HUMAN DETECTION, TRACKING AND SEGMENTATION FROM LOW-LEVEL TO HIGH-LEVEL VISION

By

CHENG CHEN

Bachelor of Science in Mechanical Engineering
Shenyang Institute of Technology
Shenyang, Liaoning, P.R.China
1992

Master of Science in Mechatronics
University of Electronic Science and Technology of China
Chengdu, Sichuan, P.R.China
1997

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2008

HUMAN DETECTION, TRACKING AND SEGMENTATION
FROM LOW-LEVEL TO HIGH-LEVEL VISION

Dissertation Approved:

Guoliang Fan

Dissertation Advisor

Damon Chandler

Martin Hagan

Jiahong Wu

A. Gordon Emslie

Dean of the Graduate College

ACKNOWLEDGMENTS

At this point I would like to thank all the people who supported my dissertation. I wish to thank the members of my committee for their support, patience. Their gentle but first direction has been most appreciated. Dr. Hagan was particularly helpful in guiding me to enter the field of pattern recognition. Dr. Wu helped me build a solid knowledge foundation in advanced linear algebra, which provides the tools to explore present ideas and issue. I am also very appreciate Dr. Chandlers encouragement. I would like to express my gratitude to my supervisors, Professor Guoliang Fan for kindly supporting this work in many ways. Without his helpful suggestions, precise reviews, and motivation, this work would have been impossible. From the beginning he had confidence in my ability to not only complete a degree, but to complete it with excellence.

I am very grateful to my group members at the Visual Computing and Image Processing Lab (VCIPL) at Oklahoma State University (OSU). Especially, I am want to thank Stephen Nilson for his hard working in collecting ground truth data to support my research.

Finally, I would like to thank my family for their love and support. I cannot thank them enough for their everlasting love, support, and understanding. My Ph.D. work would be impossible without their support behind me.

This work was supported by the National Science Foundation under NSF CAREER Award IIS-0347613

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

1.1 Motivation

Recently, there has been a surge in the global need for robust and intelligent surveillance systems. For instance, according to the world's largest market research resource [6], the world market for network video surveillance products reached $2 billion in 2006, and is forecast to continually grow by at least 40% for 5 years. Since most video surveillance systems are used to monitor human objects, automated human detection, tracking and segmentation are the key desired function components of the new generation video surveillance systems.

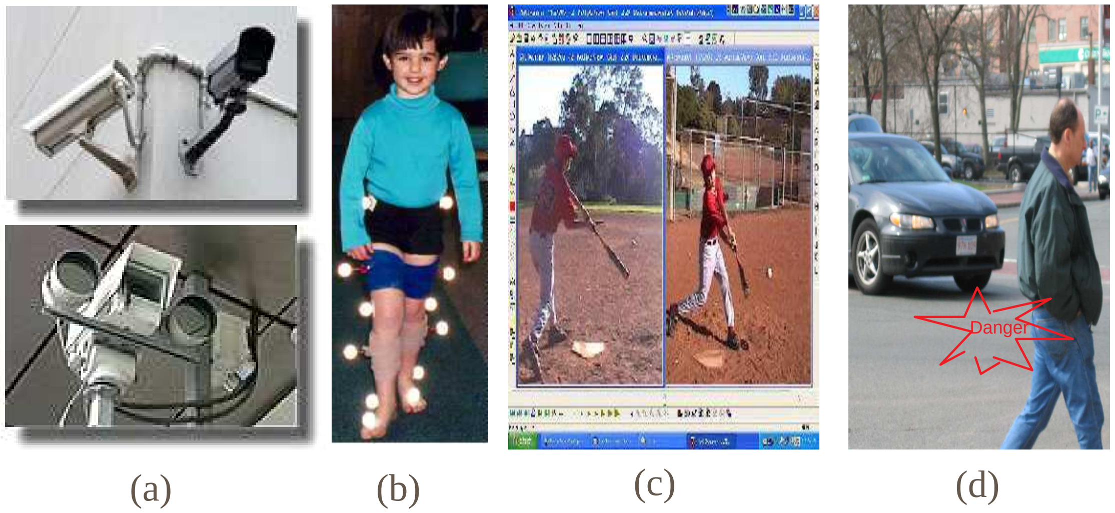


Figure 1.1: Application examples for automatic human detection, tracking and segmentation. (a)Surveillance systems. (b) Medical rehabilitation. Photo from [1]. (c) Sport analysis. (d) Driver assistance

Human detection, tracking and segmentation are also proving to be invaluable in athletics and medical rehabilitation. For example, they can provide a coach with more in-depth sport video analysis by extracting biomechanics information (as shown in Fig 1.1 c.). This information can be used to improve coaching techniques and athlete performance. Lets see another example, in a motion analysis clinic, a typical gait evaluation often takes a patient about 2 to 2.5 hours for data collection, such as changing into tight-fitting shorts, and being taped with many reflective markers, which are placed at specific anatomic locations (as shown is the Fig 1.1 b.). It usually takes 48 hours to obtain the basic gait analysis results, and at least four weeks to obtain the full gait analysis report [1]. Automated video-based human motion analysis algorithms can quickly provide accurate motion and gait information, which will make the whole diagnosis procedure much simple.

Human detection, tracking and segmentation can also have many other important applications such as vision-based Human-Computer-Interface (HCI), robotic vision, driver assistance (as shown in Fig 1.1 d.), vehicle navigation, animation, and so on. Huge potential business opportunities have elicited significant interests from both industry and academia. Automated video-based human detection, tracking and segmentation have received intensive studies in the last decade.

1.2 Research Goals and Challenges

The objective of this research is to investigate how to achieve automated and robust human detection tracking and segmentation. In a general definition, human detection is to find out whether or not human objects exist in a given image or video sequence; while human segmentation is to identify these pixels belonging to human objects. In our research, the definition of human segmentation is to detect and segment a human body as well as identify its limbs from a given image or a video sequence. In order to obtain continuous information about human detection and segmentation in a video sequence, tracking is an often used tool to facilitate the information extraction processing by incorporating the temporal context information in previous frames.

Today, an automated algorithm that can achieve robust human body detection, tracking and segmentation from generic scenes does not exist. After many years intensive studying by computer vision researchers, they remain to be ones of the most challenging research issues largely due to the ubiquitous visual ambiguities in images/videos. The other challenging factor is the ill-posed nature of the problems. In general, the raw data to a computer vision system can only be low-level signals, such as intensity or color. Sarkar and Boyer [7] classify features into four categories: the signal level, the primitive level, the structural level and the assembly level. We know that low-level physical features alone, such as color information, is not enough to represent human appearance directly. The question is how to extract intermediate features to bridge the gap between low-level features and a human object representation. The second question is how to represent and incorporate prior knowledge into high-level inference processing. We believe that these two questions should be well addressed in a success human detection, tracking and segmentation approach.

1.3 Objectives and Methodology

In the instructional paper "Why progress in machine vision is so slow", after pointing out several impediments, Pavlidis [8] gave two suggestions for current computer vision researchers. The first suggestion was to develop algorithms based on a functional understanding of the Human Vision System(HVS). Since HVS can easily partition video scenes into meaningful objects and recognize them effortlessly, we believe some hints from biological vision studies can help us to attack these mentioned challenges in computer vision. Therefore, we study human vision system (HVS) first and attempt to find some hints to guide our algorithm designing. The second suggestion was to solve very specific vision problem, which has limited context. These two suggestions can help us avoid several impediments in developing computer vision algorithms.

Following Pavlidis' first suggestion, we designed our algorithms according to the state of the art research results of cognitive psychology. According to perception organization theories [2][9][10] [11], visual perception is as a result of complex cascade part-whole hierarchy organization of visual information that involves both low-level and mid-level visions[12]. Low-level vision performs grouping processing and provides intermediate elements and features without the influence of specific domain knowledge. Once these intermediate elements and features have been constructed, they are submitted to the so-called figure-ground process, which is one of mid-level vision processes. Figure-ground process throws out the unwanted image components while keeps the relevant elements you care about for your task. The elements identified as figures are further grouped into more complex visual entities, by the action of at least some of the classical Gestalt laws such as: common fate, proximity, closure, similarity. These processes is called "mid-level vision". The cascade part-whole grouping processes in low-level and mid-level vision are often called "bottom-up" processes. At the end, semantic "understanding" of a scene can be achieved by high-level vision processes, which use both high-level prior knowledge and the provided information from mid-level vision.

According to our functional understanding of the Human Vision System(HVS), we design our three level part-whole cascaded algorithm structure and define the algorithm objectives of each level processing as shown in the Fig 1.2.

In low-level vision, our objective is to obtain compact image representation by grouping raw pixels into homogeneous regions. The criteria for homogeneousness may be in the measurement of color, motion, or intensity. In order to bridge the gap between low-level features and a human appearance representation, we expect to obtain more meaningful homogeneous regions, which can serve as building blocks for higher-level feature extraction. Therefore, both under-segmentation (one segmented

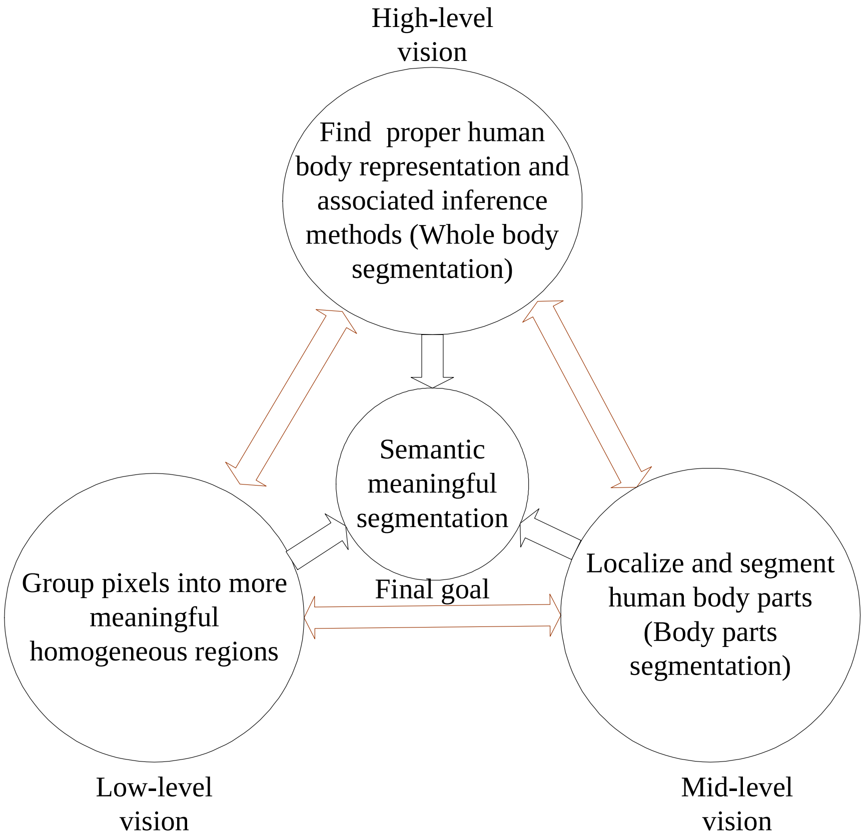

High-level vision
Find proper human body representation and associated inference methods (Whole body segmentation)
Semantic meaningful segmentation
Final goal
Group pixels into more meaningful homogeneous regions
Localize and segment human body parts (Body parts segmentation)
Low-level vision
Mid-level vision

Figure 1.2: Objectives of each level processes and their goal.

region has pixels belonging to different objects) and over-segmentation (one object is segmented into too many small pieces) are detrimental to our goal. Considering that one human figure may have different colors and one color may be shared by both background and human objects, how to balance over-segmentation and under-segmentation is a very difficult task for a low-level feature-based classifier. Our analysis shows that the problems of over-segmentation and under-segmentation relate to the two kinds of no-convex classification problems for a single layer classifier. Therefore, we extended a single-layer statistical model video segmentation algorithm into a cascaded multi-layer classification framework. Using a split-and-merge paradigm, we extracted mid-level region-based motion and color features to deal with the no-convex classification problems, moveover, more meaningful segmentation results are obtained with less over-segmentation and under-segmentation.

Numerous low-level-feature-based bottom-up approaches have tried to fill the gap between the low-level features and high-level knowledge representation, and have reached the performance ceiling where there seems little room for improvement. It seems that this gap can not be filled directly. Instead, it should have an intermediate step, middle-level vision processing, which will act as the bridge between low and high-level vision operations. In our work, the objective of mid-level vision processing is to localize and segment human body parts. For each body part, the desired output is a *map* image that indicates the likelihood that a body part will be at a given location. These *map* images will be the inputs for the high-level processing.

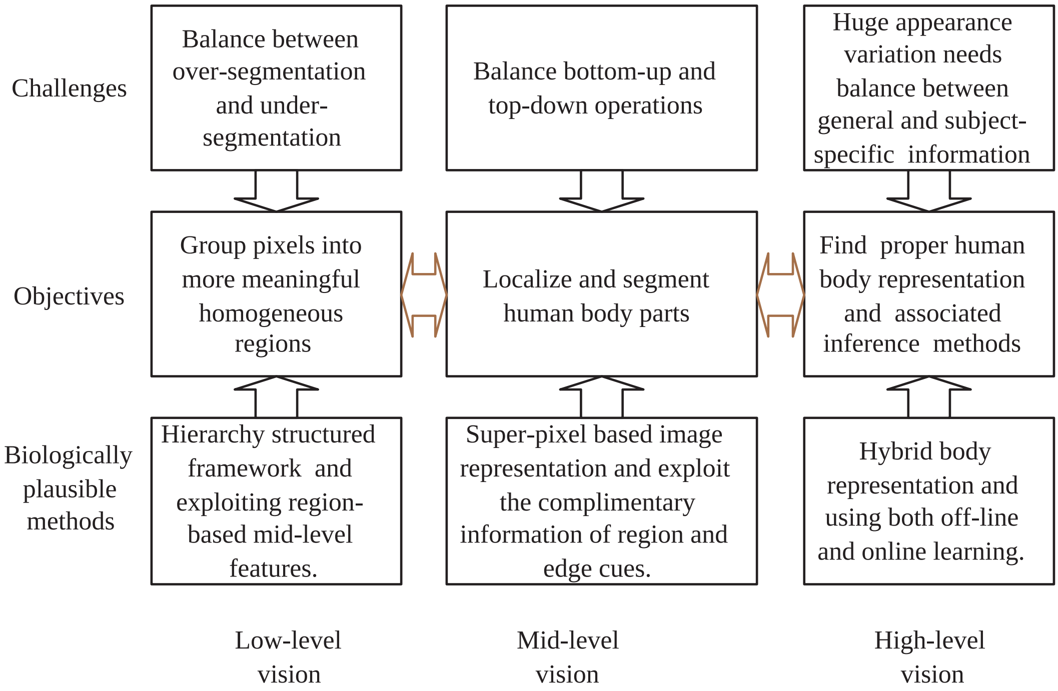
For high-level vision processing, considering the huge variation of the appearance of a human object, how to represent it in a way that can be easily understood by computers is a crucial and challenging problem. Minsky and Papert [13] pointed out

> *"No machine can learn to recognize X unless it possesses, at least potentially, some scheme for representing X." (p. xiii).*

Therefore, the objective of high-level vision processing is to find a proper prior knowledge representation and incorporate different priors to assembly middle-level outputs into a final recognition(decision) by inference.

We not only define the objectives of each level processing according to biological perception theories, but we also look for biologically plausible methods to fulfill these objectives for each level processing as shown in the Fig 1.3.

- **Low-level vision:** In low-level vision, guided by perception principles, we studied feature extraction problem in a bottom-up, low-level video segmentation process. Two kinds of non-convex video segmentation problems (related to the balance between over-segmentation and under-segmentation) can be solved in a hierarchy multi-layers classification framework, which is deeply inspired by the structured perception theory of cognitive physiology.



Challenges
Balance between over-segmentation and under-segmentation
Balance bottom-up and top-down operations
Huge appearance variation needs balance between general and subject-specific information
Objectives
Group pixels into more meaningful homogeneous regions
Localize and segment human body parts
Find proper human body representation and associated inference methods
Biologically plausible methods
Hierarchy structured framework and exploiting region-based mid-level features.
Super-pixel based image representation and exploit the complimentary information of region and edge cues.
Hybrid body representation and using both off-line and online learning.
Low-level vision
Mid-level vision
High-level vision

Figure 1.3: Objectives, challenges, and biologically plausible methods of three level processes.

- **Mid-level vision:** In middle-level vision, inspired by cognitive studies about region and edge cues, we investigated how to use the complimentary information of region and edge cues to a combined bottom-up and top-down approach. The success of this framework depends on the adoption of a super-pixel based representation strategy, which is supported by a representation element theory of cognitive psychology.

- **High-level vision:** In high-level vision, we studied representation problem for high-level computer vision. By combining the advantage of two kinds of shape representation theories in cognitive psychology, we introduced a hybrid human pose representation, which supports a joint localization segmentation and pose estimation framework. This framework can achieve significant improvement in both localization and segmentation compared with some state of the art algorithms. In chapter 6, inspired by the two perception pathways biological movement perception model [14, 15], there are two separate functional streams involved in vision perception: a ventral stream for the analysis of form (the “what” stream) and a dorsal stream for the analysis of position and motion (the “where” stream), we investigate combine both spatial prior and temporal together for articulated human tracking.

Following Pavlidis’ second suggestion, we do not attempt to look for a silver bullet that will solve human segmentation in a general scene. For example, we do not try to build one human body representation model which can be applied into the segmentation of images that are taken from different view points or/and at different scales. Instead, we attempt to investigate some fundamental problems of human detection, tracking and segmentation in images/videos that are taken from one view point.

1.4 Contributions

Before the outline of each chapter is given, we would like to summarize the main contributions we have made. This dissertation is based on two journal article drafts, three conference articles, and a technical report. There were certain motivations and contributions at the times each topic was investigated. First, we will introduce our contribution in a biologically plausible computational model for human object detection, tracking and segmentation. Then, we will introduce our contribution in low-level, mid-level and high-level vision guided by this computational model.

1.4.1 Comprehensive Computational Model

The purpose is to get some hints from biological vision studies to attack our problems. Current research about biological vision usually concentrates on only certain aspects, e.g., attention, motion perception, visual memory, low-level perception organization.

We know that the HVS is a well organized system and different perception rules need to cooperate together. However, very little effort has been made to develop a comprehensive computational model for the HVS. Moreover, biological vision studies are still at their infancy stage. These facts make it difficult to apply perception principles systematically in practice. In this work, we develop a comprehensive computational model for human motion segmentation, which allows us to gain more insights to this challenging problem.

1.4.2 Bottom-up Segmentation

For low-level and mid-level bottom-up processes, our purpose is to get compact image representation. We extend single-layer statistical model video segmentation algorithm into a cascaded multi-layer classification framework, which combines the merits of both statistical modelling and graph theory approaches. Using a split-and-merge paradigm, we extracted mid-level region-based motion and color features to deal with the no-convex classification problems, moveover, more meaningful segmentation results are obtained with less over-segmentation and under-segmentation.

1.4.3 Middle-level Vision: Part Detection

For mid-level combined bottom-up and top-down process, our purpose is to group small UC regions into more semantic meaningful regions, such as body parts, and get a confidence map for each body part. We have made three contributions.

- **An effective hypothesis-and-test paradigm:** We have developed an effective hypothesis-and-test paradigm for joint localization and segmentation. Additionally, it is able to provide a posterior density map of localization, which can support various high-level processes.

- **A new semi-parametric approach for color model online learning and figure-ground segmentation:** Based on super-pixel based image representation, we propose a new semi-parametric approach for fast online learning of figure-ground color models aided by the region-based shape prior.

- **An improved Graph-cut based segmentation method:** In our work, both region-based and edge-based shape priors are integrated into an improved Graph-cut based framework to achieve optimal segmentation. To the best of our knowledge, no research has been done on integrating both edge and region priors into the Graphic-cut framework for automated segmentation.

1.4.4 High-level Vision: Recognition, Localization and Segmentation

In Image-based high-level computer vision processing, our purpose is to make a comprehensive decision about the position of each body part by assembling map images according to offline learned spatial priors. We have contributions in develop hybrid representation for integrated pose recognition, localization and segmentation:

- **A hybrid human body representation:** A hybrid human body representation supports the online color model learning. The online learned deformable shape model can facilitate the segmentation of the whole body and parts. The proposed representation absorbs recent multifaceted advances in this field and involves shape prior guided segmentation and inference in a multi-stage fashion.

- **A three-stage cascade computational flow:** A three-stage cascade computational flow integrates pose recognition, localization and segmentation into a "biologically plausible" dynamic framework, in which low-level and middle vision parameters can online dynamically adjusted according to feedback information from high-level vision.

1.4.5 High-level Vision: Tracking and Localization

The purpose of video-based high-level computer vision processing is to integrates both spatial and temporal priors and is supported by online learning. We extend our success from image-based to video-based processing by exploiting the complementary context information in both temporal priors and spatial priors.

- **Local online learning:** In our work, "Back constraints" Gaussian process latent variable model BC-GPLVM is used to online learn a compact low dimension representation of motion trajectory in the latent space and a probabilistic reverse mapping from the low-dimension latent space to the high-dimension pose space. Online learning is more favorable and effective to deal with human motion with significant variability or even different activities

- **Combing both temporal and spatial priors:** To the best of our knowledge, there is no prior research on how to combine spatial and temporal priors in an online learning framework. The strength of our method comes from the marriage of two popular mathematical tools: GPLVM and pictorial structure graph model in an online learning context. This marriage bring complementary benefits to both sides. GPLVM brings in top-down temporal constraints and model parameters for a star-structured graph model; Star-structured graph model brings in spatial constraints for assembling bottom-up data-driven information, which will correct top-down predictions.

It is worth noting that among numerous approaches for advanced human detection, some use segmentation in their processes, while the others do not to do so. As compared with human detection, which has made significant progress over the last few years, not much progress has been made in human segmentation, and the role of segmentation has largely been ignored. In many human object related studies, such as human pose estimation and human tracking, segmentation problem is often circumambulated by assuming its results are already available, or assuming segmentation results can be obtained by a background subtraction process. This kind of assumption ignored an important fact that segmentation can be an important tool to support human object analysis in many aspects from low-level feature extraction, to mid-level human part detection, and to high-level knowledge representation and inference. Our study focuses on an unified framework for human detection, tracking and segmentation. One of the main distinguishing characteristics of our work is the role played by segmentation. Segmentation is not only a goal but also a tool in our work. In other words, our research investigates not only how to do segmentation but also how it can help us in three level vision processes, from low-level feature extraction, to mid-level part detection, and to high-level knowledge representation and inference as shown in the Fig. 1.2.

1.5 Outline

In order to provide readers better understandings of the materials and subjects investigated in this dissertation, we use this section to provide the general ideas covered in this report. The organization of this report is illustrated in Fig.1.4.

- The motivation and significance of this research as well as methodology are presented in Chapter 1.

