prod_{\mathbf{x} \in \mathbf{X}} \{\sum_j \exp[-\alpha' V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}}) = j) - \lambda_k W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) = j \mid g_{1:k})]\}, \tag{A.6}$$

where $1 \le j \le L$, α' and λ_k are the coefficients for the approximation of the potentials in (A.5). Compared to α', λ_k is assumed to be time varying for the approximation of $W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) \mid g_{1:k})$ since the observed images $g_{1:k}$ increase with time k. Combining (4.5), (A.1), (A.4), and (A.6), the posterior probability distribution of the segmentation field at time k+1 is updated as

$$p(s_{k+1} \mid g_{1:k+1}) \propto p(s_{k+1} \mid g_{1:k}) p(g_{k+1} \mid s_{k+1})$$

$$\propto \exp[-\frac{1}{2} \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))] \prod_{\mathbf{x} \in \mathbf{X}} \{\sum_j \exp[-\alpha' V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}}) = j) -$$

$$\lambda_k W_{\mathbf{x}}(s_k(M'_{\mathbf{x}}) = j \mid g_{1:k})]\} p(\mathbf{o}_{k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x})). \tag{A.7}$$

Denote α_{ij} as $\alpha' V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) = i \mid s_k(M_{\mathbf{x}}) = j)$ and combine (4.3), (A.3), and (A.7), then the posterior distribution at time k+1 can be approximated by a Markov random field with the one-pixel and two-pixel potentials in (4.6).

Appendix B Hypothesized measurements for joint tracking

Using the first order Taylor expansion and ignoring the high order terms, we have that

$$| g_k(\mathbf{d}(\boldsymbol{\theta} + v\mathbf{e}_i, \mathbf{x})) - g_k(\mathbf{d}(\boldsymbol{\theta}, \mathbf{x})) | \propto | v |, \tag{B.1}$$

where v is a small random disturbance in the ith component of the motion vector $\boldsymbol{\theta}$. For the points within the mth object,

$$E[(g_k(\mathbf{d}(\boldsymbol{\theta} + v\mathbf{e}_i, \mathbf{x})) - g_k(\mathbf{d}(\boldsymbol{\theta}, \mathbf{x})))^2] = c^{(m,i)} E[v^2], \tag{B.2}$$

where $\mathbf{x} \in D_m$, and $c^{(m,i)}$ is the proportional factor. $c^{(m,i)}$ can be learned from the reference frame by substituting r for k, $\mathbf{0}$ for $\boldsymbol{\theta}$, and fixing the variable v as 1 in (B.2). Since $\mathbf{d}(\mathbf{0}, \mathbf{x}) = \mathbf{x}$,

$$c^{(m,i)} = E[(g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x}))^2]$$

$$\approx \frac{1}{| D_m |} \sum_{\mathbf{x} \in D_m} [g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x})]^2 . \tag{B.3}$$

From (B.3) we know that $c^{(m,i)}$ is computed as the mean of the squared intensity differences in the reference region.

If the hidden state $\mathbf{z}_k$ is given, the true value of the motion parameters can be considered as $\mathbf{H}\mathbf{z}_k$ in our model. Denote $(\mathbf{H}\mathbf{z}_k)^{(m)}$ as the true motion vector for the mth object. Assume that the intensity distribution remains constant along a motion trajectory, $g_k(\mathbf{d}((\mathbf{H}\mathbf{z}_k)^{(m)}, \mathbf{x}))$ should equal $g_r(\mathbf{x})$ for a visible point of the mth object. Hence, variances of the measurement noise components can be estimated by

substituting $(\mathbf{Hz}_k)^{(m)}$ for $\boldsymbol{\theta}$, and $v_{k,j}^{(m,i)}$ for v in (B.2). Combing with (5.5) under the jth hypothesis,

$$E[(v_{k,j}^{(m,i)})^2] = \frac{1}{c^{(m,i)}} E[(g_k(\mathbf{d}((\mathbf{Hz}_k)^{(m)} + v_{k,j}^{(m,i)}\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x}))^2]$$

$$\approx \frac{1}{c^{(m,i)}} E[(g_k(\mathbf{d}((\mathbf{Hz}_k)^{(m)} + \mathbf{v}_{k,j}^{(m)}, \mathbf{x})) - g_r(\mathbf{x}))^2]$$

$$= \frac{1}{c^{(m,i)}} E[(g_k(\mathbf{d}(\mathbf{y}_{k,j}^{(m)}, \mathbf{x})) - g_r(\mathbf{x}))^2]$$

$$\approx \frac{1}{c^{(m,i)}} e_{k,j}^{(m)} = \frac{|D_m|}{\sum_{\mathbf{x}\in D_m} [g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x})]^2} e_{k,j}^{(m)}. \quad \text{(B.4)}$$

Appendix C The SHM filtering algorithm

Using Bayes' rule, we know that

$$p(s_{k+1}, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k+1})$$

$$= \frac{1}{p(\mathbf{y}_{k+1} \mid \mathbf{y}_{1:k})} p(\mathbf{y}_{k+1} \mid s_{k+1}, \mathbf{z}_{k+1})\, p(s_{k+1}, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k})$$

$$\propto p(\mathbf{y}_{k+1} \mid s_{k+1}, \mathbf{z}_{k+1})\, p(s_{k+1}, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k}). \qquad \text{(C.1)}$$

In principle, the filtering process has three stages: prediction, update, and collapsing.