- Currently related biological vision studies are reviewed and categorized in the chapter 2. Specifically, we review recent biological vision studies that are related to *human motion segmentation.* Our goal is to develop a practical and biologically plausible computational framework for the segmentation of human body in a video sequence. Specifically, we discuss the roles and interactions of bottom-up and top-down processes in visual perception processing as well as how to combine them synergistically in one computational model to guide

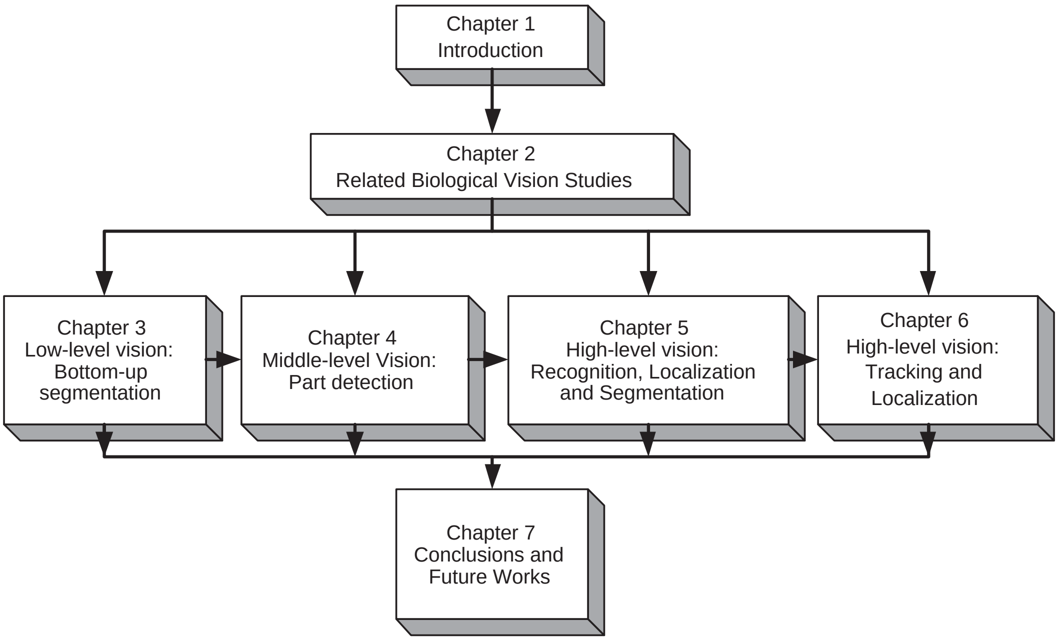


Figure 1.4: Outline of the report.

human motion segmentation. We also examine recent research on biological movement perception, such as neural mechanisms and functionalities for biological movement recognition and two major psychological tracking theories. We attempt to develop a comprehensive computational model that involves both bottom-up and top-down processing and is deeply inspired by biological motion perception. According to this model, object segmentation, motion estimation, and action recognition are results of recurrent feed-forward (bottom-up) and feedback (top-down) processes. Some open technical questions are also raised and discussed for future research.

- In Chapter 3, our research focuses on bottom-up low-level and mid-level segmentation and feature extraction problems, such as joint spatial-temporal grouping, short/middle range motion feature extraction and grouping architecture. We presented a perception principle guided unsupervised video segmentation framework, which combine the merits of statistical modelling and graph theory approach into a multi-stage classification architecture. Our simulation results verify that our new framework can achieve a more meaningful segmentation result in some complex and realistic scenarios. It is also computationally effective.

- In Chapter 4, our research focuses on combined bottom-up and top-down mid-level vision problems. We investigate how to apply the complementary information of region and edge for the shape prior constrained figure-ground segmentation. We formulate configuration estimation and figure-ground segmentation as a MAP estimation in a Bayesian framework. In order to solve the optimization problem, we resort to a segmentation-based hypothesis-and-test paradigm, in which a balance between bottom-up and top-down processing is achieved by exploiting the complementary information of region-based and edge-based shape prior. Specifically, the shape priors are represented by an implicit shape model, which unifies the representation of both region-based and edge-based shape prior. Given a configuration hypothesis, the region-based shape prior is used to guide a bottom-up segmentation. The edge-based shape prior is used to evaluate the obtained segmentation result as well as a configuration hypothesis. In this way, a correct localization will facilitate object segmentation, and a good segmentation will enhance the confidence of a localization hypotheses. The optimal segmentation and the spatial configuration can be obtained simultaneously. The obtained segmentation result is further refined through an improved Graph-cut based method, in which both region-based and edge-based shape priors are jointly involved. Our experiments demonstrate that this framework leads to significant localization and segmentation performance improvements over some state-of-the-art approaches.

- In Chapter 5, our research focuses on high-level knowledge-based human body representation problems. We propose a hybrid body representation for integrated pose recognition, localization and segmentation of the whole body as well as body parts in a single image. A typical pose is represented by both template-like view information and part-based structural information. Specifically, each body part as well as the whole body are represented by an off-line learned shape model where both region-based and edge-based priors are combined in a coupled shape representation. Part-based spatial priors are represented by a "star" graphical model. This hybrid body representation can synergistically integrate pose recognition, localization and segmentation into one computational flow. Moreover, as an important step for feature extraction and model inference, segmentation is involved in the low-level, mid-level and high-level vision.

- In Chapter 6, We integrate spatial and temporal priors for tracking an articulated human body from a monocular video sequence, where body parts can be localized and segmented simultaneously. The spatial prior is represented by a star-structured graphical model that is embedded in the temporal prior. The temporal prior is represented by a motion trajectory in a low dimensional latent space learnt from previous tracking results. The temporal prior predicts the location of each body part, and the spatial prior is used to evaluate and correct the prediction by assembling part-level detection. Both temporal and spatial priors can be online learned in a seamless fashion through the Back Constrained Gaussian Process Latent Variable Model (BC-GPLVM) that involves a moving window for training sample selection. Experimental results show that the new algorithm can achieve accurate tracking and localization results for different walking subjects with significant appearance and motion variability.

- Based on our current research results in low-level, mid-level and high-level vision, Chapter 7 states future works and concludes this dissertation.

CHAPTER 2

Related Biological Vision Studies

2.1 Overview

The human vision system (HVS) can easily partition video scenes into meaningful objects and recognize them effortlessly. We believe that hints and inspirations for reliable human detection, tracking and segmentation lie in examining the processes used in many successful biological vision systems. However. biological vision studies are still at their infancy stage. Current research about biological vision usually concentrates on only certain aspect, e.g., attention, motion perception, visual memory, low-level perception organization. Although we know that the human visual system (HVS) is a well organized system, different perception rules need to cooperate together, very little effort has been made to develop a comprehensive computational model for HVS. This fact makes it difficult to apply perception principles systematically in computer vision. Still several well-understood perception principles have been adopted in the development of computer vision algorithms, e.g., semantic video segmentation. In this work, we will review two types of biological vision studies, i.e., general perception and biological movement perception, based on which we attempt to develop a comprehensive computational model for human detection, tracking and segmentation. This work allows us to gain more insights to this challenging problem. Under the proposed framework, several open questions are raised and discussed for future research. Some of them have already started to attract researchers' attention recently.

2.2 General Perception Principles

Since the early 20^{th} century, psychologists have found a set of rules that govern the HVS. In this section, we will first introduce several important perception principles, which are closely related to the bottom-up process for video segmentation. We then develop a computational model and deduce some guidelines for feature selection and classifier design. We will also discuss the top-down process in the HVS that involves high-level knowledge or prior information for visual inferencing. State-of-the-art video segmentation algorithms have been inspired and motivated by these vision studies to some extent.

2.2.1 Joint Spatial-temporal Grouping Theory

Cognitive scientistic research shows that spatial and temporal groupings are jointly involved in the HVS [10, 16, 11, 17]. In other words, human vision recognizes salient objects in space and time simultaneously. Specifically, spatial grouping is a process that merges spatial samples to form more complex visual entities, e.g., objects. Temporal grouping is a process where visual entities are linked over time. The successive visual entities that undergo a series of spatiotemporal groupings are called *matching units* or *correspondence tokens* by Ullman [18], which are 3D volumes in space and time.

Recently, Gepshtein and Kubovy suggested that the perception of a dynamic scenes is the result of a parallel operation of spatial and temporal grouping [10]. Visual information is sampled by small receptors in HVS. Spatial grouping is a process that link samples across space to form more complex visual entities, such as objects and surfaces. Temporal grouping is a process , by which visual entities are linked over time. The successive visual entities that undergo a series spatiotemporal grouping are called *matching units* or *correspondence tokens* by ullman[18], which are 3D volumes in space and time. Matching units is the result of joint spatial and temporal grouping process, in which spatial organization and motion matching are tightly integrated. There are also neuropsychological evidences indicating that segmentation in space and time is an integrated function mediated by the posterior parietal cortex [17].

2.2.2 Perception Organization Theory

The study about perception organization can be traced back to the Gestalt school of psychology in the early 20^{th} century, and expanded by Marr [19], Palmer and Rock [2], Palmer [9], Kubovy and Gepshtein [10, 11]. According to perception organization theory, visual perception is as a result of complex cascade part–whole hierarchy organization of visual information that involves both low-level and middle-level vision [12], as discussed in the following.

- **Low-level Vision** Visual information is first sampled by small receptors: photoreceptor neurons, then, they are first grouped into small intermediate elements, which are called as *Uniform Connectedness (UC) regions* (closed regions of homogeneous properties-such as lightness, chromatic color, motion) by Palmer and Rock in[2]. This first stage construction process is generally called "low-lever vision". Although there are different viewpoints about how these intermediate elements are constructed [20] [21], it is well accepted that the first stage low-level processing depends on local visual properties [9]. It is affected only by computations on immediately adjacent areas instead of by computations on distant regions of a scene [22]. It is highly parallel over space. The retina and primary visual cortex seem wired up to perform these computations.

- **Middle-level Vision** According to Palmer and Rock's theory [2] as shown in Figure 2.1, once these intermediate elements (or UC regions) have been constructed, they are submitted to the so-called figure-ground process that is one of middle level vision processes. This figure-ground process that aims at foreground and background separation may be influenced by the feedbacks from a

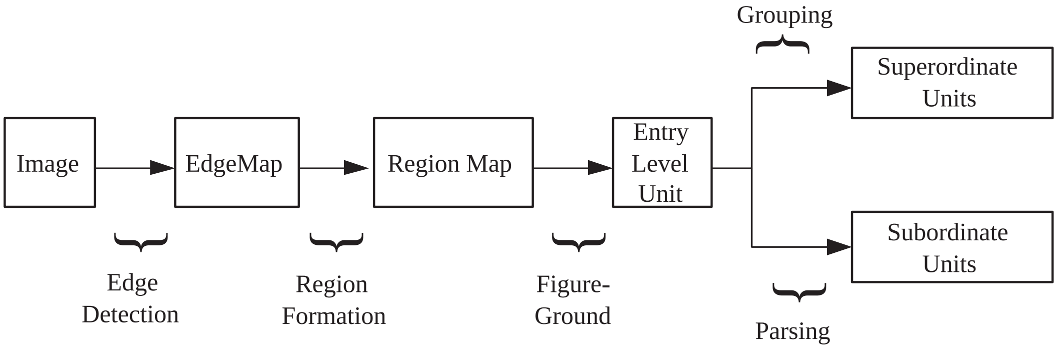


Figure 2.1: A flowchart representation of the relations among processes proposed to be involved in perception organization. (Reproduced from Palmer and Rock[2], Figure 13).

later process, especially by “common fate” grouping [23]. The elements identified as figures can be either parsed into subordinate units or further grouped into more complex visual entities, by the action of at least some of the classical Gestalt laws such as: common fate, proximity, closure, similarity. At the end, semantic “understanding” of a scene can be achieved by high-level vision [24] .

2.2.3 Motion Perception Theory

Substantial evidences in biological vision systems show that the presence of motion makes object detection, segmentation, and recognition easier, since motion cues can provide critical information for visual perception.

- **Short-range and Long-range Motion** How motion cue is used in perception is an interesting question. There are several different theories about motion perception mechanism. A dominant theoretical framework is *short-range* and *long-range* dual process theory, which was first proposed by Braddick [25]. It suggest that motion perception is mediated by *short-range* process and *long-range* process. The short-range process is a low-level process, occurring at an earlier stage in visual system, combining information over a relatively short spatial and temporal range. It occurs within brief temporal intervals and small spatial neighborhood[26]. The long-range process is a higher-level process, which operates over long distance and long durations. The outputs of short-range process can be serve as inputs to the long-range process. Cavanagh and Mather [27] classify motion stimuli into three categories: first order, second order, and third order. They think that three kind of different motion perception systems are needed to detect these three kinds of stimuli. Although there has been much debate concerning the motion perception mechanism, a common shared idea is that motion information is extracted and used in an hierarchy processing: from low-level to higher-level. The outputs of low-level processing can be taken as inputs by higher-level processing.

- **Common Fate Theory**

  Another important finds about motion perception is *common fate* theory: Elements that move together are grouped together. According to Gestalt psychology, common fate motion is a critical and robust source of information for dynamic object segmentation. Recent research suggests that common fate motion is one of the first object segmentation cues used by young infants. It can define objects in multiple object tracking [28]. In our research, we will explore an important long-range visual cue: trajectory, as a mid-level feature to attack non-convex classification problem in Chapter 3. Common fate theory is the theoretical foundation for us to define trajectory similarity, upon which object number will be estimated by trajectory merging process.

2.2.4 Bottom-up Processing in Visual Perception

Different aspects of the visual perception in HVS have been studied by different researcher with different concern. From the early work by Marr [19], Witkin and Tenenbaum [29], and Lowe [30], to more recently work by Sudeep Sarkar [7] [31] [32] [33], different perception principles are picked to instruct artificial vision research. We believe that video segmentation could benefit from a comprehensive understanding about visual perception. In order to have such a comprehensive knowledge, we will try to combine several visual related perception theories into a systematic computational framework, in which, different perception rules can efficiently work together, and benefit each other.

By comparing the conception of *matching units* in Kubovy's joint spatiotemporal grouping theory[10], with that of Kalmer's *Uniform Connectedness (UC) regions*, it is obvious that *Uniform Connectedness (UC) regions* is one kind of *matching units*, which should be the result of joint spatial and temporal grouping. Combing Braddick's motion cue study results [25] with Palmer's UC theory, It is reasonable to make a inference that short-range motion cue should be involved in the construction process of UC regions, and long-range motion cue may be involved in intermediate level grouping. Based on Palmer and Rock's flowchart representation of the relations among several different perception processes as shown in Figure 2.1, by incorporating with Braddick's motion cue study results [25] and Kubovy's joint spatiotemporal grouping theory [10], we develop a possible computational model for low-level and intermediate-level (or middle-level) vision perception processing in Fig 2.2.

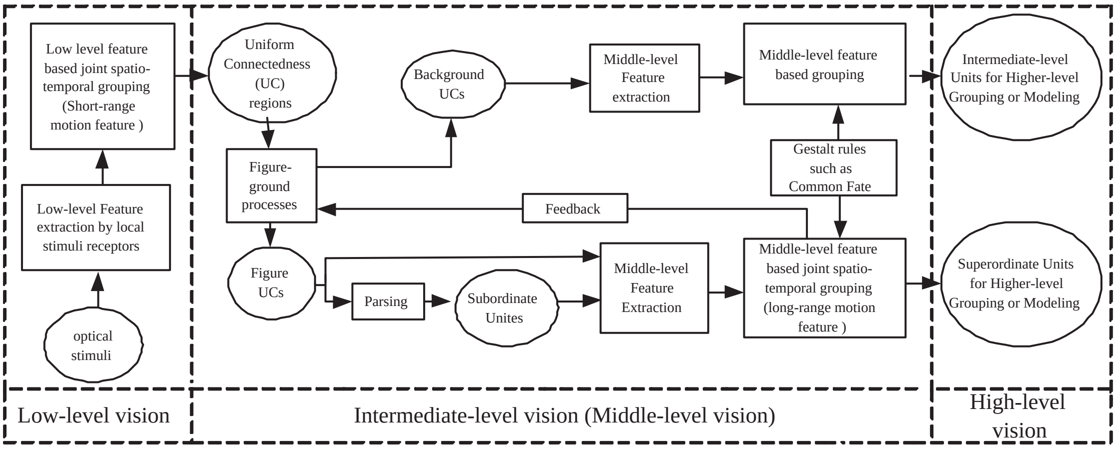

Low level feature based joint spatio-temporal grouping (Short-range motion feature)
Low-level Feature extraction by local stimuli receptors
optical stimuli
Uniform Connectedness (UC) regions
Background UCs
Figure-ground processes
Figure UCs
Parsing
Subordinate Unites
Feedback
Middle-level Feature extraction
Middle-level feature based grouping
Gestalt rules such as Common Fate
Middle-level Feature Extraction
Middle-level feature based joint spatio-temporal grouping (long-range motion feature)
Intermediate-level Units for Higher-level Grouping or Modeling
Superordinate Units for Higher-level Grouping or Modeling
Low-level vision
Intermediate-level vision (Middle-level vision)
High-level vision

Figure 2.2: Computational model of perception processing

In this computational model, optical stimuli are first sampled by small receptors:

photoreceptor neurons. Low-level visual cues (including short-range motion cues) are extracted by local stimulus receptors. Then, they are first grouped into small Uniform Connectedness (UC) regions by joint spatiotemporal grouping. Once UC regions have been established, they are submitted to figure-ground (fore-ground/background) process. This figure-ground process may be influenced by feedbacks from later higher-level grouping processes, especially by ”common fate” guided grouping [23]. These elements identified as figures can be either parsed into subordinate units or further grouped into more complex visual entities [2]. This grouping process may follow the rules of the classical Gestalt laws such as: common fate, proximity, closure [2]. Long-range motion cues should be involved in this intermediate level grouping.

The goals of low-level vision is to build UC regions based on both short-range motion cue and other physical properties of the visual environment. The main computational properties of the low-level vision is local, parallel, fast, robust to input noise, and be of bottom up [22][9].

The fundamental goals of bottom-up low-level and mid-level processes are to group entities together hierarchically into higher-level forms, upon which higher-level representation can be defined for higher-level vision processing. The goals of high-level vision is to complete the job of delivering a coherent interpretation of a scene.

2.2.5 Combined Bottom-up and Top-down Processing

The bottom-up sequential perception processing model has wide influence in the community of computer vision. But when this feed-forward model is used to process a noisy and cluttered scene, it usually fails to identify objects that can be easily recognized by human [34]. Bullier thought that the failure of bottom-up feed-forward model is due to the separation between the high-level prior information and the local bottom-up segmentation [35]. More and more evidence from neuroscience and psychology show that the top-down modulation is essential and indispensable in visual perception.

In contrast to low-level vision, which is concerned with feature extraction, the top-down modulation in high-level vision is primarily concerned with the interpretation and use of prior knowledge and information in a scene. High-level visual processes are performed on a selected portion of the image rather than uniformly across the entire scene, and they often depend upon the goal of the computation and prior knowledge related to specific objects [24]. Recently, psychologists modeled the top-down processing in high-level vision as statistical inference [36, 37, 38]. According to some recent neurophysiological evidences, and inspired by the most successful computer vision algorithms, Lee and Mumford suggested that the interactive bottom-up and top-down computations in visual neurons might be modelled by the mechanisms of particle filtering and Bayesian-belief propagation algorithms [39]. Combining the work of Lee and Mumford in [39] with these classic bottom-up sequential perception theories aforementioned, we present a computational model for visual perception in Fig. 2.3, where bottom-up and top-down processes are combined together.

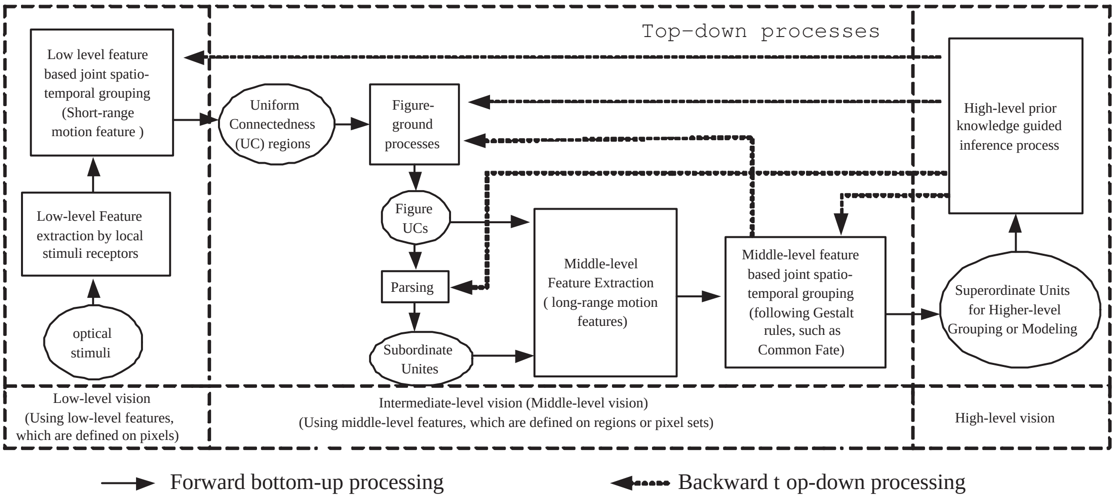

Top-down processes
Low level feature based joint spatio-temporal grouping (Short-range motion feature)
Low-level Feature extraction by local stimuli receptors
optical stimuli
Uniform Connectedness (UC) regions
Figure-ground processes
Figure UCs
Parsing
Subordinate Unites
Middle-level Feature Extraction (long-range motion features)
Middle-level feature based joint spatio-temporal grouping (following Gestalt rules, such as Common Fate)
High-level prior knowledge guided inference process
Superordinate Units for Higher-level Grouping or Modeling
Low-level vision (Using low-level features, which are defined on pixels)
Intermediate-level vision (Middle-level vision) (Using middle-level features, which are defined on regions or pixel sets)
High-level vision
Forward bottom-up processing
Backward t op-down processing

Figure 2.3: Computational model of combined bottom-up and top-down processing.

Within the framework of particle filtering and Bayesian-belief propagation algorithms, bottom-up processing generates data-driven hypotheses and top-down processing provide priors to reshape the probabilistic posterior distribution of various hypotheses. In this model, top-down processing could begin as early as the figure-ground process stage [40]. Specifically, for motion perception of human body, the prior knowledge and information about human body appearance pattern and motion pattern provide prediction and top-down prior, which play an important role in the interpretation and segmentation of human motion.

2.2.6 Application to Visual Segmentation

In the following, we will briefly review the recent research on visual segmentation, including image and video segmentation, which has been motivated and inspired by general perception principles.

- **Joint Spatiotemporal Approaches** Many video segmentation algorithms can be grouped into the catalog of joint spatiotemporal principle, such as the Normalize cuts graph partitioning method presented by Shi and Milk [41, 42], the mean shift method proposed by DeMenth [43], the Gaussian Mixture Model method proposed by Greenspan, et al [44]. A good survey of joint spatiotemporal grouping techniques for video segmentation can be found in [45]. According to the perception organization theory, low-level feature based segmentation is just at the beginning stage, i.e., “low-level vision”. Therefore, it is not a surprise that the results of single-stage spatiotemporal segmentation algorithms are still far away from “semantic video segmentation” in a general scene.

- **Multi-layer Bottom-up Approaches** The multi-layer framework is an effective approach for video segmentation. In [46, 47, 48, 49], pixels are first grouped into small homogeneous regions in each frame. Then, segmented 2D regions are merged (or tracked) into 3D volumes in space and time. Recently, some multi-layer algorithms construct 3D space-time regions at the first stage.

They match well with both the joint spatiotemporal principle and the hierarchical perception organization theory. In [50], a region growing method was proposed to construct the smallest homogeneous 3D blobs at the first stage, then, new features such as boundary, trajectory and motion of these blobs are extracted. Based on the extracted new features, these over-segmented small blobs can be further grouped into more advanced structures. The 3D watershed method was proposed in [51] to generate 3D blobs, and these blobs are then merged into more semantically meaningful objects. Multi-layer segmentation algorithms can use relevant visual features in different layers to achieve the final segmentation progressively. However, all multi-layer algorithms face some problems when there is a cluttered scene or objects have complicated visual properties and behaviors. Another challenge is how to decide when the cascaded merging processes should stop. Possible solutions to these problems may be provided by top-down processing.

- **Combined Bottom-up and Top-down Approaches** Although the segmentation scheme that combines both bottom-up and top-down processing is commonly advocated in the computer vision community [52], there has not yet been a widely accepted computational framework to achieve that goal. In [53], Borenstein et al. proposed an example of how to combine both bottom-up and top-down approaches into a single figure-ground image segmentation process. The unified segmentation, detection, and recognition framework proposed in [54] might be one of the most successful examples of applying this scheme in image segmentation. How to apply the similar idea to video segmentation remains to be a very challenging research topic.

2.3 Biological Movement Perception

Until now, little work has been done in combining bottom-up and top-down processes for semantic *video* segmentation, even though we know that it is one of the most promising directions. Since we are interested in human detection, tracking and segmentation, we will review several neurophysiological and physiological literatures on biological movement perception, particularly human motion perception. Through this study, we try to find some inspiring hints and general guidelines to develop a practically plausible computational model to guide our future research.

2.3.1 Neural Mechanisms of Biological Movement Perception

- **Two-pathways Vision Perception** Physiological studies found that there are two separate functional streams involved in vision perception: a ventral stream for the analysis of form (the "what" stream) and a dorsal stream for the analysis of position and motion (the "where" stream) [14, 15]. This discovery is called one of the major breakthroughs of the last few decades of vision research [55]. Inspired by this discovery, Giese and Poggio proposed a neural model for biological movement recognition [56]. This model has two separated parallel hierarchical pathways: *form pathway* and *motion pathway*, which are specialized for the analysis of form and motion information respectively. This model has a feed-forward architecture. The form and motion pathways consist of hierarchies of neural detectors that are connected unidirectionally in a bottom-up fashion.

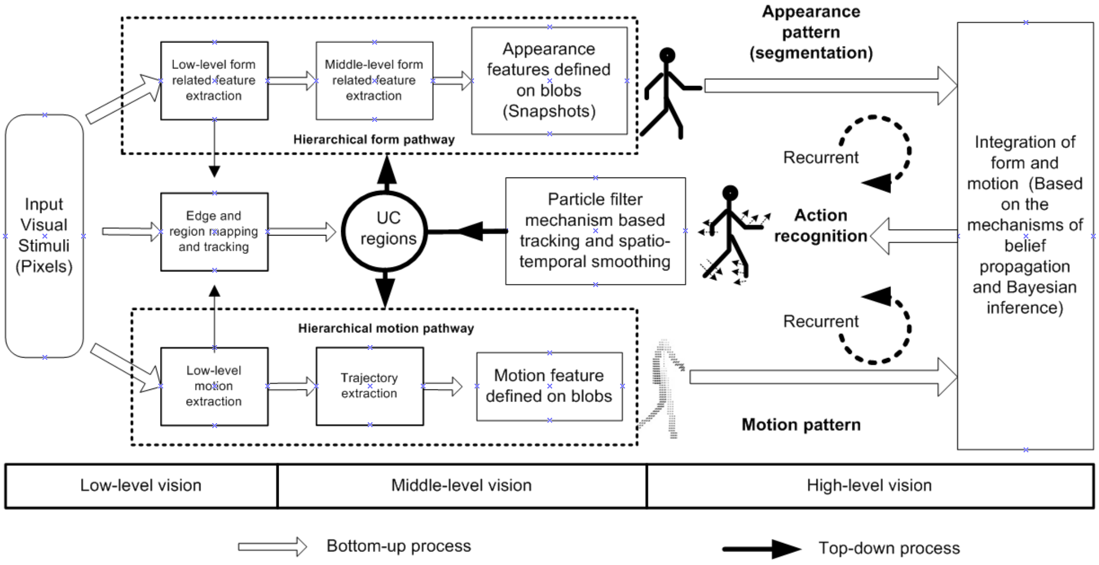
- **Interaction and Convergence of Two Pathways** Further physiological and neuropsychological studies show that the two processing streams interact at several levels. Oram and Perrett found evidence for the convergence of two pathways in the anterior part of the superior temporal polysensory area (STPa) of the macaque monkey [57]. The integration of two separate aspects of information about a single object has been referred to as the binding problem. In[58], Sajda and Baek describe a probabilistic approach for the binding problem, which uses a generative network model to integrate both form and motion cues using the mechanism of belief propagation and Bayesian inference. The discovery of the convergence of two pathways in higher-level vision neural region is very important because the interaction of the two pathways may be realized by a feedback from the convergence place. A similar mechanism is used to explain the exchange of information between two distant neuron regions, where direct information exchange is inefficient or difficult [59].

2.3.2 Visual Tracking Theory

In computer vision, visual tracking is a common method for human motion and pose estimation that can provide prior information for the next frame. This strategy is well supported by a psychological theory called "object specific preview benefits" (OSPB) [60], which states *that the detection of a dynamic object's features are speeded when an earlier preview of those features occurs on the same object, ...*. Although the nature of visual tracking has not been systematically studied, there are two main theories for Multiple Object Tracking (MOT): Pylyshyn's visual index theory [61] and Kahneman's "object file" theory [60]. The visual index theory matches the construction process of 3D UC blobs very well. It is a low-level automated vision process, and no attention is needed. In contrast with the visual index theory, the "object file" theory suggests that the effortful attention is needed for a successful tracking process. An object file is a middle-level visual representation stored in the shot-term working memory, which collects spatiotemporal properties of the tracked objects, and the content of object files will be updated when the sensory situation changes.

2.3.3 A Comprehensive Computational Model

Based on previous studies and discussions, we hereby propose a comprehensive computational model for human detection, tracking and segmentation, as shown in Figure 2.4. This computational model is deeply inspired by previously reviewed perception principles and biological movement perception. Related to the hierarchical neural model proposed in [56], the top flow is the bottom-up hierarchical form pathway; the bottom flow is the hierarchical motion pathway. This model is developed from the the general computational model as shown in Figure 2.3. Moreover, the bottom-up processes begin from the input visual stimuli (pixels). The output of form pathway is the appearance pattern (object segmentation), and that of motion pathway is the motion pattern (motion estimation). Action recognition is achieved by the integration of both pathway outputs.



Input Visual Stimuli (Pixels)
Low-level form related feature extraction
Middle-level form related feature extraction
Appearance features defined on blobs (Snapshots)
Hierarchical form pathway
Appearance pattern (segmentation)
Edge and region mapping and tracking
UC regions
Particle filter mechanism based tracking and spatio-temporal smoothing
Recurrent
Action recognition
Recurrent
Integration of form and motion (Based on the mechanisms of belief propagation and Bayesian inference)
Hierarchical motion pathway
Low-level motion extraction
Trajectory extraction
Motion feature defined on blobs
Motion pattern
Low-level vision
Middle-level vision
High-level vision
Bottom-up process
Top-down process

Figure 2.4: A comprehensive computational model of human motion analysis.

For human motion perception, top-down processing begins from stored (or learned) prior knowledge about the human appearance and motion patterns, and then combines with the outputs of bottom-up processing to recognize human action via inference [58]. The recognized human action (appearance+motion) is used as prior for tracking. UC region construction is the entry-level unit for the part-whole hierarchy, and we argue that the top-down inference should not go beyond the UC region. In other words, the UC region is the fundamental unit for top-down tracking. In the proposed model, Kahneman's "object file" theory in [60] can explain visual tracking after UC regions are generated. This is also supported by Kahneman's suggestion that "visual index" might be the initial phase of a sample object file. Therefore, in the proposed model, the two tracking theories can be integrated at different levels and in a serial flow.

2.4 Discussions and Conclusions

The model in Fig. 2.4 could help us understand some difficult problems in human detection, tracking and segmentation. For example, in a bottom-up framework, motion estimation and object segmentation are often considered as a chicken-egg problem. Here object segmentation, motion estimation, and action recognition are results of recurrent and interwound feedforward/feedback processes. Also, model order estimation in bottom-up processing can be better understood under this framework. According to [24]: *the top-down process tells the bottom-up process where to stop.* As we mentioned earlier, particle filtering algorithms can help physiologists better understand human perception. In turn, studies of biological vision may help us to improve our computer vision algorithms. Guided by the model as shown in Fig. 2.4, we also raise the following open questions for future research.

- **How to combine both appearance and motion prior information about a moving human into statistical inference?** Most current particle filter algorithms use only a dynamic appearance model, few of them consider motion features or patterns. It has been already proved that combining both motion and appearance information can achieve very promising results for human motion detection and recognition [62, 63].

- **How to use bottom-up results as the inference unit in the top-down process?** In other words, how can we build a data driven particle filtering structure to combine both bottom-up and top-down processing? In [64], Tu and Zhu proposed a data driven method for image segmentation, where the fundamental units for inference are pixels. In human detection, tracking and segmentation, the UC region should be the fundamental unit for object tracking processing.

- **May we use the idea of "object file" theory to attack the occlusion problem by registering tracked objects across frames?** It is well known that occlusion is a difficult problem for particle filtering-based human tracking. Decomposing a complex object into several independent moving parts, tracking them individually, and building an "object file" for each of them may be a good way to deal with the occlusion problem. A similar idea has been implemented in the tracking algorithm proposed in [65].

- **How to use the idea of matching between adjacent frames into the particle filter to enhance the tracking performance?** In the proposed framework, tracking acts as a bridge between objection segmentation, motion estimation, and action recognition. Tracking processing plays a key role to sustain and stimulate recurrent interactions among them. However, matching between adjacent frames, as a nature of tracking in the HVS, start to received attention in the community of particle filtering research. An appearance-adaptive method proposed in [66] is an exemplary effort in this direction.

- **How to represent and learn prior knowledge for statistical inference?** In general, we assume that prior knowledge is known by certain off-line learning algorithm. When the off-line learned prior information does not fit current observations, tracking performance could suffer. Therefore, good balance between online and off-line learning may greatly improve the robustness and effectiveness of tracking. In [63], Lim et. al. proposed an example of online manifold learning, which has archived promising results.

CHAPTER 3

Low-level Vision: Bottom-up Segmentation

3.1 Overview

As compared with human detection, which has made significant progress over the last few years, not much progress has been made in human segmentation, and the role of segmentation has largely been ignored. In many human object related studies, such as human pose estimation and human tracking, segmentation problem is often circumambulated by assuming its results are already available, or assuming segmentation results can be obtained by a background subtraction process. This kind of assumption ignored an important fact that segmentation can be an important tool to support human object analysis in many aspects from low-level feature extraction, to mid-level human part detection, and to high-level knowledge representation and inference.

For low-level and mid-level bottom-up segmentation processes, our purpose is to get compact image representation by grouping pixels into more meaningful segmentation results with less over-segmentation and under-segmentation. When objects can be depicted in term of semantical words, such as the crown of a tree, a car,or a tree's trunk, we call this object-based video segmentation is at semantical level or in the term of semantical-level video segmentation. Both new multimedia standards MPEG-4 [67] and MPEG-7 [68], which provide users with the flexibility of content-based video representation and description, can benefit from semantically meaningful object-based segmentation. Object-based segmentation has received intensity interest during the past decade, but it is still one of most challenging tasks in video processing due to the complexity of real-world video data. We know that human vision system (HVS) can easily partition video scenes into meaningful objects and recognize them effortlessly. Few person doubts that good knowledge about HVS can benefit research in video segmentation.

Cognitive science research shows that spatial grouping and temporal grouping are jointly involved in HVS [10][16] [11]. Supported by this perception principle, joint spatiotemporal video segmentation algorithms are particularly attractive. In [41, 42], Shi and Milk have presented normalize cuts graph partitioning method for joint spatiotemporal segmentation. But this classification is achieved at the cost of heavy computational burden. $Nystr\ddot{o}m$ method was presented by Fowlkes et al. [69] to alleviates computational load of normalize cuts approach. In general, the major limitations of normalize cuts approach are the heavy computational load, sensitivity to noise, and manual estimation of the cluster number.

As an appealing alternative to graph theory methods, statistical method can also be used for joint spatiotemporal segmentation. In general, statistical method classify a multidimensional feature vector space. In [43], DeMenthon proposed a nonparametric statistics paradigm, in which hierarchical mean shift method was used to cluster a seven-dimensional feature vector space. A parametric statistics paradigm was proposed by Greenspan, et al [44], which is based on Gaussian Mixture Model(GMM) learning for a six-dimensional feature space. The order of GMM is estimated by the Minimum Description Length (MDL) criterion. This statistical modelling method tend to be more computationally efficient and very robust to video noise. However, it assumes each cluster has a gaussian distribution. This assumption make it be not suit for the classification of feature space that has complicated manifold structure. In other words, it has difficult to deal with non-convex classification problem, which can make for a either over-segmentation or under-segmentation result. This is the main hindrance to achieve semantically meaningful segmentation results.

Up to now, compared with the requirement of automatic semantic-level video segmentation, all the single stage low-level feature based algorithms have achieved only a very limited success. Their results are still far away from semantically meaningful in more complex and realistic scenarios. According to Sarkar and Boyer's feature level based classificatory structure [7], one step low-level feature based joint space-time algorithms are just at the beginning stage: signal level processing.

The limitation of single layer classification is acknowledged by more and more people. For complex classification problem, multi-classifier strategy is advocated by many researchers, such as, Kittler[70], Fred and Jain [71]. Basically, there are two structures for combining multiple classifiers, i.e., parallel and cascaded. Supported by Marr's sequential perception processing theory[19], multi-layer cascaded classification approaches are very popular for video segmentation, in which, pixels are first grouped into small homogeneous regions based on features such as color, position; then these regions are further grouped according to some new features extracted from these regions. In [46][47][48][49], segmentation is operated frame by frame at the first stage. It's outputs are 2-D regions. Then, these 2-D regions are merged (or tracked) into 3-D volumes in the following stage operation. Very recently, some multi-stage algorithms construct 3-D space-time regions at the first stage processing. In [50], Porikli etc. use region growing method to construct smallest homogeneous 3-D blobs at the first stage, then, new features such as boundary, trajectory and motion of these blobs are extracted by so called, self descriptors. Based on extracted new features, these over-segmented small regions can be further grouped into more advance structures by hierarchical clustering method. Instead of region growing method, 3-D watershed method is used by Tsai et cl. [51]to generate 3-D blobs at the first stage, then blobs are merged by a Bayesian framework based on new features extracted from these blobs. Unlike the single stage algorithm, multi-stage algorithms can use different level features into classification. However, many multi-stage algorithms discard the conception of joint spatio-temproal grouping. Moreover, motion information, as the important visional cue, has not been enough explored in many multi-stage algorithms.

In this chapter, we attempt to build an effective computational framework, by which the limitations of different classifiers can be evaded by the cooperation of each other, different level features or representation can be effectively extracted and involved in segmentation. We extended Greenspan's statistical modelling method [44] into a perception principle guided multi-layer framework, which combines both the merits of multidimensional approach for joint spatiotemporal grouping, which was believed to be the most promising direction recently [32], and the merits of a multi-stage process. In our algorithm, feature selection and classifier design, the two key issues of pattern recognition, are inspired by cognitive science studies: what and how visual cues are extracted and used in HVS and how visual information is transformed in HVS. Guided by a possible perception computational model, no-convex classification problem can be attacked in a cascade multi-layer classification framework. Object number estimation is carried out at the last layer classification, when higher-level feature: "trajectory" , which capsulate more information about object number, is available. The most time consume operation: MDL criterion-based model order estimation is discarded. Our simulation results verify that our new framework can achieve a more meaningful segmentation result in some complex and realistic scenarios. It is also computationally effective.

In Section 3.2, we will first review the Gaussian Mixture Model(GMM)-based statistical modelling approach for video segmentation proposed by Greenspan et. al' in [44]. In order to get some hints from HVS to attack limitations of single-layer statistical modelling approach, we will introduce some related perception principles in Section 3.3 and deduce some guidelines and perception computational model to instruct our algorithm designing. The detail implementation of our algorithm is given in Section 3.4. Experiment results are presented in Section 3.9 to validate our algorithm. At the end, conclusions and future work discussion will be given in Section 3.10.

3.2 Statistical Video Modelling

3.2.1 GMM-based Video Modelling

By assuming that all pixel-wise feature vectors are generated from a multivariate Gaussian mixture model (GMM), each homogeneous region can be represented by a multi-dimension Gaussian distribution. GMM parameters can be estimated by the Expectation Maximization(EM) algorithm. We now give a brief review of the statistical model-based video segmentation technique proposed by Greenspan, et al in [44]. It has mainly three steps. First, given a video sequence, a six-dimensional (6-D) feature vector, i.e., 3-D (L, a, b) color descriptor, 2-D position $(x,\ y)$, and time or frame index (t), is extracted for each pixel. The second step is EM-based GMM model learning. Let o be the feature vector in R^d, the mixture density is defined as: $f(o|\theta) = \sum_{j=1}^{K} p_j \varphi_j(o|p_j, \mu_j, \Sigma_j)$ Given a set of feature vectors $o = \{o_i; 1 \leq i \leq N\}$, the maximum likelihood estimation of θ is: $\theta_{ML} = \arg\max_\theta f(o_1, ..., o_N|\theta)$. The number of model components K value is decided according to the Minimum Description Length (MDL) criterion [72]. Specifically, a whole set of candidate GMMs with different component numbers ranging from k_{max} to k_{min} has been obtained by using the EM algorithm $(k_{max} - k_{min} + 1)$ times.

$$\log p_y(y|K, \theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} p_j \varphi_j(o|p_j, \mu_j, \Sigma_j), \tag{3.1}$$

$$MDL(K, \theta) = -\log p_y(y|K, \theta) + \frac{1}{2} L \log NM, \tag{3.2}$$

where the parameters $\theta = \{p_j, \mu_j, \Sigma_j\}_{j=1}^{K}$ are to be estimated and $0 < p_j < 1$, $\sum_{j=1}^{K} p_j = 1$. The model of order k_{opt} is the one that can minimize the MDL criterion.

In other worlds, the optimal K_{opt} is found by searching a set of candidate models with different orders K ranging from K_{max} to K_{min}. After GMM model training, the last step is that we segment the video by assigning each pixel to the most probable Gaussian cluster, which maximizes the *a-posteriori* probability(MAP). Model-based object segmentation actually implements the MAP classification of all pixel-wise feature samples derived from a video shot. Each Gaussian component in the feature space corresponds to a video region, whose certain properties, such as position, color, or velocity, can be calculated by associated Gaussian model parameters.

3.2.2 Further Developments

Since the covariance matrix coefficients of GMM model can only describe the mean direction and velocity of complex motion patterns, in order to get precise description of nonlinear motion patterns, Greenspan [73] proposed an extended scheme, termed *Piecewise* GMM framework. The computational load of these object-based segmentation methods is normally very high. Recently, improved methods were proposed by our research group [74][75][76]. We assume that there is a trade-off between efficiency and robustness of the EM training. Particularly, we have introduced key-frame extraction prior to model training, and GMM estimation is only based on a set of pre-selected key-frames whose optimal number depends on the complexity video data. Then trained GMM is applied to whole data set for video segmentation. It was found that extracted key-frames can contain sufficient training samples with much reduced redundancy and outliers, leading to robust and efficient model training.

3.2.3 More Discussions

Although these works just mentioned in last subsection can improve the video segmentation performance dramatically, the statistical modelling scheme described thus far still has several limitations which we would like to address.

The first limitation is that the MDL-based GMMs model order estimation is not only very time consuming, but also not suit for estimation of the number of semantic objects. The optimal Gaussian order K_{opt} is found by searching a set of candidate models, which need to run EM algorithm to learn GMMs model parameter several time. Moreover, Low-level pixel-wise feature based MDL method can only give a estimation of how many homogeneous regions existing in a video sequence. In general, a sematic object may content several homogeneous regions. So, the estimation of the number of sematic objects is not a low-level grouping issue, but need higher level features.

The second mainly limitation is non-convex classification problem, which is the main hindrance for achieving semantically meaningful object segmentation. A non-convex video segmentation example is shown in Fig. 3.1. The first row is the original video sequence. Object segmentation results are shown in the following rows. Every row is associated with one segmented component. It is obvious that the second row are over-segmented results of moving objects. The third and fourth row are under-segmentation results. Non-convex classification problem always causes a either over-segmentation or under-segmentation result. Neither over-segmentation nor under-segmentation is our desired result. In order to achieve more meaningful segmentation result, non-convex classification problem deserves further investigation.

In this chapter, we classify non-convex classification problem of GMM-based clustering into two categories. The first kind of non-convex classification problem is shown in Fig. 3.2(a). The learned GMMs with $K = 2$ and $K = 3$ are shown in Fig. 3.2(b) and (c) respectively. Since one object concave into another object in feature space, it is impossible to find two Gaussian models, by which the two objects can be correctly classified. In video segmentation, this kind of non-convex classification problem mainly affects accuracy of object boundary segmentation, especially for background objects, which have long and narrow shapes and concave into other objects.

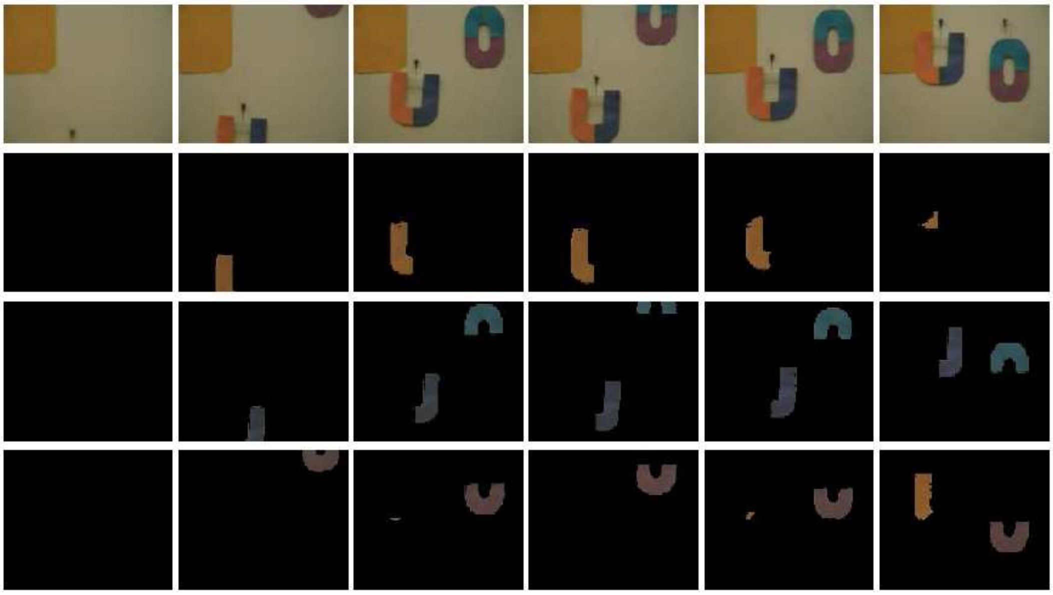
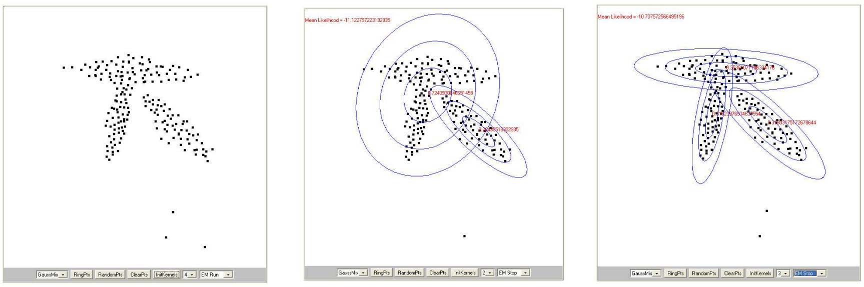


Figure 3.1: An example of non-convex video segmentation. The first row images are several frames of the input video. The second, the third, and the fourth rows are these detected and segmented moving objects. The third row and the fourth row are under-segmented results.