With the transition probabilities in (5.3) and (5.4), the predictive distribution for time k+1 is computed as

$$p(s_{k+1} = i, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k})$$

$$= \sum_j \int p(s_{k+1} = i, \mathbf{z}_{k+1} \mid s_k = j, \mathbf{z}_k)\; p(s_k = j, \mathbf{z}_k \mid \mathbf{y}_{1:k}) d\mathbf{z}_k$$

$$= \sum_j p(s_{k+1} = i \mid s_k = j) p(s_k = j \mid \mathbf{y}_{1:k}) \int p(\mathbf{z}_{k+1} \mid \mathbf{z}_k) p(\mathbf{z}_k \mid s_k = j, \mathbf{y}_{1:k}) d\mathbf{z}_k$$

$$= \sum_j \alpha_{i,j} \beta_{k,j} \int N(\mathbf{z}_{k+1}; \mathbf{F}\mathbf{z}_k, \mathbf{Q}) N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j}) d\mathbf{z}_k$$

$$= \sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1|k,j}, \mathbf{P}_{k+1|k,j}) . \qquad \text{(C.2)}$$

After receiving the measurement set $\mathbf{y}_{k+1}$ at time k+1, the posterior density is updated as follows,

$$p(s_{k+1} = i, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k+1})$$

$$\propto p(\mathbf{y}_{k+1} \mid s_{k+1} = i, \mathbf{z}_{k+1}) p(s_{k+1} = i, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k})$$

$$\propto \sum_j \alpha_{i,j}\beta_{k,j}\ N(\mathbf{y}_{k+1,i};\mathbf{H}\mathbf{z}_{k+1},\mathbf{R}_{k+1,i})\ N(\mathbf{z}_{k+1};\mathbf{m}_{k+1|k,j},\mathbf{P}_{k+1|k,j})\,. \tag{C.3}$$

If the covariances in $\mathbf{P}_{k+1|k,j}$ are small [1], the product in (C.3) can be approximated by

$$N(\mathbf{y}_{k+1,i};\mathbf{H}\mathbf{z}_{k+1},\mathbf{R}_{k+1,i})N(\mathbf{z}_{k+1};\mathbf{m}_{k+1|k,j},\mathbf{P}_{k+1|k,j})$$

$$\approx N(\mathbf{y}_{k+1,i};\mathbf{H}\mathbf{m}_{k+1|k,j},\mathbf{S}_{k+1,i|j})N(\mathbf{z}_{k+1};\mathbf{m}_{k+1,i|j},\mathbf{P}_{k+1,i|j})\,. \tag{C.4}$$

The conditional probability of the switching state is updated as

$$\beta_{k+1,i} = p(s_{k+1} = i \mid \mathbf{y}_{1:k+1})$$

$$= \int p(s_{k+1} = i, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k+1})d\mathbf{z}_{k+1}$$

$$\propto \sum_j \alpha_{i,j}\beta_{k,j}\ N(\mathbf{y}_{k+1,i};\mathbf{H}\mathbf{m}_{k+1|k,j},\mathbf{S}_{k+1,i|j})\,. \tag{C.5}$$

Since $\sum_i \beta_{k+1,i} = 1$, (5.9) can be obtained by normalizing. From (C.3) – (C.5), the pdf $p(\mathbf{z}_{k+1} \mid s_{k+1} = i, \mathbf{y}_{1:k+1})$ becomes a mixture of L Gaussians.

$$p(\mathbf{z}_{k+1} \mid s_{k+1}= i, \mathbf{y}_{1:k+1})$$

$$= \sum_j \beta_{k+1,i|j} N(\mathbf{z}_{k+1};\mathbf{m}_{k+1,i|j},\mathbf{P}_{k+1,i|j})\,. \tag{C.6}$$

It could be derived that

$$p(\mathbf{z}_{k+1} \mid s_{k+1}=i, s_k = j, \mathbf{y}_{1:k+1})$$

$$= N(\mathbf{z}_{k+1};\mathbf{m}_{k+1,i|j},\mathbf{P}_{k+1,i|j})\,. \tag{C.7}$$

At time k, the distribution $p(\mathbf{z}_k \mid \mathbf{y}_{1:k})$ is represented as a mixture of L Gaussians, one for each hypothesis of s_k. Then each Gaussian is propagated through state transition,

so that $p(\mathbf{z}_{k+1} \mid \mathbf{y}_{1:k+1})$ will be a mixture of L^2 Gaussians. The number of Gaussians grows exponentially with time. To deal with this problem, the mixture of Gaussians in (C.6) is collapsed to a single Gaussian in (5.10) using moment matching [42]. Collapsing is processed under each hypothesis of s_{k+1}. Therefore, the possibility of each hypothesis will not be cast throughout the propagation.