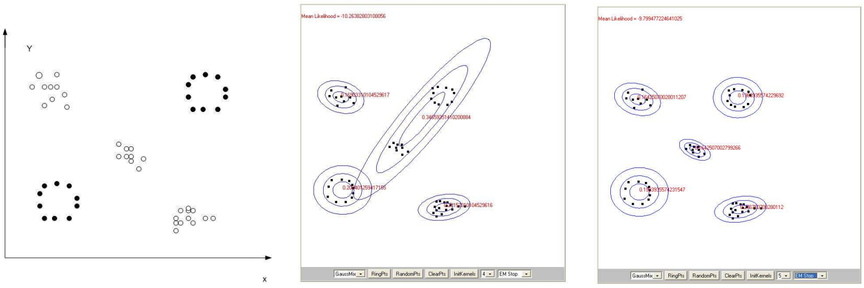
To attack the first kinds of non-convex classification problem, we can over segment the objects first, as showing in Fig. 3.2(c), then, according to the similarity of Gaussian models, merge the two components, which have the shortest distance. We will get the correct classification. Above motioned method is generally used for dealing with non-convex classification problem. It can be found in many literatures, such as in [77] and [78]. However, it has difficult to deal with more general case non-convex classification problem as shown in Fig. 3.3(a).

Object I are represented by two black dot sets, which have circle shape. Other objects are represented by light dots. Assuming the only features that we know are position (x, y). Obviously, one step statistical modelling classification will always give us a either over-segmentation result, like Fig. 3.3(c) or a under-segmentation result, like Fig. 3.3(b). Since the model distance between the two circles is not the shortest one, GMMs model distance-based merging can not correctly merge the two circles.



(a) (b) (c)

Figure 3.2: (a) An example of the first kind of non-convex classification problem. (b) GMM learning result for $K = 2$. (c) GMM learning result for $K = 3$. (Sofeware provided by [3])



(a) (b) (c)

Figure 3.3: (a) An example of the second kind of non-convex classification problem. (b) GMM learning result for $K = 4$. (c) GMM learning result for $K = 5$. (Sofeware provided by [3])

Fig. 3.3(a) illustrates the second kind of non-convex classification problem.

Before the attempt of designing a algorithm to attack the motioned limitations of single layer statistical modelling scheme, especially the non-convex classification problem, let us have look at how HVS do the complex classification work in the next Section.

3.3 Hints from Perception Principles for Video segmentation

A real-world video is usually orderly and rule-governed instead of visually chaotic. Since the early 20^{th} century, psychologists have found a set of rules that are followed by HVS. In general, there are at least two kinds of perception principles, which can bring us inspirations for video segmentation. The first kind is about what visual cues are used in HVS; Another is about how visual information is transformed in HVS. These two kinds of knowledge associate with the two key issues in pattern recognition: feature selection and classifier design. In this section, Based on the developed computational model in Chapter 2 as shown in Fig 2.2, we deduce some guidelines for feature selection and classifier design. Then, a possible solution for non-convex video segmentation is given.

In a view of computational study of vision, Hildreth and Ullman [79] state that we can take the vision process as the construction of a series of representations of visual information with explicit computation that transforms one representation into next. The earliest representations are first extracted simply and directly from the initial image. Subsequent representation capture the characters of visible surfaces. Since cognitive science conception "representation" associates with the term of "feature or feature vectors" in computer vision, in this chapter, features are classified into the following three categories: low-level feature, which is defined on pixels (such as color,position)and can be extracted simply and directly from the initial video, mid-level feature, which is defined on visible surfaces or homogenous small regions(such as color, position, edges, contours), and high-level feature, which is used for more intelligent processing. A similar feature classification was proposed by Sarkar and Boyer's in [7]. Another important point that we can learn from HVS is that motion features should be extracted in the form of short-range and long-range; further more, short-range features should be involved in the construction of UC regions. It is obvious that low-level features alone are often insufficient for semantical video segmentation. From the analysis in Chapter 2, we know HVS does not achieve a semantical object segmentation of a scene in a single step. The single step low-level feature based joint space-time algorithms are just at the beginning stage of a cascade part-whole hierarchy visual information transform processing. These perception studies support a multi-layer cascade algorithm structure for semantic object-based segmentation, where both low-level features and higher level features can be extracted and effectively involved into classification.

From the studies of the HVS, we know that correct machine recognition depends on sufficient feature representing. For the example shown in Fig. 3.1, depending only on pixel-wise low level feature, we always get either under-segmented or over-segmented results in one step classification. Therefore, we will explore an important long-range visual cue: trajectory, as a mid-level feature to attack this non-convex classification problem. According to *common fate* rule of Gestalt psychology, elements that move together should be grouped together, elements belong to same object should move in the same direction at every frame. So, based on the similarity of motion trajectories, the elements that belong to same object could be merged together. In general, a motion trajectory is extracted as a coordinate sequence, which records the region center in every frame. So, the preliminary condition for a good trajectory feature extraction is an over-segmented "blob". For example, we cannot extract trajectory feature directly from an under-segmented blob as shown in the third row of Fig. 3.1. From the computational model of perception process as shown in Fig. 2.2, we know that there is parsing process before middle-level feature extraction and grouping. After a connectivity based parsing processing, under-segmentation result as shown in third row of Fig. 3.1 should be separated into over-segmentation blobs, upon which trajectory can be extracted correctly. After common fate-based trajectory grouping, moving object number could be also estimated, since common fate motion can define objects in multiple object tracking [28].

3.4 Multi-layer Framework Video Segmentation Algorithm

Encouraged by the case studies in the previous sections, a new perception principle guided video segmentation framework will be presented in this section. This framework is derived from the computational model shown in Fig. 2.2. The flowchart of our designed three-layer classification framework is shown in Fig. 3.4.

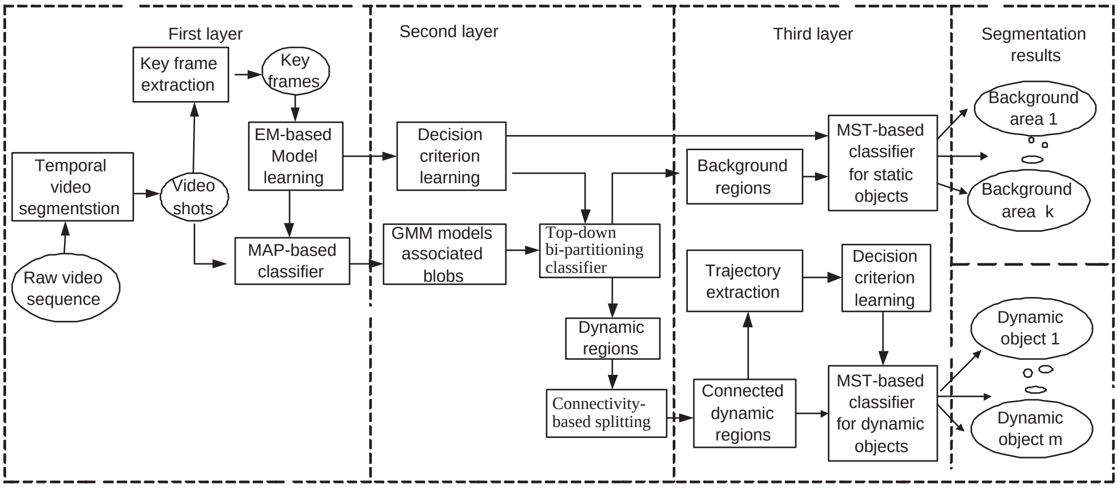


Figure 3.4: The proposed multi-layer and cascade video segmentation framework.

The first layer classification is built on recent research results of single-layer GMM-based segmentation framework described by Greenspan [44], and our previous work in [74]. We use key-frames-based EM algorithm to obtain GMMs model parameters. Here, the purpose of GMM Modelling-based clustering is to segment image sequence into small space-time Uniform Connectedness (UC) volumes or called blobs, which are entries of the following stage processing. In the second layer classification, the output blobs are classified into two groups: static (background) groups and dynamic groups based on corresponded Gaussian model parameters. This top-down motion detection process guarantees the blob merging in the last stage would not happen between static blobs and dynamic blobs. By an another top-down process: parsing , dynamic blobs will be split into connected regions. This splitting process can facilitate trajectory extraction. In the third layer, still and moving regions are merged separately by using the MST method based on different similarity measures.

3.5 First Layer: GMM Modelling-based Segmentation

Bayesian approaches have enjoyed a great deal of recent success in their application to problems in computer vision. Some psychologists, such as: Kersten, Knill and Rao [36] [37] [38], also prefer a unified Bayesian framework to characterize human perceptual organization. GMM and EM-based model training are effective and robust in dealing with noisy multi-dimension feature space. From Section3.3, we know that the computation within the low-level vision is local, parallel, fast, robust to input noise, and be of bottom up. Bayesian-based Gaussian mixture model (GMM) method has similar computational properties as low-level vision. This fact makes them an excellent choice for the first stage classification.

The first layer classification of our framework is similar to proposed Grennspan's method in [44]. But there are two significant differences. The first difference is that we use different features. Besides three-dimensional YUV color feature, and two dimension position feature (x, y), the motion feature, i.e., *the intensity change over the time dY* is used in our algorithm, instead of the time feature: frame index t, which is used by Grennspan in [44]. The feature dy is extracted from the pixel-wise luminance difference of two consecutive frames in a video shot. It includes short-range motion information. The reason for us to make this feature changing is that feature t is a uniformly distributed. According to Law et al. [80], this kind of features make it difficult for Gaussian mixture learning algorithm to recover the underlying clusters. Our experiments also verified Law's statement. The involvement of feature "t" make segmentation results without time coherence. Without feature "t", our segmentation results are more time coherent, therefore, we can extract a stable mid-level feature "trajectory" for further grouping to achieve more meaningful segmentation results. After this feature changing, our algorithm can detect one kind of second-order motion as shown in Fig. 3.5, the detected moving object is shown in Fig. 3.5 (d),(e) and (f), which is very difficult to be detected by Greenspan's algorithm and all the algorithms belonging to the category of *segmentation with spatial priority* in the survey of Megret etc.[45].The second difference is the our GMMs learning is based on extracted Key frames. Key-frames contain the salient and important video content structures, therefore, not only key-frame based model training has a faster speed, but also the learned GMMs models can better characterize salient video content and has a better segmentation performance. For the video "UO", as well as the segmentation results of the first layer statistical modelling-based classification are shown in Fig. 3.1

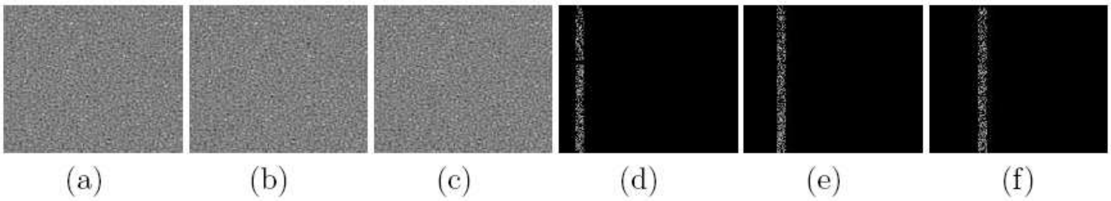


Figure 3.5: (a),(b),(c), are the first, third and 5th frames of a second-order motion video sequence. (d),(e) and (f) are detected motion from the frame (a),(b) and (c) respectively (Original video data comes from: http://www.psych.ndsu.nodak.edu/mccourt/Psy460/)

3.6 Second Layer: Bi-partitioning and Spatial Connectivity-based Splitting

3.6.1 Region Bi-partitioning

A bi-partitioning method is used in the second layer. Static GMM blobs and dynamic blobs are separated by threshold of their motion magnitudes in GMMs models, which are extracted from the first layer. Given a Gaussian component l and its motion vector μ_{dY}, the bi-partition is implemented as follows:

$$\begin{cases} \text{If } \mu_{dY} > \lambda, & \text{it is a moving blob;} \\ \text{Otherwise} & \text{it is a static blob,} \end{cases} \tag{3.3}$$

where λ is a threshold for moving object detection, and μ_{dY} is the mean value of motion vector dY in the GMM. This step is a top-down classification process, which makes it possible for the following layer to merge small static regions and dynamic regions into background and moving objects respectively.

3.6.2 Spatial Connectivity-based splitting

As mentioned before, the preliminary condition for a good trajectory feature extraction is an over-segmented blob. In other words, every space-time blob contents only one object or several parts of a single object. Because under-segmented space-time blobs content several objects, which may have different dynamic patterns as shown in the fourth row of Fig. 3.1, thus the extracted trajectory do not characterize any useful information. In order to achieve a better over-segmentation result, we need to increase GMMs model order K. However, the computational load of EM learning is in direct proportion to K^2. Therefore, it is not good idea to use very large value k to achieve a over-segmentation result. In the proposed framework, We apply a 4-neighbor connected component labelling process to divide undersegmentation into spatially connected components. This operation associates with the parsing processing of HVS. Since artifacts often take the form of small disconnected groups, we use size filter to eliminate noisy areas in binary map, ie. any blob smaller than a threshold size is removed. Using this parsing operation, the under-segmentation example as shown in the third row of Fig. 3.1 can be separated into two over-segmentation blobs as shown in Fig. 3.6.

3.7 Third Layer: MST-based Merging

3.7.1 Graph-based Approach for Classification

Graph theories have long been an important tool in computer vision, especially because of their representational power and flexibility. At the later stage of classification,

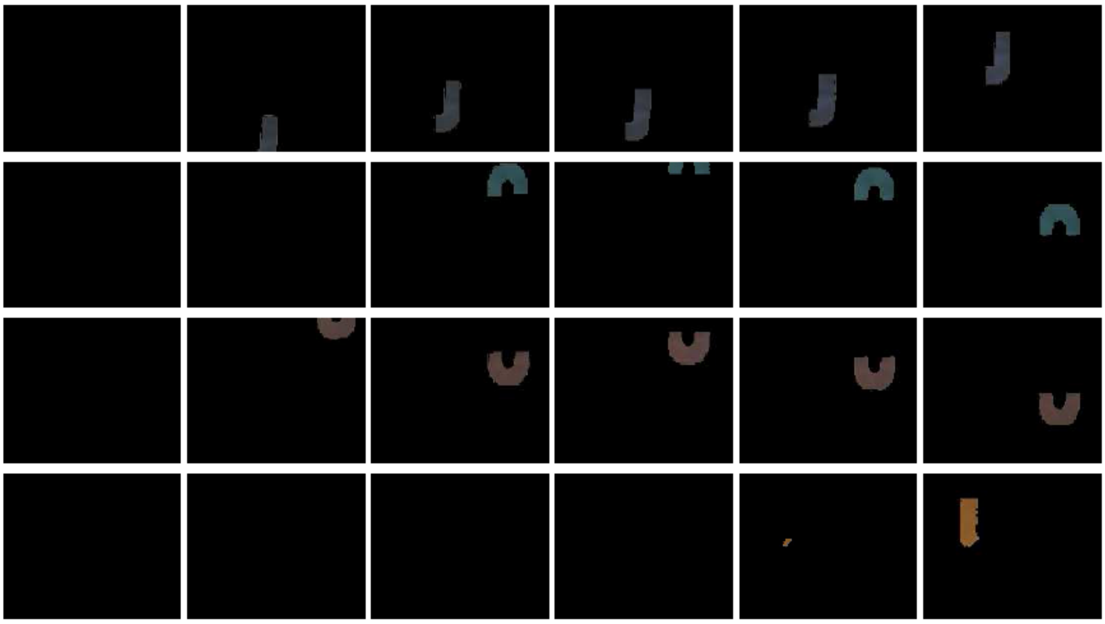


Figure 3.6: Connectivity-based splitting results of the second layer classification. The first and second row are connectivity-based splitting results of the blob shown in the third row of Fig. 3.1. The third row and fourth row are the connectivity-based splitting results of the blob shown in the fourth row of Fig. 3.1.

the number of input entries is too small to use statistical grouping method to group them. For computational convenience, we take a graph-based approach for the third layer classification. In a typical graph-based approach such as [81, 41, 42], pixels to be clustered are represented by an undirected adjacency graph $G = (V, E)$; every pixel is a vertices $v_j \in V$; edges $(v_i, v_j) \in E$ representing the link between neighboring vertices. Each edge $(v_i, v_j) \in E$ is associated with a non-negative measure of dissimilarity, called weight $w((v_i, v_j))$, which reflects the similarity between the linked vertices. A segmentation is achieved by remove edges $(v_i, v_j) \in E$ to form mutually exclusive subgraphs. There are different methods can be used to generate these subgraphs, such as graph cut approach, minimum spanning tree (MST) approach.

In the third layer of our algorithm, Kruskals algorithm-based MST approach is used to get subgraphs. Based on different similarity measures, static regions and dynamic regions are merged separately This MST-based merging method not only can handle some non-convex segmentation problems, but also dramatically speed up the segmentation process. Here the final segmentation component number is obtained by MST-based merging that depends on the thresholds of vertex similarity. The proposed MST-based merging method bypasses the MDL-based GMM model retraining process, so it can save the computational load drastically.

3.7.2 MST-based Merging for Static Regions

After the secondary layer, all the static region are separated from moving regions. Every static region corresponds to one Gaussian model in the GMMs. We consider every Gaussian model as a vertex of a graph, $G = (V, E)$, where G, V, and E denote a graph, a set of vertexes, and a set of edges respectively. Every edge has a weight, which is a similarity measurement between two Gaussian models. The detailed MST edge weight definition for static regions is showed in Equation 3.4. Kruskals algorithm is used to develop the MST tree. It is a typical bottom-up approach. In this layer, MST-based clustering is performed on finite Gaussian components of the trained GMM rather than on individual pixels. The edge weight function $D_s(\cdot)$ between two static regions characterized by Gaussian models l and m is defined to be an upper bound on the change in the MDL criterion due to the merging of two Gaussians. This upper bound is defined as a distance between two models in [82]:

$$D_s(l,m) = \frac{N\pi_l}{2} \log \frac{|\Sigma_{(l,m)}|}{|\Sigma_l|} + \frac{N\pi_m}{2} \log \frac{|\Sigma_{(l,m)}|}{|\Sigma_m|}, \tag{3.4}$$

where Σ_l and Σ_l are the covariance matrices of Gaussians l and m, respectively; $\Sigma_{l,m}$ is the covariance matrix of a new Gaussian obtained by merging Gaussians l and m; π_l and π_m are two prior probabilities, and N is the number of feature samples. The merging operation will stop if $D_s(l,m)$ is larger than a given threshold. That means only the merging that leads to insignificant MDL decrease will be accepted.

3.7.3 MST-based Merging for Dynamic Regions

After the connectivity-based splitting, trajectory is extracted as a set of tuples (x^n, y^n); where (x^k, y^k) is the center of blob in the kth frame. Dynamic blobs merging depends on trajectory similarity, but it's definition is very difficult, Although there are many literatures about how to measure the trajectory similarity, (Good review can be found in [83]), most of them deal with one dimensional and same length time-series. In video segmentation, extracted trajectories are two dimensional time-series and in general have different length. In our algorithm, since the final moving objects number is decided by trajectory similarity-based merging instead of by MDL criterion, we need a trajectory similarity definition, which has clear perception physical meaning.

Combining Gestalt rule: "common fate" and a probabilistic perceptual principle i.e.,the Helmholtz principle, we design a simple trajectory similarity measurement based on motion direction. "Common fate" rule state that elements that move together are grouped together. The Helmholtz principle is a general perception law. it was recently applied to image feature detection by Desolneux et al. in [84]. The Helmholtz principle states that an event is perceptible, that is to say significant, if its occurrence being a random situation is very small.

Let A and B be two moving regions co-exist in N continuous frames, we can compute the motion trajectory of the common part of them: $\mathbf{t}_A = \{(x_A^h, y_A^h)\}$ and $\mathbf{t}_B = \{(x_B^h, y_B^h)\}$, $h = 1, ..., N$, which are used to calculate the trajectory similarity between A and B. Three steps are involved as follows. First, motion sequences $\mathbf{m}_A$ and $\mathbf{m}_B$ of two regions are computed based on their trajectories. For example, $\mathbf{m}_A = \{u_A^h, v_A^h | h = 2, ..., N\}$ where $u_A^h = x_A^h - x_A^{h-1}$ and $v_A^h = y_A^h - y_A^{h-1}$. Then direction sequences $\mathbf{\Psi}_A$ and $\mathbf{\Psi}_B$ are estimated for two trajectories. For example, $\mathbf{\Psi}_A = \{\psi_A^h | h = 2, ..., N\}$ where $\psi_A^h = \arctan(u_A^h / v_A^h)$ and $\psi_A^h \in [180°, -180°)$. Thirdly, a sequence of direction matching between $\mathbf{t}_A$ and $\mathbf{t}_B$, i.e., $\mathbf{\Phi}_{A,B} = \{\phi_{A,B}^h | h = 2, ..., N\}$, is obtained as follows.

$$\phi^h_{l,m} = \begin{cases} 1 & |\psi^h_l - \psi^h_m| \leq \alpha \ \text{ or } \ |\psi^h_l - \psi^h_m| \geq 360^\circ - \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

where α is a threshold for direction matching, e.g., $\alpha = 20^\circ$.

Let j to be the frame number of two blobs having the same moving direction. It can be calculated by $j = \sum_{h=0}^{N} \phi^h_{l,m}$. Assuming the object moving directions at each frame are random, the probability of this event happening can be compute by:$C^j_N p^j (1-p)^{N-j}$, where, $p = \frac{\alpha}{360}$ is the probability of two blobs having the same moving direction randomly at each frame. According to Helmholtz principle, smaller $C^j_N p^j (1-p)^{N-j}$ value means more significant event happened. Here, that is to say, more likely the two trajectories are extracted from the deferent parts of a same object. So this probability can be a distance between two trajectories. We call it Helmholtz distance.

$$d_{helm}(l,m) = C^j_N p^j (1-p)^{N-j} \tag{3.6}$$

When N and p are fixed, $C^j_N p^j (1-p)^{N-j}$ is only function of j. In order to simplify calculation, we can use $\frac{1}{j}$ to measure the trajectory distance. Considering that different trajectory pair may have different N, we use a normalized definition $\frac{N}{j}$ as trajectory distance, which ranges from 1to ∞. When two trajectories are extracted from different parts of a same objects, the value trajectory distance $\frac{N}{j}$ is only affected by trajectory extraction noise. When there is no outlier, $\frac{N}{j} = 1$. Therefore, the distance between blobs l and m in terms of motion trajectory can be defined as in $d(l,m) = \frac{N}{\sum_{h=0}^{n} \phi^h_{l,m}}$.

Usually, it is unlikely for two moving regions of distinct motion trajectories to be one object. Similarly, it is also less likely for moving regions disconnected in space or time to be one object. Hence we define the MST edge weight function $D_d(A,B)$

between two moving regions A and B as following:

$$D_d(A,B) = \frac{N-1}{(\epsilon + 0.001)\sum_{h=2}^{N} \phi_{A,B}^{h}}, \tag{3.7}$$

when A and B are connected, $\epsilon = 1$, otherwise $\epsilon = 0$. This connectivity information can be obtained in the second layer.

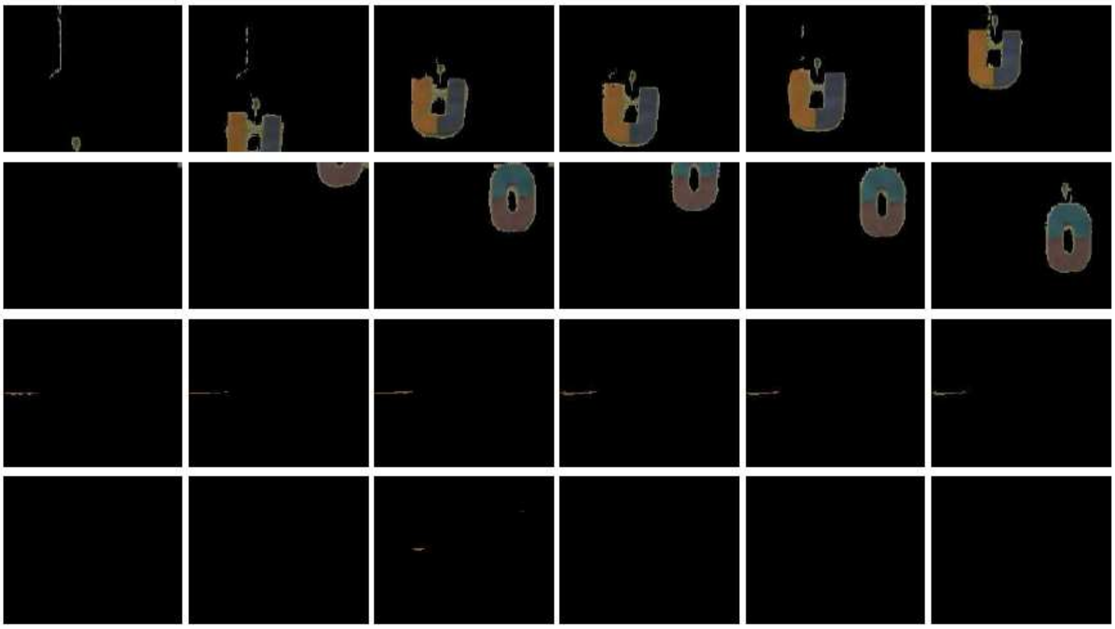


Figure 3.7: The trajectory-based merging results of the third layer classification. The first and second row are moving objects. The third row are parts of background, which are miss classified into dynamic blobs at the first layer. The fourth row are the noise of dynamic blobs.

For the non-convex classification example shown in Fig. 3.1. The trajectory-based merging results are shown in Fig. 3.7.The first and second row are moving objects. The third row are parts of background, which are miss classified into dynamic blobs at the first layer. The fourth row are the noise of dynamic blobs. From the results we can see the two non-convex moving objects are correct segmented. In the third layer classification, before merging process, we also check whether or not this blobs is a static blobs according to the position variance of trajectory. all miss classified static blobs will be merged together as a special background blob,as shown in the last row. So part of motion detection error happened in the first layer can be corrected at the third layer. In the first row of Fig. 3.7, some segmentation noise still can be seen. Most of them can be erased by sample morphological operation such as opening or closing. The simulation results of proposed algorithm is very robust to GMMs model order. In this example, we set GMMs model order K as 7, when K is changed from 7 to 18, we still can get the two correctly segmented moving objects. This multi-stage classification algorithm is very computationally efficient, it take only 60 seconds to process this 76 frames video sequence.

3.8 Further Discussion

The proposed framework does not purport to model human vision. Under the guideline of perception process computational model, this framework has a flexible structure. New developed technologies for feature extraction or classifier designing can be adopted into the framework. For example, in the first layer classification, the background registration technique proposed by Chien et al. in [85] may be adopted to extract short-range motion feature process to improve aperture problem. According to the flowchart in Fig. 2.1, UC regions are generated by both edge detection and region formation process. The idea in [86] may be useful for combining region and edge information into the generation process of UC regions. There are still many open spaces left for further research, such as how to extract more and reliable middle-level feature to achieve more robust and meaningful segmentation. Only the first kind of non-convex classification is discussed for background segmentation. How to extract more middle level features and design higher-level grouping process for background is beyond the scope of this chapter.

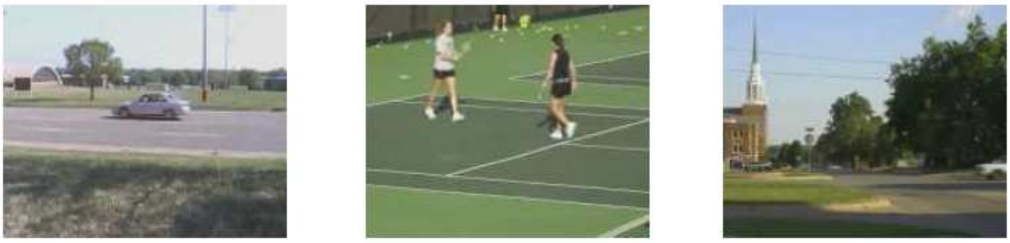
3.9 Experimental Result

In order to validate the effectiveness of the proposed video segmentation algorithm, a set of experiments is presented in this section. An efficient C-code implementation is developed for the proposed framework. We evaluate the proposed multi-layer segmentation framework by comparing it with the single-layer method proposed in [74]. Both of them was tested on a set of color video sequences. The frame size of each sequence is 176 × 144. The test platform is a PC computer (Intel Pentium 3.0GHz CPU and 1GB RAM). The experimental purpose is to test what kind of merits this perception principle guided framework can bring to us, and how well the proposed algorithm can mitigate limitations of single-layer statistical modeling algorithm as mentioned in Section 3.2.3. All the segmentation results can be clearly observed only in color image.

3.9.1 Computational Efficiency and Moving Object Number Estimation

As mentioned earlier, the MDL-based GMMs model order estimation is not only very time consuming, but also not suit for estimation of the number of sematic objects. After discarding MDL-based GMMs model order estimation, the proposed algorithm is expected to be more computationally effective. At the same time, we expect the common fate-based trajectory merging method could give us a robust estimation of moving object number. We start our experiment on several video sequences, i.e., Car, Tennis, and Church as shown in Fig. 3.8. Performance comparisons between the single-layer method and the multi-layer method are listed in Table 3.1. From Table 3.1, we can see the proposed multi-layer framework is much faster than the single layer method. This is due to the factor that we discard MDL-based model order estimation method. We run EM algorithm only one time.

In the segmentation example of the video as shown in Fig. 3.1, when preset initial GMMs model order K is changed from 7 to 18, the proposed multi-layer algorithm



(a) "Car" (39 frames) (b) "Tennis" (47 frames) (c) "Church" (37 frames)

Figure 3.8: Three input videos. (a),(b),(c) are the first frame of three input videos repetitively.

Table 3.1: Performance comparisons between the single-layer method and the proposed method.

Videos	# Frame	Time (second)		β index	
		Single-layer	Multi-layer	Single-layer	Multi-layer
Car	32	120	45	2.30	3.0
Tennis	47	142	50	2.30	3.08
Church	42	150	56	1.96	2.23

still can give us the two correctly segmented moving objects. In contrasty, single-layer algorithm will give us dramatically changed segmentation results. In proposed algorithm, object number estimation is depend on middle-level feature: "trajectory", which is more closely related to the number information of moving objects than low-level pixel-wise feature such as color, position. Therefore, trajectory similarity-based model order estimation method of the proposed algorithm is robust to preset initial value of Gaussian model order. Low-level pixel-wise feature based MDL method can only give a estimation of how many homogeneous regions existing in a video sequence. We know that homogeneous region number has little relation with the number of semantically meaningful object in a complex video sequence. Our experiment shows that low-level pixel-wise feature based MDL method is also very sensitive to the preset initial value of Gaussian model order. Our experiment result of MDL-based GMMs model order estimation method on the video sequence: Car, Tennis, and Church as shown in Fig. 3.9.

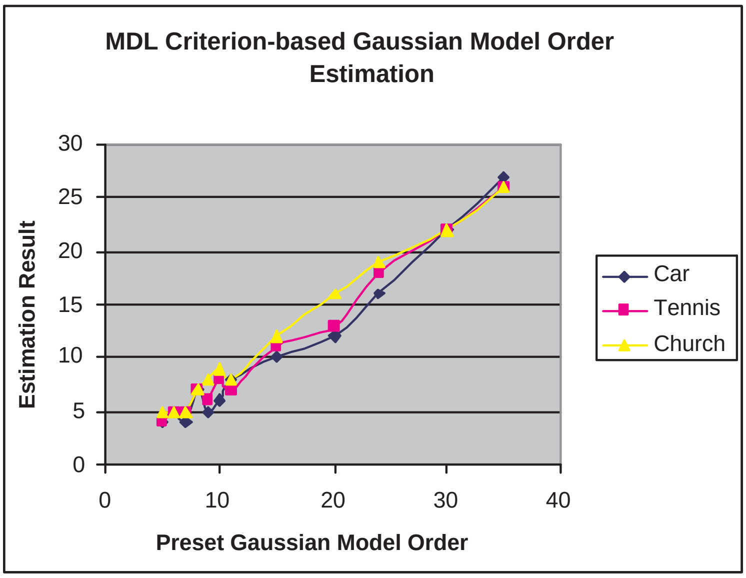


Figure 3.9: MDL-based GMMs model order estimation for three videos. The horizontal axle shows the initialization values for a Gaussian model order. The vertical axle shows the estimated model orders.

From Fig. 3.9, we can see that the MDL-based Gaussian model order estimation result almost has a linear relation with the preset Gaussian model order. This experimental result further support our discarding MDL-based Gaussian model order estimation method. In our algorithm Gaussian model order is affected only by the desired size scale of segmented background regions.

3.9.2 Non-convex Classification

As discussed earlier, we hope our algorithm can address the non-convex classification problem, which is one of the main limitations of single-layer algorithm. In this work, we classify non-convex classification problem of GMM-based clustering into two categories. In general, two kinds of non-convex classification problem exist in both background and moving objects. For background, only the first kind of non-convex classification problem is addressed in our algorithm. That is, static blobs are merged by the similarity of Gaussian models. For moving blobs, mid-level feature: trajectory will be extracted. A trajectory-based joint spatial-temporal grouping will be used to merge dynamic blobs into meaningful moving objects. In the following section, we will discuss background segmentation results and foreground segmentation results separately.

Background Segmentation Results

In this section, The goal of our experiment is to evaluate how well the proposed multi-layer segmentation framework can mitigate the first kind of non-convex classification problem in background segmentation. Both objective and subjective evaluations are conducted.

In order to make a quantitative comparison between the proposed algorithm and the single-layer algorithm proposed in [74], β index is calculated in every simulation. β index is the ratio of total variation and inter-region variation. It is a widely used method for classification evaluation and was first introduced by Fisher in [87]. β index is defined as:

$$\beta = \frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\|X_{ij}-\overline{X}\|}{\sum_{i=1}^{c}\frac{n_i}{n}\times\frac{1}{n_i}\sum_{j=1}^{n_i}\|X_{ij}-\overline{X_i}\|}, \tag{3.8}$$

where is n is the size of a video; n_i is the number of pixels in region i $(i = 1, 2, ..., c)$; $X_{i,j}$ denotes the feature vector of jth pixel $(j = 1, 2, ..., n_i)$ in region i; $\overline{X}$ represents mean feature vector of the video; $\overline{X_i}$ is the mean of n_i feature vectors of region i. Since the numerator is constant for an video, the value is dependent only on the denominator. The denominator decreases with increase in homogeneity in the region. Therefore, for a given video sequence and c value(number of region), the higher the homogeneity within the segmented regions, the higher would be the β value. The value of β also increases with cluster number c. For the same cluster number, the higher the homogeneity within the segmented regions, the higher would be the β value. Compared with the GMM-based method proposed in [74], numerical comparisons are presented in Table 3.1. Simulation results show that β indexes from multi-layer method are larger than the ones from the single-layer method. This means that, at the same cluster number, the proposed multi-layer algorithm can give a more homogeneous segmentation than single-layer algorithm.

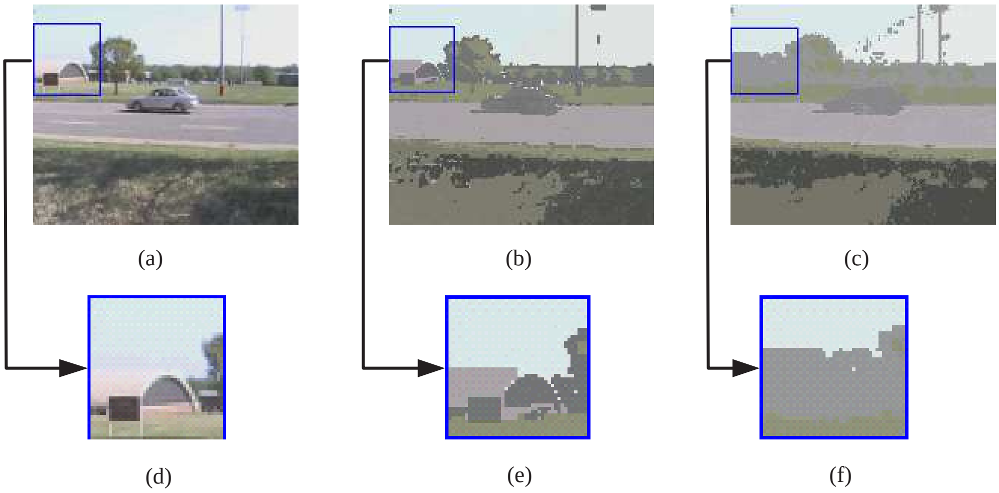


Figure 3.10: (a) is one frame of original video sequence: "Car"; (b) is the segmentation result of the proposed multi-layer method; (c) is the segmentation result of single-layer method; (d) is the extended building region of (a); (e) is the extended building region of (b). (f) is the extended building region of (c). (Every color represents one segmentation component)

As we mentioned earlier, after mitigating the first kind of non-convex classification problem, segmentation results should have a more accurate boundary. Our experiment in the video sequence "Car" verify that the proposed algorithm truly can do so. From simulation results of proposed algorithm as shown in Fig. 3.10(b)(e), we can find the more clear-cut profile of the building and more clear boundary between trees and sky than in Fig. 3.10(c)(f), which are generated by the single-layer algorithm. From Fig. 3.10(c), we can also find that the single-layer algorithm produces a lot of segmentation noise in the sky. These noise lay in the boundary region of two different color sky. In general, the proposed multi-layer classifier can produce more accurate background boundaries. As we mentioned before, the reason for above results is that the multi-layer classification framework can deal with the first kind of non-convex classification problem and generate more homogenous segmentation results. Our observation is associated with quantitative analysis of β index.

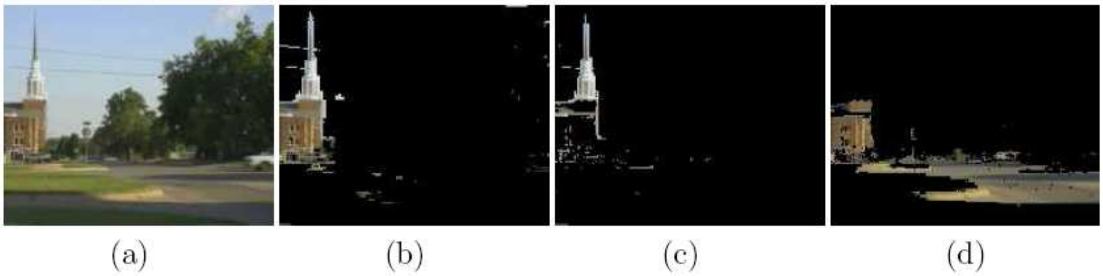


Figure 3.11: The first frame segmentation results of the church building in video sequence "Church". (b) is one of the segmented components of the proposed algorithm. (c) and (d) are the two segmented components of the single-layer algorithm. Obviously, (d) is an under-segmented region, part of church building and road are grouped into one component, which has no semantical meaning

For video sequence "Church", the church building has a non-convex feature space. It is composed of three different parts: gray color tower peak, white color tower roof and red color building body. The proposed algorithm successfully merging these three parts of church building into a one component, which has semantical meaning: church building, as shown in Fig. 3.11(b). While single-layer algorithm gives us an under-segmented region, part of church building and road are grouped into one component, which has no semantical meaning , as shown in Fig. 3.11 (d).

For video sequence "Tennis", a desired background segmentation result is the one, in which, outside court regions should be segmented from inside court regions. From the simulation results shown in Fig. 3.12(only one frame background segmentation

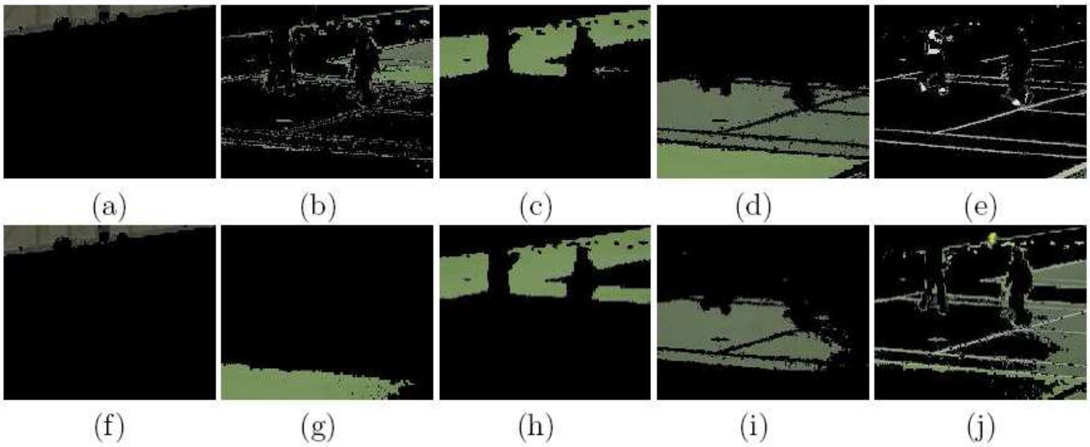


Figure 3.12: The background segmentation results for video sequence: Tennis. The first row (a),(b),(c),(d),(e): segmentation results of single-layer algorithm ; the second row (f),(g),(h),(i),(j): segmentation results of the proposed algorithm.

results), we can see that the proposed multi-layer algorithm can achieve this goal (As shown in Fig. 3.12 (g), (h), and (i)), while the single-layer algorithm fail to do so (As shown in Fig. 3.12 (c) and (d)). Obviously, (h) represents a more complete outside court region than (c) does. From (b) and (d), we can see that single-layer algorithm has not segmented the outside tennis court region from the inside tennis court region. It gives under-segmentation results, which include both inside and outside court regions.

Therefore, not only from β index, but also visually, we can see that the proposed MST-based background region merging framework has better performance than single-layer algorithm. After mitigating the first kind of non-convex classification problem, at the same preset component number, proposed algorithm gives us less under-segmented results than the single-layer algorithm. In general, over-segmented building blocks can severe as better entry-level units for higher-level video processing than under-segmented blocks. With certain component number, we hope to get over-segmented building blocks as long as possible at the low-level stage of video segmentation.

Moving Object Segmentation Results

The goal of our experiments on moving objects is to test whether or not the second kind of non-convex classification can be mitigated by trajectory-based merging process. The proposed algorithm should result in more semantically meaningful segmentation results for moving objects. In order to test object segmentation results, simulations have been carried out on several video sequences, i.e., Tennis, Multi-car, and Multi-Pedestrian.

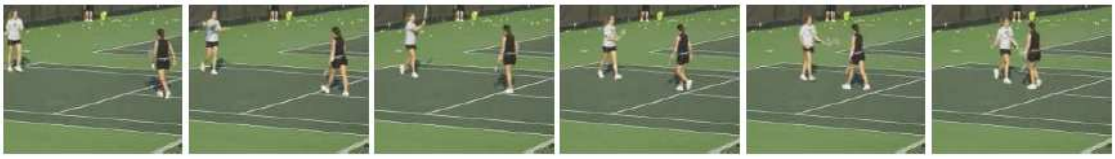


Figure 3.13: Video sequence: Tennis

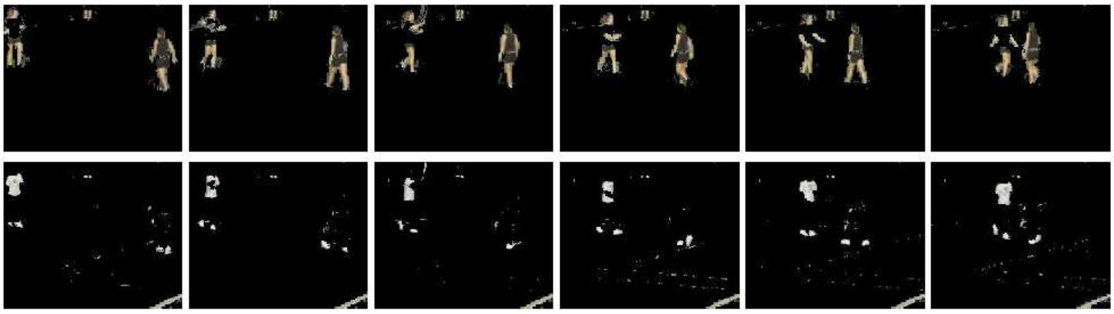


Figure 3.14: Segmentation results of single layer classification.

In the video sequence “Tennis”, as shown in Fig. 3.13, the two tennis players wear the same color shots and have same color skin. In a desired object segmentation, the two player should be segmented from each other, in other word, one segmented blob content one complete player. In order to achieve the desired segmentation, the second kind of non-convex classification problem must be solved. Since the single layer algorithm can not deal with the non-convex classification problem, from its segmentation results as shown in Fig. 3.14, we can see that the detected two dynamic blobs are under-segmented results, which have few semantical meaning.

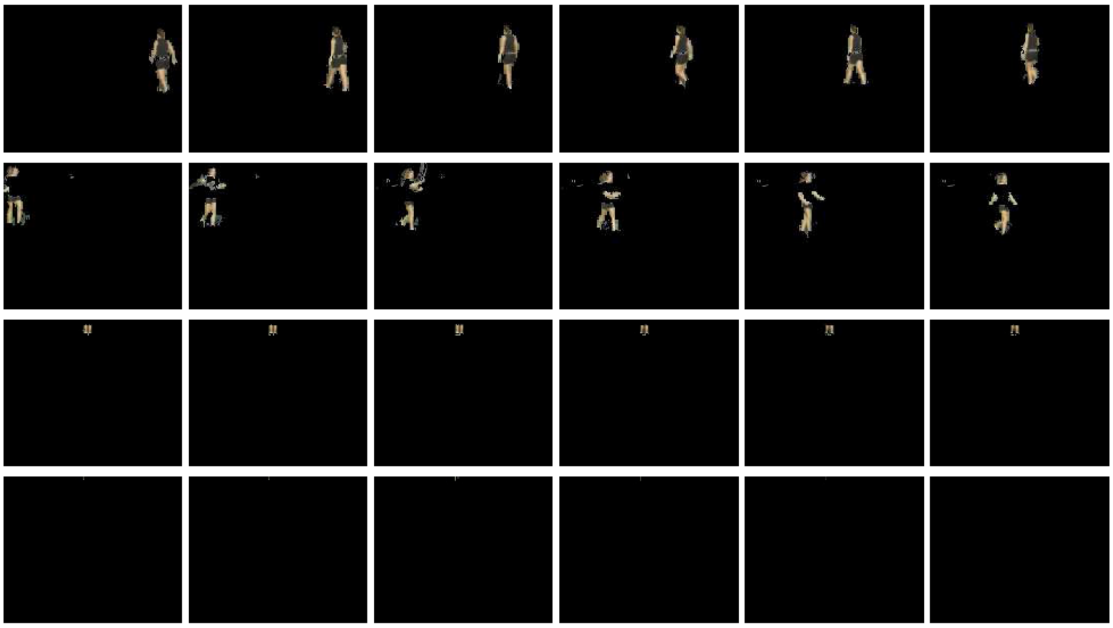


Figure 3.15: Connectivity-based splitting results of the second layer classification.

In the proposed algorithm, single-layer GMM modeling based classification is only the first stage process, which associate with low-level vision. After connectivity-based splitting in the second layer classification, we get splitting results for the two under-segmented dynamic blobs as shown in Fig. 3.15 and Fig. 3.16 respectively.Based on trajectory based merging at the third layer classification, we get results as shown in Fig. 3.17. Obviously, this is our desired segmentation results. As shown in the first and second rows of Fig. 3.17, every moving object (tennis player) is correctly and completely segmented out as one segmentation component.