References

[1] B. D. O. Anderson and J. B. Moore, *Optimal filtering*, Prentice-Hall, 1979.

[2] S. Ayer and H. S. Sawhney, "Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding," *Proc. Int'l Conf. Computer Vision*, pp. 777-784, 1995.

[3] Y. Bar-Shalom and T. E. Fortmann, *Tracking and data association*, Academic Press, 1988.

[4] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Statist. Soc. B*, vol. 48, pp. 259-302, 1986.

[5] G. Borgefors, "Distance transformation in digital images," *Computer Vision, Graphics, and Image Processing*, vol. 34, pp. 344-371, 1986.

[6] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann, "Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings," *Proc. IEEE*, vol. 89, pp. 1382-1402, 2001.

[7] R. G. Brown, *Introduction to random signal analysis and Kalman filtering*, John Wiley & Sons, 1983.

[8] T.-J. Cham and J. M. Rehg, "A multiple hypothesis approach to figure tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 239–245, 1999.

[9] M. M. Chang, A. M. Tekalp, and M. I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. Image Processing*, vol. 6, pp. 1326-1333, 1997.

[10] P. B. Chou and C. M. Brown, "The theory and practice of Bayesian image labeling," *Int'l J. Computer Vision*, vol. 4, pp. 185-210, 1990.

[11] I. J. Cox and S. L. Hingorani, "An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 18, pp. 138–150, 1996.

[12] S. L. Dockstader and A. M. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proc. IEEE*, vol. 89, pp. 1441-1455, 2001.

[13] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, pp. 1151-1163, 2002.

[14] P. A. Flach, "On the state of the art in machine learning: A personal review," *Artificial Intelligence*, vol. 131, pp. 199-222, 2001.

[15] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *Proc. Conf. Uncertainty in Artificial Intelligence*, pp. 175-181, 1997.

[16] Y. Fu, A. T. Erdem, and A. M. Tekalp, "Tracking visible boundary of objects using occlusion adaptive motion snake," *IEEE Trans. Image Processing*, vol. 9, pp. 2051–2060, 2000.

[17] B. Galvin, B. McCane, and K. Novins, "Virtual snakes for occlusion analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 294–299, 1999.

[18] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 6, pp. 721-741, 1984.

[19] M. Gerlgon, and P. Bouthemy, "A region-level graph labeling approach to motion-based segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 514-519, 1997.

[20] Z. Ghahramani, "Learning dynamic Bayesian networks," in *Adaptive processing of temporal information*, Lecture notes in artificial intelligence, pp. 168–197, Springer-Verlag, 1998.

[21] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, pp. 963–996, 1998.

[22] G. Gordon, T. Darrell, M. Harville, and J. Woodfill, "Background estimation and removal based on range and color," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 459-464, 1999.

[23] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 20, pp. 1025–1039, 1998.

[24] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 809-830, 2000.

[25] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," *Proc. FRAME-RATE Workshop*, 1999.

[26] M. Isard and A. Blake, "A mixed-state Condensation tracker with automatic model-switching," *Proc. Int'l Conf. Computer Vision*, pp. 107–112, 1998.

[27] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," *Proc. European Conf. Computer Vision*, pp. 343–356, 1996.

[28] Y. Ivanov, A. Bobick, and J. Liu, "Fast light independent background subtraction," *Int'l. J. Computer Vision*, vol. 37, pp. 199-207, 2000.

[29] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Detection and location of people in video images using adaptive fusion of color and edge information," *Proc. Int'l Conf. Pattern Recognition*, vol. 4, pp. 627-630, 2000.

[30] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer-Verlag, 2001.

[31] A. D. Jepson, D. J. Fleet, and M. J. Black, "A layered motion representation with occlusion and compact spatial support," *Proc. European Conf. Computer Vision*, pp. 692-706, 2002.

[32] N. Jojic and B. J. Frey, "Learning flexible sprites in video layers," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 199-206, 2001.

[33] M. I. Jordan (Ed.), *Learning in graphical models*, MIT Press, 1999.