The third row of Fig. 3.17 deserved further detail description. They are the results of false-positive classification. These objects belong to background, but mistakenly classified as moving objects by figure-ground process in the second layer operation. After connectivity-based splitting in the second layer, these false-positive classification results are separated from moving objects. After middle-level feature: trajectory, is

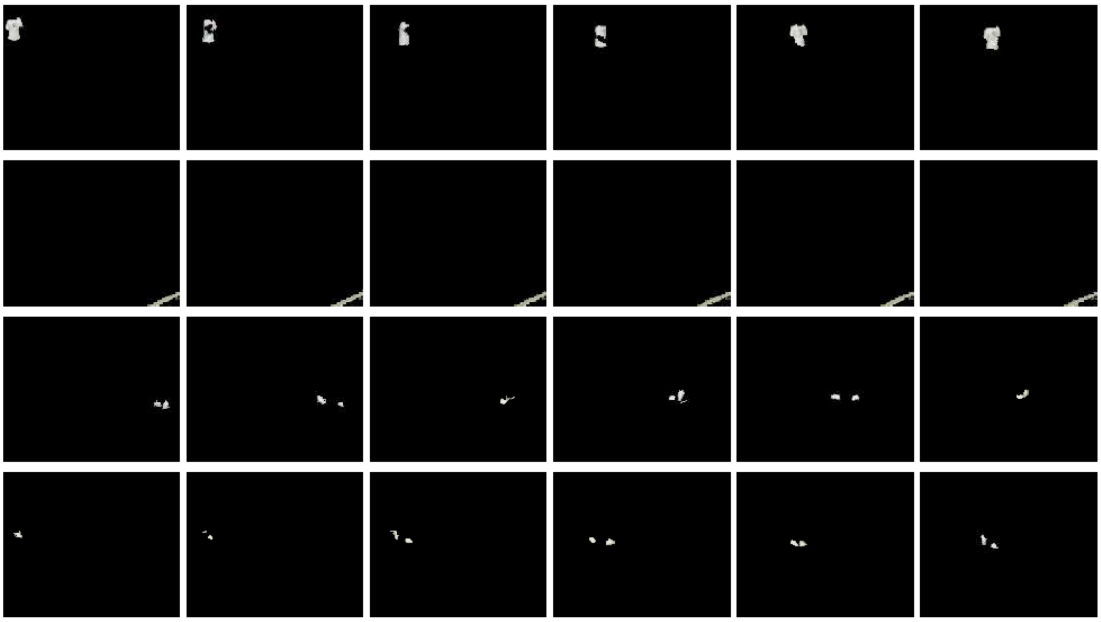


Figure 3.16: Connectivity-based splitting results of the second layer classification.

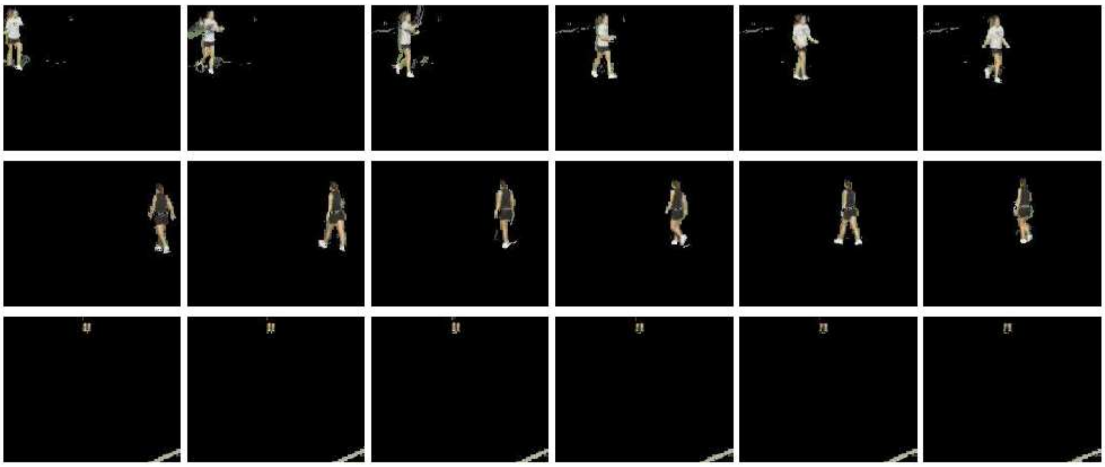


Figure 3.17: Segmentation results of proposed algorithm. The first and second row are moving objects. The third row are background components, which are miss classified into dynamic blobs at the first layer. The fourth row are the noise of dynamic blobs.

extracted, trajectory-similarity based grouping process can detect and group all these miss-classified background objects together, because they share the common character of a static background object: little position changing. This correction process can also be found in visual perception of HVS. It associates with the feedback from higher-level grouping process to figure-ground process as shown in Fig. 2.2. This feedback correction process can benefit our background extraction processing, and result in a more complete background image. This statement is supported by our experiments results as shown in Fig. 3.18. Fig. 3.18 (a) is the extracted background of the single-layer algorithm; (b) is that of the proposed algorithm.

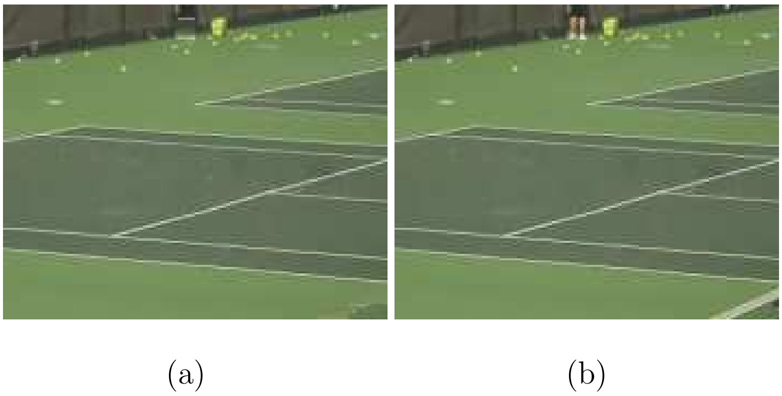


Figure 3.18: Background extraction. (a) is the extracted background of the single-layer algorithm; (b) is that of the proposed algorithm.

The video sequence "Multi-car" is shown in Fig. 3.19. There are four vehicles in this video sequence. The bus has a non-convex feature space. The results of single layer algorithm and proposed algorithm are shown in Fig. 3.20 and Fig. 3.21 respectively. By connectivity-based splitting and trajectory-similarity based space-time grouping, the top part and bottom part of the bus are merged together. Since both motion information (trajectory similarity) and space information (connectivity) are involved in our merging process; merging process happens only in two space-connected blobs, these four vehicles are correctly segmented out in different components, notwithstanding they have very similar trajectories.

A more experiment is conducted on the video sequence "Multi-pedestrian", as shown in Fig. 3.22. Segmentation results of single layer algorithm and proposed algorithm are shown in Fig. 3.23 and Fig. 3.24 respectively. The proposed algorithm truly can mitigate the second kind of non-convex classification problem, and achieve more semantically meaningful segmentation.

3.10 Conclusions

Guided by a biologically plausible computational model, in this chapter, we extended single-layer statistical model video segmentation algorithm into a cascaded multi-layer classification framework. For background segmentation, the first kind non-convex classification problem was mitigated. Its segmentation results are more homogenous and less likely to be under-segmented than the single-layer algorithm. For dynamic objects segmentation, both short-range and long-range motion information are used in classification. By combining the merits of statistical modelling and graph-theoretic approaches, the second non-convex classification problem can be attacked by the proposed algorithm. Therefore, more semantically meaningful segmentation results can be obtained, which can better support many content-based applications or higher-level video processes, such as object recognition or behavior modelling. Experimental results show that this cascaded multi-classifier approach is also computationally efficient.

In the future, we need to look for more robust parsing process to deal with the occlusion problem in the step of connectivity-based splitting. We know that vision perception is a cascade processing, where high-level knowledge based top-down inference play an important role for semantical video segmentation. How to develop a top-down feedback loop or inference to guide low-level and mid-level bottom-up classification will be discussed in the following several chapters.

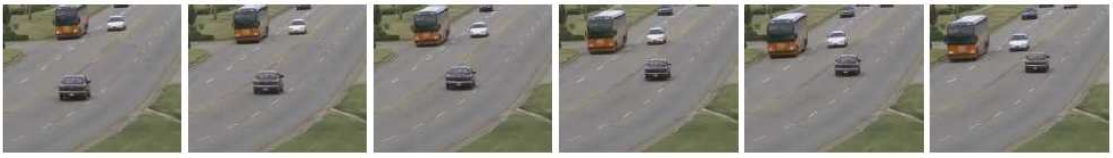


Figure 3.19: Video sequence:Multi-car

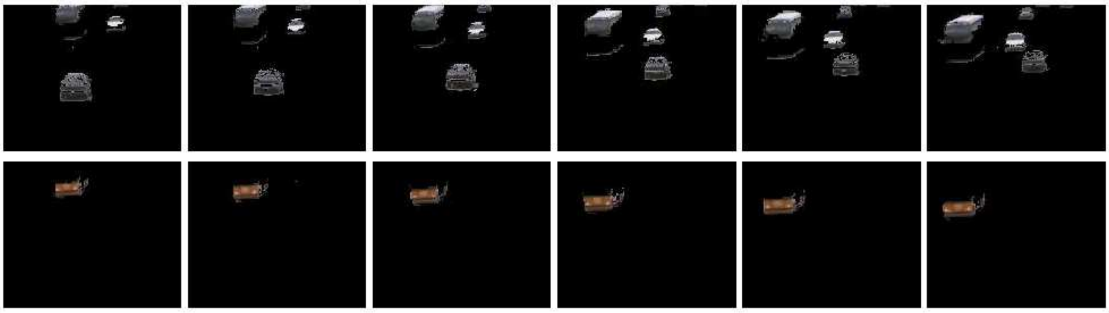


Figure 3.20: Segmentation results of Single-layer framework for video sequence:Multi-car

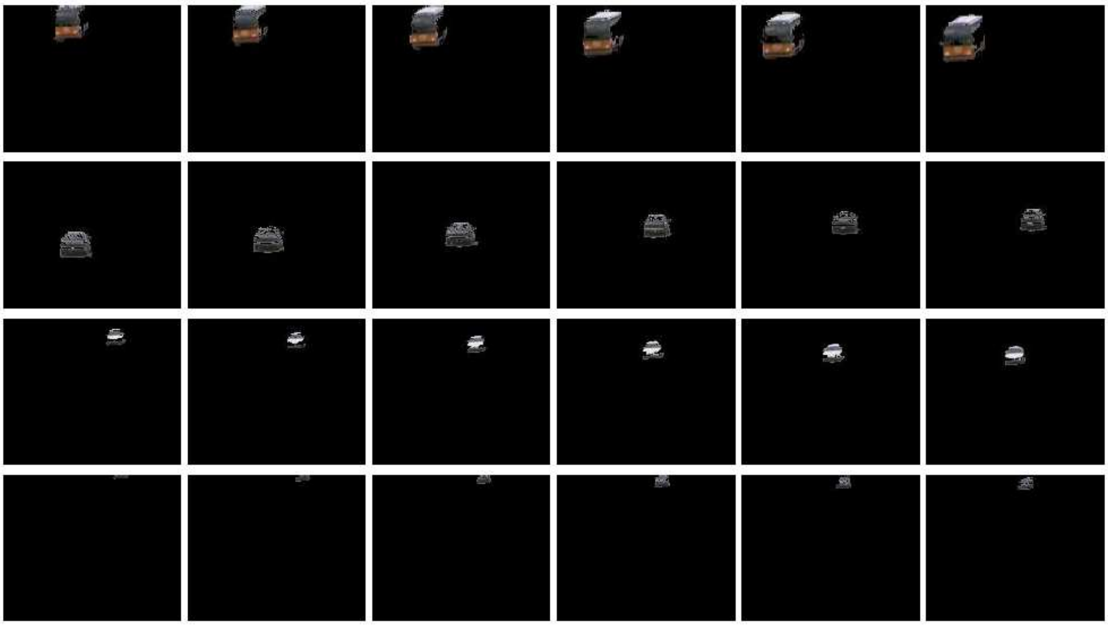


Figure 3.21: Segmentation results of the proposed framework for video sequence: Multi-car.

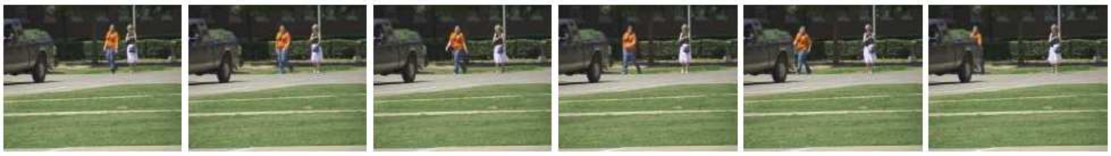


Figure 3.22: Video sequence:Multi-pedestrian

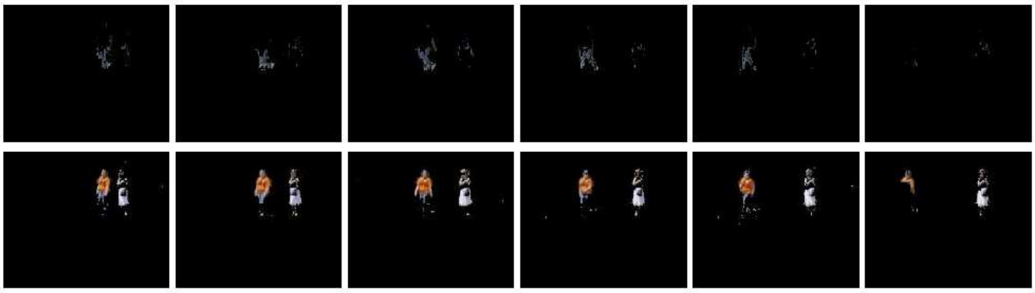


Figure 3.23: Segmentation results of Single-layer framework

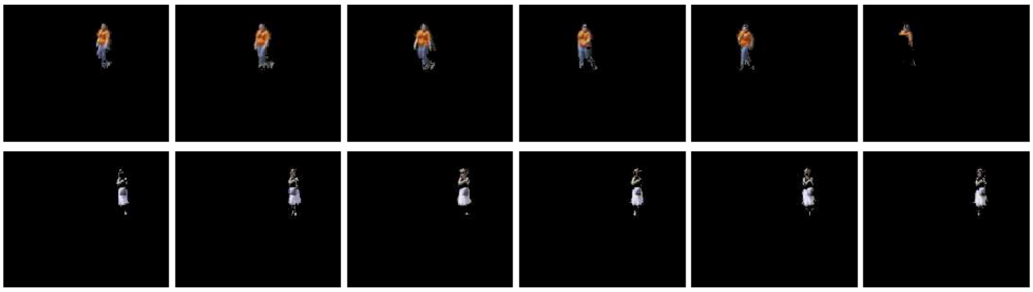


Figure 3.24: Segmentation results of proposed framework

CHAPTER 4

Middle-level Vision: Part Detection

4.1 Overview

In middle-level vision, our goal is to group small UC regions into more semantic meaningful regions, such as body parts, and get a confidence map for each body part. In other words, we want to jointly obtain localization and segmentation of human body parts. However, segmentation itself is an important and long-standing research topic in the fields of image analysis and computer vision, and it can be done at different levels. At low-level vision, it is called *image segmentation* that is to group pixels into regions of homogeneous properties based on various low-level region-based cues (e.g., intensity, color, or texture) and/or edge-based cues (e.g., boundaries or local gradients). Combining both region-based and edge-based cues has led to significant successes for image segmentation due to their complementary nature [86]. At mid-level or high-level vision, segmentation is usually referred to as *figure/ground segmentation* that is to partition an image into foreground and background regions [88], where object-specific priors are usually involved, such as a shape prior. Most current shape constrained segmentation methods, such as [53, 89, 90, 91, 92], require manual initialization of the object configuration (position and orientation). As a separate but related topic, object localization is usually discussed outside the context of segmentation. Our research goal is to integrate localization and figure-ground segmentation into one unified framework where the two tasks can be jointly formulated and optimized in a synergistic way.

In this chapter, the issue of joint localization and figure/gorund segmentation is formulated as a Bayesian estimation problem where we search for the optimal configuration and segmentation of an object of interest (OOI) in the sense of maximum *a posteriori* (MAP). Different from some recent techniques where figure/ground segmentation is usually optimized in a spatially implicit fashion, our objective function is directly defined and optimized in the 2-D spatial space and provides spatially explicit indication of the existence of an OOI. In particular, we resort to a segmentation-based *hypothesis-and-test paradigm* where the coupled region-edge shape priors are involved with two different but complementary roles. Specifically, the region-based shape prior is used to form a segmentation (given a configuration hypothesis), while the edge-based shape prior is used to evaluate the validity of the formed segmentation (in terms of the similarity and smoothness of the boundary). It is believed that *a correct location hypothesis will encourage a valid shape-constrained segmentation while a valid segmentation will enhance the confidence of the location hypothesis.* This makes the proposed algorithm a suitable tool for mid-level vision computation in two ways. First, the prior knowledge about object configuration can be directly used to prune the search space, such as in the video tracking case, where the object configuration at the previous frame provides useful contextual information for the present frame. Second, the algorithm outputs a *map* image that indicates the likelihood of an OOI at each pixel location in an image.

Additionally, we propose two techniques that ensure the efficiency and effectiveness of the proposed algorithm. Specifically, at the *hypothesis* stage where the region-based shape prior is used, a new *semi-parametric* kernel-based color model learning method is proposed that can efficiently learn the figure/ground color models *online* at each hypothesized location and support effective figure/ground segmentation. At the *test* stage where the edge-based shape prior is used, we develop a *mixed* edge-based evaluation criterion that measures both the similarity and smoothness of the formed boundary and is helpful to reject false positives with rugged boundaries for an OOI with a smooth boundary. Our study is focused on mid-level vision, where intermediate results are obtained that can support various high-level vision tasks. As a case study, the proposed method is examined in the context of body part detection that has many applications for human detection and tracking as well as pose recognition and localization [5].

4.2 Related works

There is rich literature on localization and figure/ground segmentation. We will present a brief review from two different but related perspectives, a *methodological review* and a *technical review*. The former one focuses on the background and development of this area, and the latter one discusses the technical connections and distinctions between the proposed algorithm and existing ones.

4.2.1 Methodological Review

Broadly speaking, recent works on figure/ground segmentation and localization can be classified into three categories, i.e., the *bottom-up dominant* approaches, the *top-down dominant* ones, and the *combined bottom-up/top-down* ones. As a bottom-up dominant approach, Srinivasan and Shi [93] provided a set of bottom-up parsing rules to segment human body parts guided by a parse tree. Mori et al. [94] used the contour, shape, shading, and focus cues to find the body parts by searching the optimal segmentation from all possible combinations of super-pixel segments according to a scale constraint defined by a rectangular-shaped bounding box. Wang et. al. [95] proposed a shape-based object recognition and image segmentation algorithm where a shape prior is represented in a multi-scale curvature form. Target objects are identified and segmented by grouping over-segmented image regions in a probabilistic way that is influenced by the image information and the shape similarity constraint. Generally speaking, the *bottom-up dominant* approaches do not depend on a well defined object model, so it is robust to shape variability due to different views or poses. The other side of the same coin is that false positives or negatives may occur because of the limitation of using low-level features only. In contrast, the *top-down dominant* approaches rely on how good an object model matches the OOI in an image. Borenstein and Ullman proposed a fragment-based object representation in [96] that is to cover as closely as possible the images of different OOIs from a given class using a set of primitive shapes. This representation was used for combined object recognition and segmentation in [97] where a probabilistic segmentation map can be computed as the output of top-down inference. As noted in [96], bottom-up cues could be used to refine the boundary of a segmented OOI. This kind of extension naturally leads to the *combined* approaches.

A combined bottom-up/top-down approach usually involves bottom-up features as well as an informative object representation that can be encoded in a global template-like view [98] or a set of fragments [88]. The template-like representation has global shape information that well suits figure/ground segmentation. In order to accommodate more shape variability, various deformable template models have been developed recently [99]. The fragment-based object representation introduced in [96] was used for combined top-down/bottom-up segmentation in [88], where Yu and Shi proposed an integration model to integrate bottom-up pixel grouping and top-down patch matching. It was shown that incorporating bottom-up constraints improves the boundary smoothness of the segmented OOI compared with top-down dominant methods and reduces the false positives/negatives compared with bottom-up dominant approaches.

The combined bottom-up/top-down approaches could be further classified into two classes according to how the OOI is localized. The methods of the first class require manual initialization, such as the active contour model based approaches [100], which is widely used in medical image analysis. The methods of the second class can obtain segmentation and localization simultaneously. Usually, these approaches jointly formulate the two tasks by defining one optimization problem where both low-level features and top-down priors are integrated into an objective (energy) function, e.g., [101, 102, 54]. There are two ways to optimize the energy function. The first way is to optimize it in a spatially implicit space via statistical modeling and inference, such as Conditional Random Field (CRF) [101] [102], or Monte Carlo Markov Chain (MCMC) methods [54]. The second way is to optimize it in a spatially explicit space, such as the hypothesis-and-test approach proposed in [103] that searches through the 2D space to find the global solution. This kind of optimization will facilitate the incorporation of spatial priors and provide an intermediate and spatially sensible outputs for high-level vision tasks.

4.2.2 Technical Review

As a combined top-down and bottom-up approach, our method is inspired by prior research. First, we adopt a segmentation-based hypothesis-and-test paradigm that is similar in spirit with [103] where region-based segmentation is used as an intermediate step for object detection and recognition. While segmentation here is not only the *approach* but also the *goal* where coupled region-edge shape priors are involved. Second, we use the super-pixel-based image representation. Unlike [104, 94, 93] where Normalized-cut is used, and we adopt the watershed transform to create super-pixels with well defined boundaries that are essential for edge-based evaluation. Third, for each segmentation hypothesis, an evaluation is involved to examine its validity. Unlike [94] where the scale constraint defined by a bounding box is used for segmentation optimization in a spatially implicit space, we involve edge-based segmentation evaluation in the 2-D spatial space.

Our approach has two features that make it especially suitable for mid-level vision. (1) In general, a segmentation-based hypothesis-and-test approach is computationally expensive. We propose an efficient kernel-based learning technique to online learn the figure/ground color models at each hypothesized position that can take advantage of the super-pixel image representation. (2) To reduce false positives for OOIs with smooth boundaries, we develop a mixed edge-based evaluation criteria that combines the similarity and smoothness. The synergistic use of the coupled region-edge shape priors is the highlight of this work. The algorithm outputs a spatially sensible *map* image that can be further used for various high-level vision tasks.

4.3 Overview of Our Approach

Our fundamental assumption is that *the optimal shape-constrained segmentation that maximizes the agreement with the edge-based shape prior occurs at the correctly hypothesized location.* The research overview is presented in Fig. 4.1. In this work, joint localization and segmentation are formulated as a Bayesian estimation problem that can be optimized in the 2D space by a hypothesis-and-test approach.

4.3.1 Problem Formulation

We represent an input image G by a set of super-pixels i.e., $G = \{\mathcal{C}_i | i = 1, 2, ..., N\}$. Given an OOI, its shape prior has two components, i.e., the region-based shape prior Y_r and the edge-based shape prior Y_e, both of which are learned together from a set of training images where the OOI has been manually segmented. Y_r and Y_e have a complementary nature for shape representation. Specifically, Y_r is effective for region-based segmentation by grouping multiple super-pixels, and Y_e is efficient for edge-based evaluation to examine the boundary of the formed segmentation. Assuming an OOI is present in an image, we look for the OOI configuration L^* (the position and orientation) and the optimal segmentation X^* by maximizing the posterior probability as follows.

$$\{X^*, L^*\} = \arg\max_{X,L} P(X, L | G, Y_e, Y_r). \tag{4.1}$$

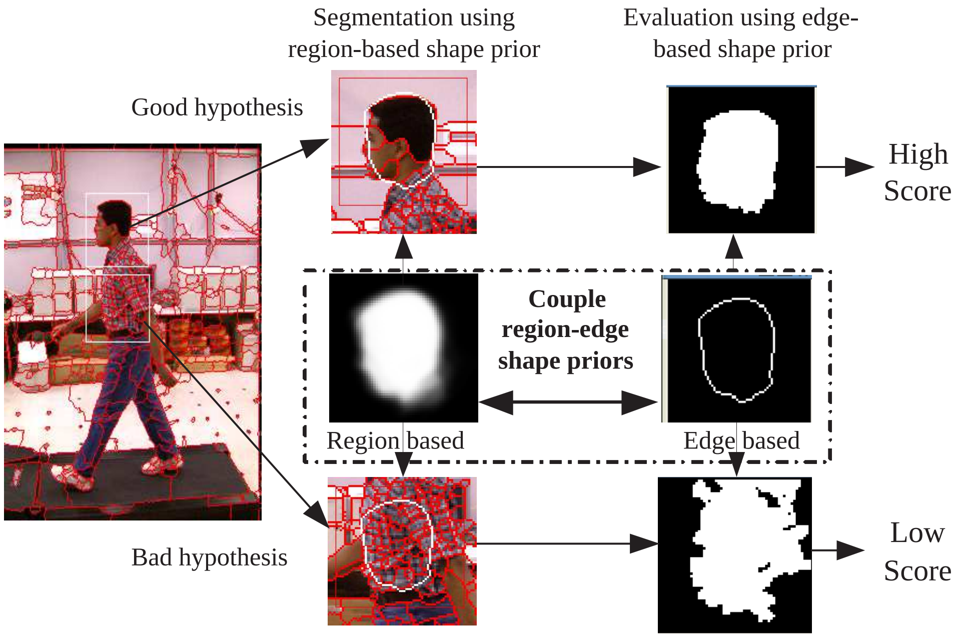

Segmentation using region-based shape prior
Evaluation using edge-based shape prior
Good hypothesis
High Score
Couple region-edge shape priors
Region based
Edge based
Bad hypothesis
Low Score

Figure 4.1: The segmentation-based hypothesis-and-test paradigm where coupled region-edge shape priors are involved with two different roles, i.e., *forming a segmentation* and *evaluating a formed segmentation*. The algorithm outputs a spatially sensible *map* image that reveals the possibility of the existence of an OOI in each position.

Using Bayes' law, the joint probability of a segmentation X with a specific configuration L is written as:

$$P(X, L|G, Y_e, Y_r) = \frac{P(Y_e|X, L, G, Y_r)P(X, L|Y_r, G)}{P(Y_e|Y_r, G)}, \tag{4.2}$$

where the denominator $P(Y_e|Y_r, G)$ is a constant depending on the given image G and the learned shape priors, and the second term of nominators can be written as

$$P(X, L|Y_r, G) = P(X, |L, G, Y_r)P(L|G, Y_r). \tag{4.3}$$

For simplicity, we omit the condition on G, and we have:

$$P(X, L|Y_e, Y_r) \propto P(Y_e|X, L, Y_r)P(X, |L, Y_r)P(L|Y_r), . \tag{4.4}$$

Then (4.1) becomes

$$\{X^*, L^*\} = \\ \arg\max_{X,L} P(Y_e|X, L, Y_r)P(X|L, Y_r)P(L|Y_r). \tag{4.5}$$

We interpret (4.5) intuitively as follows. Given the coupled region-edge shape priors Y_r and Y_e, $P(X|L, Y_r)$ models the posterior probability of obtaining a segmentation X that corresponds to a shape-constrained segmentation under Y_r with configuration L. Having a segmentation X, the first term $P(Y_e|X, L, Y_r)$ models the relationship between Y_e and X conditioning on L. It can be approximated by the edge-based evaluation of the formed segmentation X, and can be further computed by checking the validity of the boundary of X with respect to Y_e. The last term $P(L|Y_r)$ is the prior of the spatial configuration that indicates the possible configurations of the OOI in the image.

4.3.2 Optimization

We develop an effective hypothesis-and-test paradigm to optimize (4.5) that consists of two phases: *hypothesis generation* and *hypothesis test.* The former one generates a hypothesis of object configuration L and creates a corresponding segmentation X based on Y_r specified by L; and the latter one evaluates L by comparing the boundary of the formed segmentation X with Y_e configured by L. In the first phase, a configuration hypothesis can be generated from a prior probability distribution $p(L|Y_r)$ (the last term in (4.5)) that are the main focus for some research, such as [105, 106]. In this work, we assumed $p(L|Y_r)$ to be uniform, indicating a full search strategy. With the help of an efficient online figure/ground color model learning and edge-based evaluation methods, the optimization process can still be computationally feasible in practice.

Object configuration L contains two terms, position L_p and orientation L_r, both of which need to be estimated. A two-step estimation method is used here. First, we marginalize L_r by summing over all possible orientation, then find the optimal L_p^* which maximizes

$$\{\mathbf{X}, L_p^*\} = \arg\max_{X, L_p} \sum_{L_r} P(Y_e|X, L_p, L_r, Y_r) P(X|L_p, L_r, Y_r), \tag{4.6}$$

where $\mathbf{X}$ is a set of candidate segmentations corresponding to different possible orientations at position L_p^*. Second, the optimal rotation L_r^* and the optimal segmentation X^* can be obtained by maximizing

$$\{X^*, L_r^*\} = \arg\max_{X, L_r} P(Y_e|X, L_p^*, L_r, Y_r) P(X|L_p^*, L_r, Y_r). \tag{4.7}$$

The above two-step method is embedded in the two phases of our hypothesis-and-test approach.

4.4 Proposed Algorithm

4.4.1 Watershed-based Super-pixels

There are several commonly used algorithms for super-pixel generation, such as Normalized-cut [104], watershed [107] and mean-shift [108]. Specifically, we choose the watershed transform due to its many "biologically plausible" properties [109]. Moreover, it is fast, local, and has the potential for parallel processing. However, the severe over-segmentation problem is the main concern of using the watershed method. Many studies showed that this problem can be largely mitigated by some preprocessing techniques, such as geodesic reconstruction [110].

Given an input image I, the immersion-based watershed algorithm [107] and geodesic reconstruction preprocessing [111] are used to obtain Z watershed cells $I = \{\mathcal{C}_i | i = 1, 2, ..., Z\}$. Each watershed cell $\mathcal{C}_i$ consist of its pixel members $\mathcal{C}_i = \{p_1^{(i)}, p_2^{(i)}, ..., p_{\eta_i}^{(i)}\}$, where η_i is the number of pixels in the cell. For each watershed cell $\mathcal{C}_i$, we also record its edge pixels by $\Gamma(\mathcal{C}_i)$ that will be used for edge-based evaluation.

Moreover, we use a 3-D Gaussian model $\mathcal{N}(x|\mu_i, \Sigma_i)$ to represent its color distribution in the $L * a * b$ color space. (μ_i,Σ_i) are estimated simply by a maximum-likelihood estimator (MLE) that will be used to online learn the color models for figure-ground segmentation.

4.4.2 Offline Learning of Shape Priors

We use the shape histogram to represent the shape prior, in which the shape prior is embedded implicitly into an "image" [90]. Given a set of manually aligned and segmented OOIs defined in a window Ω, the shape histogram $SH(p)$ can be obtained by adding and these binary image windows followed by appropriate normalization, i.e., $SH(p) \in [0, 1]$ and $p \in \Omega$ is a pixel location in the window Ω where the shape prior is defined. $SH(p)$ reflects the the probability that pixel $p \in \Omega$ belongs to the object, and $1 - SH(p)$ indicates the the probability that the pixel p belongs to the background. Given a threshold ε (say 0.5), an *average* object boundary $\mathcal{M}$ can be extracted from $SH(p)$ by a level-set like method,

$$\mathcal{M} = \{p|SH(p) = \varepsilon\}. \tag{4.8}$$

Therefore, $\mathcal{M}$ defines two regions in Ω, namely the object region $\mathcal{R}_{\mathcal{M}}$ enclosed by $\mathcal{M}$ and the background region $\Omega \setminus \mathcal{R}_{\mathcal{M}}$, as defined below:

$$SH(p) = \begin{cases} \varepsilon, & \text{if } p \in \mathcal{M}; \\ > \varepsilon, & \text{if } p \in \mathcal{R}_{\mathcal{M}}; \\ < \varepsilon, & \text{if } p \in \Omega \setminus \mathcal{R}_{\mathcal{M}}. \end{cases} \tag{4.9}$$

Given $SH(p)$, pixels in $\mathcal{R}_{\mathcal{M}}$ more likely belong to the foreground, and those in $\Omega\setminus\mathcal{R}_{\mathcal{M}}$ the background. Therefore, $\mathcal{M}$ can be used as an edge-based shape prior. Such coupled shape representation by $SH(p)$ and $\mathcal{M}$ facilitates the interface between region-based segmentation (bottom-up) and edge-based evaluation (top-down).

4.4.3 Hypothesis Step: Region-based Segmentation

Given the shape prior $SP(p)$ $p \in \Omega$ for an OOI where Ω is a rectangular window, we can use Ω as a sliding-window to scan through the whole image to examine the existence of the OOI at each location. For a hypothesized location, we use $SP(p)$ to induce a local figure-ground segmentation that is composed by some watershed cells covered by Ω. This segmentation will be used to validate the existence of the object at that location. In order to take advantage of watershed cells and their built-in color models, we propose a new *semi-parametric* kernel-based model learning techniques to online learn the figure/ground color models from the watershed cells directly. We treat the Gaussian model learned from a watershed cell as a kernel center, and learn the figure/ground color models as follows,

$$\begin{aligned} \hat{f_{ob}}(x) &= \sum_{i=1, \mathcal{C}_i \cap \ \Omega \neq \emptyset}^{Z} \alpha_i K_i(x), \\ \hat{f_{bg}}(x) &= \sum_{i=1, \mathcal{C}_i \cap \ \Omega \neq \emptyset}^{Z} \beta_i K_i(x), \end{aligned} \tag{4.10}$$

where x is a color vector; $\mathcal{C}_i$ is one of Z watershed cells that overlap with window Ω; $K_i(x) = \mathcal{N}(x|\mu_i, \Sigma_i)$ is the color model associated with $\mathcal{C}_i$; α_i and β_i denote the contribution of cell $\mathcal{C}_i$ to the object and background respectively that can be calculated from $SP(p)$ $(p \in \Omega)$ and the overlapping watershed cells as

$$\alpha_i = \frac{1}{\mathcal{T}} \sum_{p \in (\mathcal{C}_i \cap \Omega)} SP(p), \tag{4.11}$$

$$\beta_i = \frac{1}{\mathcal{T}} \sum_{p \in (\mathcal{C}_i \cap \Omega)} (1 - SP(p)), \tag{4.12}$$

where $\mathcal{T}$ is the size of shape prior window Ω. Based on the figure/ground color models, we can use the maximum *a posterior* (MAP) criterion to identify the watershed cells that belong to the object. Let τ_i be the class label for $\mathcal{C}_i$:

$$\tau_i = \begin{cases} 1 \text{ (object)}, & \alpha_i \hat{f_{ob}}(\mu_i) > \beta_i \hat{f_{bg}}(\mu_i); \\ 0 \text{ (background)}, & \alpha_i \hat{f_{ob}}(\mu_i) < \beta_i \hat{f_{bg}}(\mu_i). \end{cases} \tag{4.13}$$

Therefore, we can obtain the corresponding segmentation for a position hypothesis as, $X = \{\bigcup \mathcal{C}_i | \tau_i = 1\}$. Different from the one in [103] where the shape prior is used once for online figure-ground color model learning, here we use the region-based shape prior $SP(p)$ twice. The first time is for the online color model learning as defined in (4.10), and the second time is for MAP-based segmentation as defined in (4.13). Considering the false negative is more detrimental than the false positive at mid-level vision, we encourage more object-like segmentations by fully incorporating the region-base shape prior into the segmentation process. This may lead to some false positives due to the double usage of the region-based shape prior. However, the later edge-based evaluation will mitigate this problem.

4.4.4 Test step: Edge-based Evaluation

After segmentation X is formed, we evaluate it according the edge prior $\mathcal{M}$. Let $\Gamma(X)$ to be the boundary of X, we compare $\Gamma(X)$ with $\mathcal{M}$ in terms of shape similarity and boundary smoothness. The score of X with respect to its compliance with $\mathcal{M}$, i.e., $\rho_{\mathcal{M}}(X)$ is given by,

$$\rho_{\mathcal{M}}(X) = \exp(-d_{chamfer}(\Gamma(X), \mathcal{M})) + \zeta(1 - \mathcal{S}(\Gamma(X), \mathcal{M})), \tag{4.14}$$

where the first term is the chamfer distance indicating the shape similarity; the second term measures the boundary smoothness; and ζ balances the relative importance between the two terms. It is expected that a valid segmentation should have a smooth boundary that matches with $\mathcal{M}$ well. The first term is sensitive to the transition, rotation and scale. This is desired for rejecting false hypotheses. The second term aims to reject false segmentations with rugged boundaries for the OOI with smooth boundary.

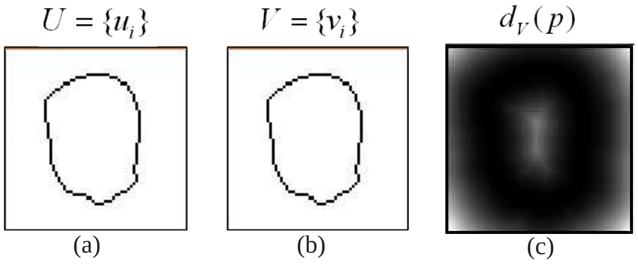


Figure 4.2: The computation of the Chamfer distance via the distance transform. (a) and (d) show two sets of edge pixels, and (c) shows the distance transform between (a) and (b).

Boundary Similarity

In order to eliminate the effect of outliers, we use a modified chamfer distance [112],

$$d_{chamfer}(U,V) = \frac{1}{n}\sum_{u_i \in U} \min(\min_{v_i \in V} \parallel u_i - v_j \parallel, \eta), \tag{4.15}$$

where η is a factor controling the tolerance of mismatching. Furthermore, Equ. (4.15) can be efficiently computed using the distance transform (DT) [113]. Given two sets of edge points $U = \{u_i\}$ and $V = \{v_i\}$ in a window Ω, the distance transform $d_V(p)$ specifies the distance from each pixel $p \in \Omega$ to the nearest pixel $v_i \in V$ (as shown in Fig. 4.2). Therefore the chamfer distance based shape similarity between U and V can be calculated by

$$d_{chamfer}(U,V) = \frac{1}{\#(U)}\sum_{u_i \in U} d_V(u_i), \tag{4.16}$$

where $\#(U)$ denotes the number of pixels in U, and $d_{chamfer}(U,V)$ the average distance between U and V.

Boundary Smoothness

As shown in Fig. 4.3, assume that $\Gamma(X)$ touches n cells $\{\mathcal{C}_1, ..., \mathcal{C}_n\}$, and we define $H_i = \{h_1^{(i)}, ..., h_{n_i}^{(i)}\}$ to be the set of n_i boundary pixels shared between $\mathcal{C}_i$ and $\Gamma(X)$. Let $\phi_{\mathcal{M}}(p) : \mathrm{R}^2 \rightarrow \mathrm{R}$ be the signed Euclidian distance transform that is

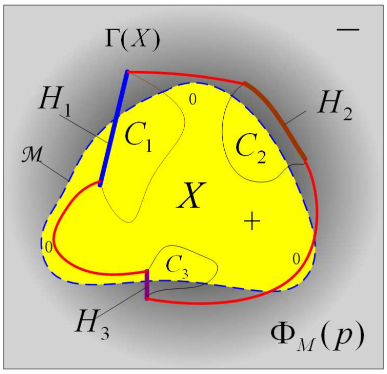


Figure 4.3: The computation of edge smoothness based on the signed Euclidian distance transform of the edge prior $\mathcal{M}$, i.e., $\phi_{\mathcal{M}}(p)$. X is a segmentation and $\Gamma(X)$ is the boundary of X that touches several watershed cells. H_i is the set of edge pixels shared by watershed cell C_i and $\Gamma(X)$. Specifically, H_2 has the high parallelness (good smoothness), while H_3 has low parallelness (bad smoothness).

"+" or "-" for p inside or outside $\mathcal{M}$, respectively. The maximum and minimum distances from H_i to $\mathcal{M}$ are obtained by $d^{(i)}_{max} = \max(\phi_{\mathcal{M}}(h^{(i)}_1), ..., \phi_{\mathcal{M}}(h^{(i)}_{n_i}))$, and $d^{(i)}_{min} = \min(\phi_{\mathcal{M}}(h^{(i)}_1), ..., \phi_{\mathcal{M}}(h^{(i)}_m))$, respectively. The degree of parallelness between $\Gamma(X)$ and H_i is defined as

$$\mathcal{S}_{\mathcal{M}}(H_i) = \frac{d^{(i)}_{max} - d^{(i)}_{min}}{n_i}. \tag{4.17}$$

When H_i is parallel to $\mathcal{M}$ (e.g., H_2 in Fig. 4.3(b)), $\mathcal{S}_{\mathcal{M}}(H_i) \cong 0$, indicating good local smoothness. When H_i is perpendicular to $\mathcal{M}$ (e.g., H_3 in Fig. 4.3(b)), $\mathcal{S}_{\mathcal{M}}(H_i) \cong 1$, indicating poor local smoothness. In general, the smaller the value, the more parallel between H_i and $\Gamma(X)$. Therefore, we define the overall smoothness of $\Gamma(X)$ as

$$\mathcal{S}(\Gamma(X), \mathcal{M}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\mathcal{M}}(H_i). \tag{4.18}$$

If a full search is involved, the score function (4.14) will return a *map* image that records the existence possibility of the OOI at every pixel location. The larger the value, the more likely there is an OOI. It is worth noting that at each position hypothesis, the shape prior is hypothesized with different angles around the mean orientation, and we use the winner-take-all strategy to generate the *map* image. The optimal angle at each location is also recorded.

Essentially, this computation is at mid-level vision, and it could support high-level vision by incorporating high-level knowledge. In this work, we focus on joint localization and segmentation of a non-articulated object with a relatively well defined shape, such as human body parts. In the next chapter, we proposed a hybrid body representation for integrated pose recognition, localization and segmentation, where this algorithm is used as the inference engine at mid-level vision to locate and segment the body parts.

4.5 Segmentation Refining via Graph-cut

Recently, the graph-cut approach has achieved considerable success in image segmentation. It has the capacity to fuse both boundary and regional cues in an unified optimization framework [114]. Several existing methods, such as [89], only incorporate a single shape prior (edge-based or region-based) into the segmentation process. Our contribution here is to combine two shape priors into segmentation where the image is represented by watershed cells.

Given image $I = \{\mathcal{C}_i | i = 1, ..., Z\}$, $\mathbf{l} = \{l_i | i = 1, ..., Z\}$ denotes the set of binary class labels for all watershed cells ($l_i = 0$: background and $l_i = 1$: object). Following the segmentation energy definition from [114]

$$E(\mathbf{l}) = \lambda . \sum_{i=1}^{Z} R(l_i) + \sum_{\mathcal{C}_i \bigcap \mathcal{C}_j \neq \emptyset} E(H_{i,j}) \delta(l_i, l_j), \tag{4.19}$$

where $R(l_i)$ is the regional term, which relates to the posteriori probability of $\mathcal{C}_i$ belonging to class l_i; $E(H_{i,j})$ is the boundary term, which represents the consistence between the edge-based shape prior $\mathcal{M}_w^{L^*}$ and local boundary formed by two cells, $H_{i,j} = \mathcal{C}_i \bigcap \mathcal{C}_j$; $\delta(l_i, l_j) = 1$ when $l_i \neq l_j$ otherwise $\delta(l_i, l_j) = 0$; λ specifies a relative importance between two terms.

The calculation of $R(l_i)$ involves online learning of figure/ground color models where region-based shape prior Y_r is involved for kernel-based density estimation, as discussed in Section 4.4.3. Let $\hat{f}_{ob}^{(w)}(x)$ and $\hat{f}_{bg}^{(w)}(x)$ be the figure/ground color models, and $\alpha_i^{(w)}$ and $\beta_i^{(w)}$ are computed from Y_r that denote the prior probabilities of $\mathcal{C}_i$ belonging to the object and background respectively. Therefore, $R(l_i)$ is defined as

$$R(l_i = 1) = -\ln \alpha_i^{(w)} \hat{f}_{ob}^{(w)}(\mu_i^{(c)}), \tag{4.20}$$

$$R(l_i = 0) = -\ln \beta_i^{(w)} \hat{f}_{bg}^{(w)}(\mu_i^{(c)}), \tag{4.21}$$

where $\mu_i^{(c)}$ is the mean color vector of $\mathcal{C}_i$. Using the same idea of edge-based shape evaluation defined in (4.14), let $X = H_{i,j}$, and we can define $E(H_{i,j}) = \rho_{\mathcal{M}}(X)$, which evaluates the consistence between $H_{i,j}$ and edge-based shape prior $\mathcal{M}$ in terms of the degree of parallelness and the shape similarity.

After object configuration estimation,the using of the improved Graph-Cut method will further improve our segmentation results.

4.6 Experiments

A set of experiments were conducted to validate the proposed algorithm. The algorithm was programmed in C++, and the test platform is a PC with Pentium-IV 3.2GHz CPU and 1GB RAM. Our experiments were based on the CMU Mobo database [115], which contains image sequences of 25 individuals walking on a treadmill. Each image is resized to 240 × 320 pixels. In particular, we are interested in joint localization and segmentation of six body parts, i.e., (the *head*, *torso*, *left-arm*, *right-arm*, *left-leg*, and *right-leg*) as our OOIs, each of which is defined in a window of

61×61 pixels [1]. For each OOI, 200 manually segmented images from six individuals were used for learning the coupled region-edge shape priors, and all training images share similar poses (i.e., recoil/contact) and the same side-view. The reason of using one pose is because that the shape of some body parts (legs and arms) may deform under different poses, and that of using the recoil/contact poses is due to the fact that they have the least occlusion problem compared with other poses. How to handle shape deformation and occlusion is beyond our scope. The coupled region-edge shape priors of six body parts are shown in Fig. 4.4.