[34] S. Kamijo, K. Ikeuchi, and M. Sakauchi, "Segmentations of spatio-temporal images by spatio-temporal Markov random field model," *Proc. EMMCVPR Workshop*, pp. 298-313, 2001.

[35] C.-J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.

[36] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, "Towards robust automatic traffic scene analysis in real-time," *Proc. Int'l Conf. Pattern Recognition*, vol. 1, pp. 126-131, 1994.

[37] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer-Verlag, 2001.

[38] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 572–578, 1999.

[39] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, pp. 42-56, 2000.

[40] I. Mikic, P. C. Cosman, G. T. Kogut, and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," *Proc. Int'l Conf. Pattern Recognition*, vol. 1, pp. 321-324, 2000.

[41] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 20, pp. 897-915, 1998.

[42] K. P. Murphy, "Learning switching Kalman filter models," Technical Report 98-10, Compaq Cambridge Research Lab, 1998.

[43] H. T. Nguyen, M. Worring, and R. van den Boomgaard, "Occlusion robust adaptive template tracking," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 678–683, 2001.

[44] T. N. Papps, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Image Processing*, vol. 4, pp. 901-914, 1992.

[45] N. Paragios and V. Ramesh, "A MRF-based approach for real-time subway monitoring," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 1034-1040, 2001.

[46] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by MAP labeling of watershed segments," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 326-332, 2001.

[47] V. Pavlovic and J. M. Rehg, "Impact of dynamic model learning on classification of human motion," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 788–795, 2000.

[48] V. Pavlovic, J. M. Rehg, T.-J. Cham, and K. P. Murphy, "A dynamic Bayesian network approach to figure tracking using learned dynamic models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 94–101, 1999.

[49] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 25, pp. 918-923, 2003.

[50] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257-286, 1989.

[51] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 560–576, 2001.

[52] J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," *Proc. European Conf. Computer Vision*, vol. 2, pp. 336-350, 2000.

[53] K. Rohr, "Towards model-based recognition of human movements in image sequences," *Computer Vision, Graphics, and Image Processing: Image Understanding*, vol. 59, pp. 94–115, 1994.

[54] M. Seki, T. Wada, H. Fujiwara, and K. Sumi, "Background subtraction based on cooccurrence of image variations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 65-72, 2003.

[55] R. H. Shumway and D. S. Stoffer, "Dynamic linear models with switching," *Journal of the American Statistical Association*, vol. 86, pp. 763–769, 1991.

[56] J. Stauder, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation," *IEEE Trans. Multimedia*, vol. 1, pp. 65-76, 1999.

[57] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 747-757, 2000.

[58] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. M. Buhmann, "Topology free hidden Markov models: Application to background modeling," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 294-301, 2001.

[59] C. Stiller, "Object-based estimation of dense motion fields," *IEEE Trans. Image Processing*, vol. 6, pp. 234-250, 1997.

[60] H. Tao, H. S. Sawhney, and R. Kumar, "Object tracking with Bayesian estimation of dynamic layer representations," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 24, pp. 75–89, 2002.

[61] A. M. Tekalp, *Digital Video Processing*, Prentice Hall, 1995.

[62] P. H. S. Torr, R. Szeliski, and P. Anandan, "An integrated Bayesian approach to layer extraction from image sequences," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 297-303, 2001.

[63] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 255-261, 1999.

[64] Y. Tsaig and A. Averbuch, "Automatic segmentation of moving objects in video sequences: A region labeling approach," *IEEE Trans. Circuit Sys. Video Technol.*, vol. 12, pp. 597-612, 2002.

[65] N. Vasconcelos and A. Lippman, "Empirical Bayesian motion segmentation," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 217-221, 2001.

[66] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Processing*, vol. 3, pp. 625-637, 1994.

[67] Y. Wang, K.-F. Loe, T. Tan, and J.-K. Wu, "A dynamic hidden Markov random field model for foreground and shadow segmentation," *Proc. IEEE Workshop on Applications of Computer Vision*, 2005, in press.

[68] Y. Wang, T. Tan, and K.-F. Loe, "Joint region tracking with switching hypothesized measurements," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 75-82, 2003.

[69] Y. Wang, K.-F. Loe, T. Tan, and J.-K. Wu, "Spatio-temporal video segmentation based on graphical models," *IEEE Trans. Image Processing*, in press.

[70] Y. Wang, T. Tan, and K.-F. Loe, "Switching hypothesized measurements: A dynamic model with applications to occlusion adaptive joint tracking," *Proc. Int'l Joint Conf. Artificial Intelligence*, pp. 1326-1331, 2003.

[71] Y. Wang, T. Tan, and K.-F. Loe, "Video segmentation based on graphical models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 335-342, 2003.

[72] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 19, pp. 780-785, 1997.