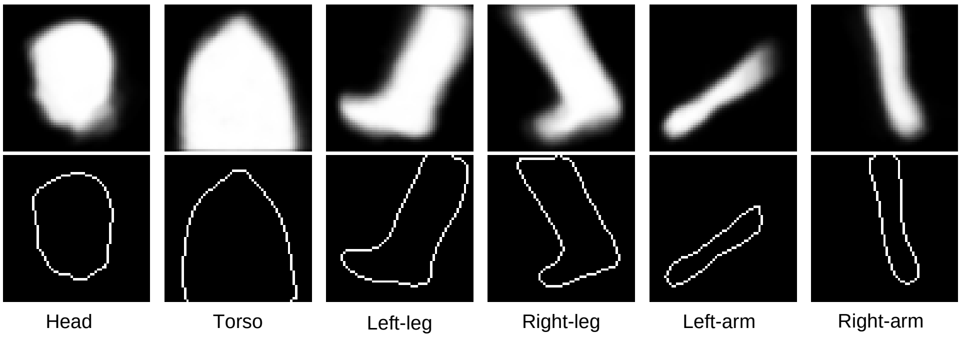


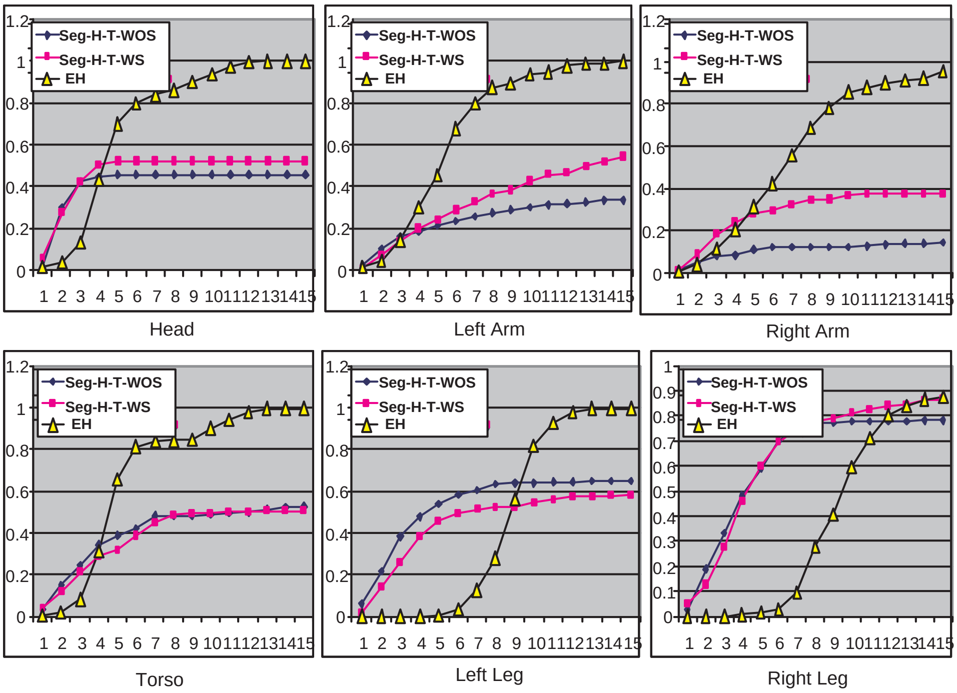
Figure 4.4: The coupled region-edge shape priors for six body parts.

Our algorithm was evaluated in two aspects, i.e., localization and segmentation. For localization, the competing algorithm is the state-of-the-art edge histogram (EH) method in [116] that is a mid-level computation and generates an intermediate localization *map* image for an OOI. For segmentation where the online learning of figure/ground color models is the key issue, we compare our semi-parametric technique with the Fast Gaussian transform (FGT) that is a non-parametric learning technique [117]. 180 test images are from six individuals that have the similar poses with the training data, and we also manually obtained the ground-truth segmentation and localization information for all test images with respect to six body parts.

[1] For simplicity, *Left* and *right* here are defined according to the relative position of two arms or legs to the viewer.

4.6.1 Localization

As a mid-level vision task, our algorithm outputs a *map* image of a given OOI in an image. Its pixel value indicates the likelihood or confidence of the OOI at each pixel location in the image. Given a *map* image, we expect that its maximum value locates at (or close enough to) the ground truth position. In order to evaluate the quality or saliency of *map* images, we define a localization accuracy function that shows the relationship between *the tolerance error* and *the hit rate*. A tolerance error defines an acceptable region around the ground truth position, and the hit rate records the percentage of the optimal values fall in the acceptable region. As shown in Fig. 4.5, our algorithm is compared with the SH method in terms of localization accuracy for six body parts. Specifically, we further analyze our algorithm by considering the case with (Seg-H-T-WS) and without the smoothness term (Seg-H-T-WOS) in the edge-based segmentation evaluation in (4.14).



Head

Left Arm

Right Arm

Torso

Left Leg

Right Leg

Figure 4.5: The comparison of localization results for six body parts.

There are two observations from Fig. 4.5. (1) the Seg-H-T offers better localization precision (i.e., higher hit rates at small tolerance errors) for all six body parts, while the EH method is more robust (i.e., higher hit rates at large tolerance errors). This is understandable due to the nature of the two methods. One uses deterministic edge-based evaluation, and the other involves statistical edge histograms. (2) The smoothness term is more useful for the OOI with smooth boundaries, such as the head and two arms. As for the OOIs with less well defined boundaries, such as the two leg (due to different shoes and pants worn by the subjects), the smoothness term is less useful. It is is possible that the usefulness of the smoothness term may not be fully exploited because the imperfect segmentation (with rugged boundaries) at the true location could be mis-judged by edge-based evaluation.

On the other hand, the EH method is more efficient due to a direct 2D convolution involved, while the Seg-H-T involves a segmentation at each pixel location. Some preprocessing could be used to trim the candidate locations. For example, we used a simple shape matching method that convolves the edge map (consists of all watershed boundaries) with the edge-based shape prior (under different orientations) and selects only 2% pixel locations of the best shape matching. Then it takes about 10 seconds per image for the Seg-H-T, and about 1 second for the EH method. It is interesting to find out that there is little chance for both methods to achieve successful localization in the same image at low tolerance errors. It implies that there is a complementary nature between the two methods, and they could be combined together for more accurate (due to Seg-H-T) and efficient (due to EH) localization.

As mentioned before, the proposed Seg-H-T method accomplishes a mid-level computation, and Fig. 4.5 only partially reveals its advantages over the EH method for localization. We have compared the two methods in their usefulness for high-level vision, such as pose recognition and localization, in [5], where we incorporated the spatial prior of six body parts represented by a “start” model proposed in [116]. It was

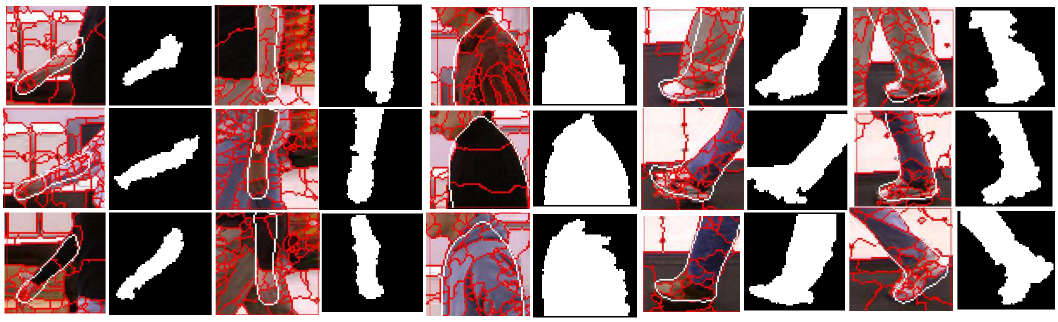


Figure 4.6: Segmentation examples of five body parts (from left to right: the left/right arms, the torso, and left/right legs) where super-pixels and the edge-based shape prior is shown.

shown that the new method provides significant advantages over the EH method for pose recognition and whole-part localization. Here, we show part localization results in Fig. 6.1 which is obtained by averaging the results over a walking cycle, *H-point*, *Contact*, and *Passing*. We can see that significant improvements are achieved for two arms and two legs that undergo major movement during a walking cycle. The proposed Seg-H-T offers salient mid-level outputs, i.e., *map* images, which ensures precise body/part localization. At the same time, we can achieve the segmentation results for each body part given the correct localization which are not available from the EH-based method [116].

Table 4.1: The comparison of localization errors (in pixel) in one walking cycle.

Methods	Head	Torso	L-arm	R-arm	L-leg	R-leg
EH	5.23	6.83	12.17	11.07	12.73	12.77
Seg-H-T	5.07	6.03	9.80	8.63	4.83	5.63
Improvement	3%	11.7%	19.5%	22.0%	62.1%	55.9%

4.6.2 Segmentation

One key idea in our Seg-H-T approach is the semi-parametric kernel-based method for online color model learning that takes advantage of the super-pixel representation and supports efficient figure/ground segmentation at each hypothesized position. This is essential to the computation at mid-level vision. We compare the proposed color model learning technique with the Fast Gauss Transform (FGT) [117] that was used for non-parametric color model learning and tracking in [118]. We have downloaded the online FGT code from [119], and used it for comparative studies. Also, we have implemented the pixel-wise FGT (FGT-P) and super-pixel-wise FGT (FGT-SP). We evaluate the three segmentation algorithms on 180 test images for six body parts within the 7 × 7 region centered around the ground truth position. Segmentation results are evaluated by the ratio of the falsely detected region size (including both false positives and false negatives) to the object size [120]. The experimental results are shown in the Table 4.2.

Table 4.2: The comparison of segmentation errors (%) for different shape prior guided methods at ground truth location.

Methods	Time	Torso	Head	L-arm	R-arm	L-leg	R-leg
FGT-P	150 ms	35.8	31.18	43.59	51.63	34.04	33.35
FGT-SP	25 ms	3.83	1.85	17.74	9.89	7.72	6.55
Our method	2.3 ms	3.08	1.49	15.90	7.77	7.26	6.36

From Table 4.2, we can see that significant improvements can be achieved by using super-pixels in stead of raw image pixels and our method slightly outperforms the FGT-SP approach. One possible reason is that our method uses soft decision while the FGT involves hard decision to learn the color models. Moveover, it costs only about 2.3 ms for our method, and 150 and 25 ms for the FGT-P and FGT-SP methods respectively, making the new method more appropriate in the hypothesis-and-test paradigm. Some segmentation examples of the *head* for the three methods are shown in Fig. 4.7, where the first column shows the input images where the watershed cells and the edge-based shape prior are shown.

More segmentation results for other body parts are shown in Fig. 4.6. There are two main possible causes to the segmentation errors. (1) Some super-pixels are under-segmented which cover both the foreground and background regions (e.g., the arms in the second row); (2) The OOI has an irregular shape that violates the shape prior (e.g., the legs in the first row).

4.6.3 Graph-cut Based Segmentation Refining

Table 4.3: The segmentation comparison.

Methods	Torso	Head	Left arm	Right arm	Left leg	Right leg
Our method	4.52%	4.14%	4.36%	4.1%	5.31%	4.74%
OriGCUT	7.16%	3.85%	4.39%	5.2%	7.18	5.20%

In this chapter, we introduced both region-based and edge-based shape priors into a Graph-cut framework for segmentation through the two terms of the energy function defined in (4.19). In order to evaluate our approach, it is necessary to compare it with the standard method proposed in [114](OriGCUT), where the second term of energy function(4.19) is replaced by a standard color similarity-based boundary penalty term. We performed Graph-cut segmentation with ($\lambda = 1/30$ in (4.19)) on 184 test images for six objects (the *head*, *torso*, *left-arm*, *right-arm*, *left-leg*, and *right-leg*). Segmentation results are evaluated by the ratio of the falsely detected region size (including both false positives and false negatives) to the ground truth region size. The same evaluation method is used in [120]. Statistical segmentation results are shown in Table 4.3. The results of our method are listed in the first row, and the standard method proposed in [114] without using the edge-based shape prior are listed in the second row. From Table 4.3, significant segmentation improvement can be observed for the Torso and Legs. Since the segmentation results of the standard method are already very good for the object of Head, only a slight improvement can be achieved by our method. Fig. 4.6.3,Fig. 4.6.3 and Fig. 4.6.3 show some examples of segmentation results from our method and the standard Graph-cut method proposed in [114]. Watershed cells, which belong to the object, are marked with dot array. It is worth noting that not all the properties of using an edge-based shape prior are desirable. As shown in the (c.) of Fig. 4.6.3, a more smoother segmentation boundary ia obtained at the cost of a false negative segmentation.

4.7 Conclusion and Future Research

In this chapter, we have presented a new joint localization and figure/ground segmentation algorithm that involves coupled region-edge shape priors and is implemented in a segmentation-based hypothesis-and-test paradigm. Specifically, our research focuses on mid-level vision and can produce a spatially sensible *map* image that reveals the possibility of the existence of an OOI in each pixel location. One possible improvement of this work is to introduce a deformable shape model or a more powerful edge-based shape representation, such as the multiscale-curvature shape model used in [95], which could enhance the flexibility and adaptability of edge-based segmentation evaluation. The proposed method can be incorporated into other high-level vision research tasks [121][122][123][124][125], where the human body is modeled as an assembly of body parts. It can also be directly used for part-based human detection, pose recognition, localization, tracking and segmentation.

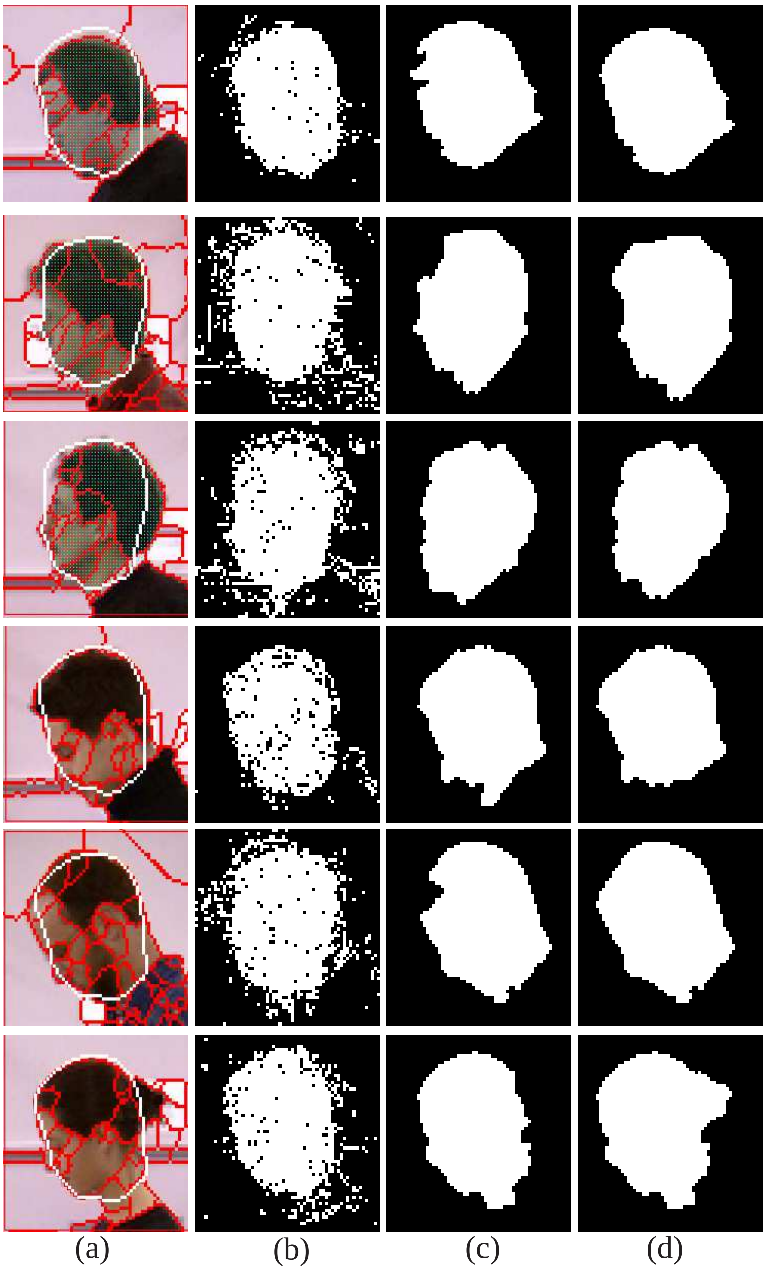


Figure 4.7: Segmentation results for the input images (a) using the FGT-P (b), FGT-SP (c), and our method (d).

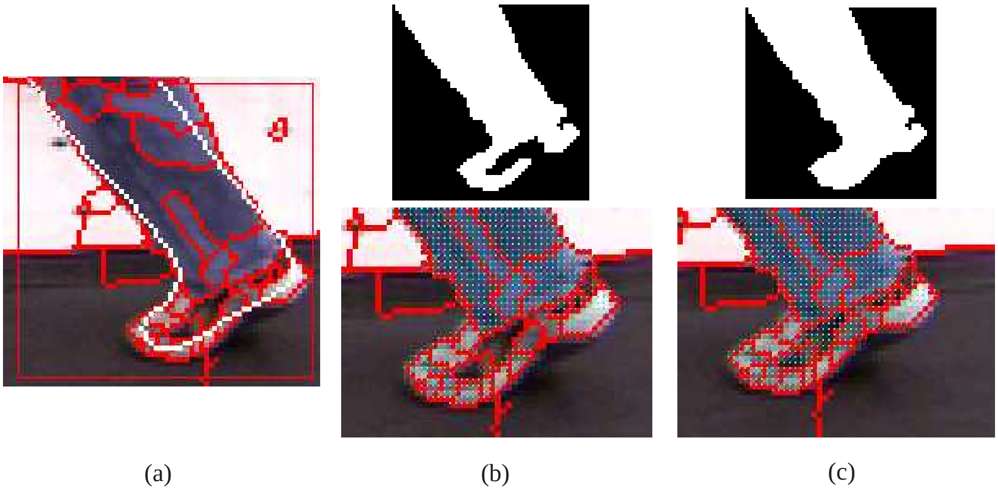


Figure 4.8: Right leg Segmentation results. (a): Edge-based shape prior at the detected configuration. (b): The segmentation results of without using edge-based shape prior. The top one is the segmented binary image in the size of 61x61 pixel; (c): The segmentation results of using edge-based shape prior.

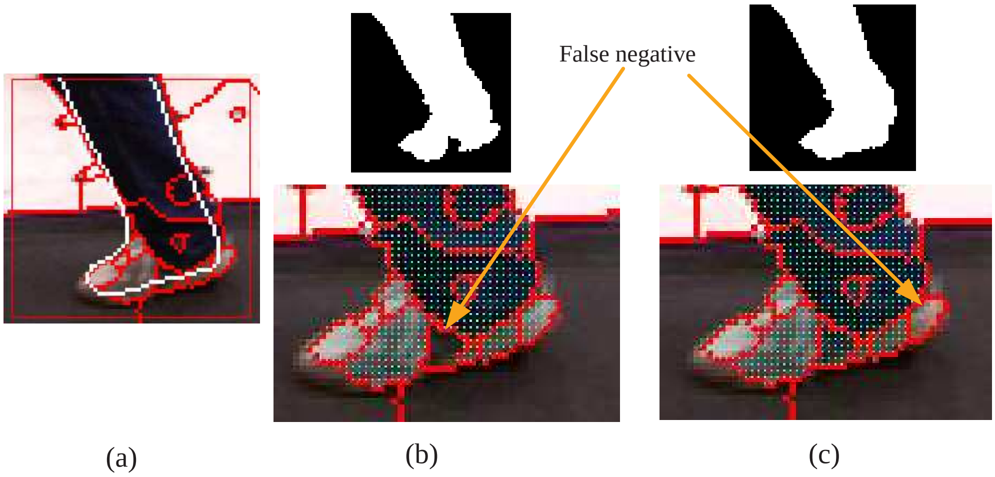


Figure 4.9: Right leg Segmentation results. (a): Edge-based shape prior at the detected configuration. (b): The segmentation results of without using edge-based shape prior. (c): The segmentation results of using edge-based shape prior.

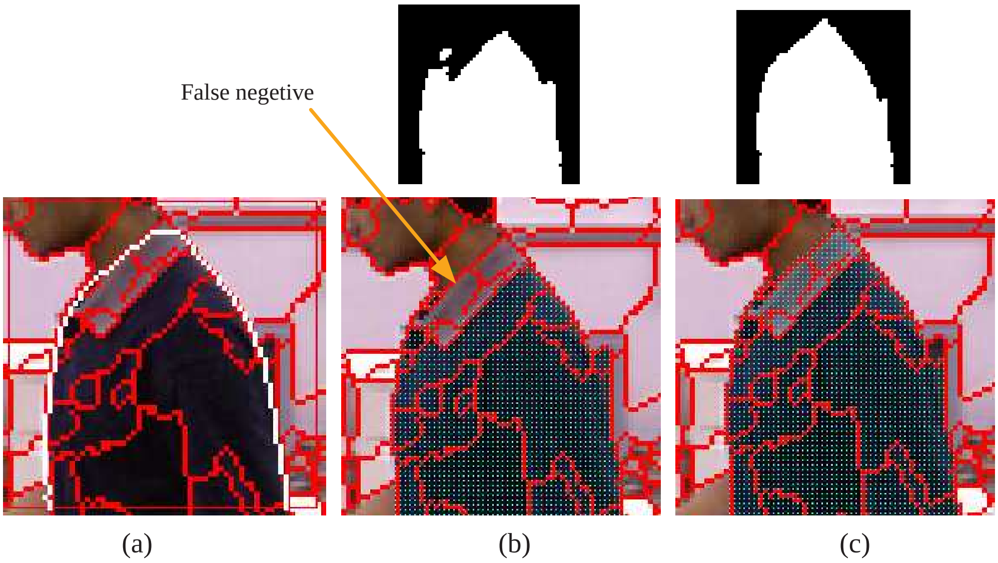


Figure 4.10: Torso segmentation results. (a): Edge-based shape prior at the detected configuration. (b): The segmentation results of without using edge-based shape prior. (c): The segmentation results of using edge-based shape prior.

CHAPTER 5

High-level vision: Recognition, Localization and Segmentation

5.1 Overview

Based on middle-level vision processing output results: *map* images, in this chapter, we consider a comprehensive decision about the position of each body part. We will deal with pose recognition, localization and segmentation of the whole body as well as body parts in a single image. Our objective is to develop a hybrid human representation and the corresponding processing to assemble three tasks into one integrated framework. We propose a *hybrid body representation*, as shown in Fig. 5.1, where the four images show *the input image represented by watershed cells*, *the proposed hybrid body representation*, *the online learned whole shape prior*, and *the part/whole segmentation results*, respectively. Specifically, segmentation is involved for learning and inference, since more and more evidences show that segmentation can boost the recognition and localization performance.

The proposed research is deeply inspired and motivated by shape representation theories in cognitive psychology where there are two prevailing theories, i.e., the structural description-based and the view-based representations [126]. The former one suggests that a complex object is represented by a collection of simpler elements with specific inter-relationships. The latter one postulates a very simple template-like representation in which an objects is holistically represented by a simple vector or matrix feature without an intermediate representational step. Current cognitive studies indicate that none of these two representation schemes alone can provide a complete characterization of the human vision system for object recognition [127].

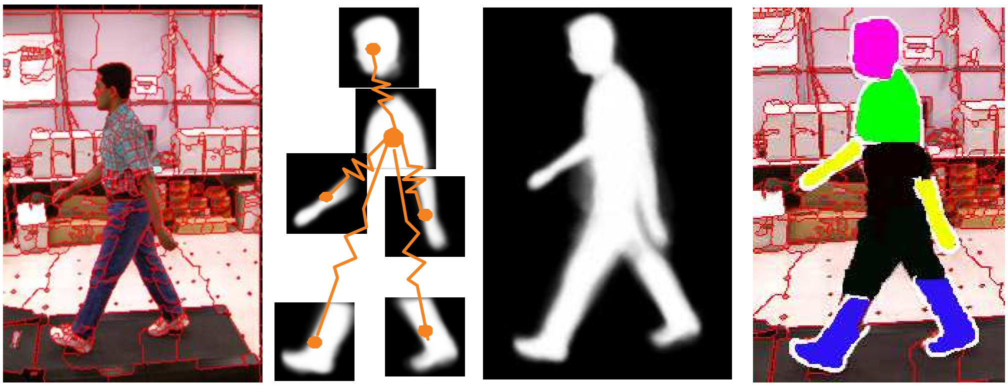


Figure 5.1: The input image represented by watershed cells, the hybrid body representation, the online learned whole shape prior, and the part-whole segmentation results (from left to right).

5.2 Related work

Similarly as in cognitive psychology studies, existing shape representations in computer vision can be roughly grouped into two categories. One is template-like or silhouette-based methods, which are suitable for shape prior-based segmentation. The other is the part-based methods, which can capture the intra-class variability. *The main idea of our research is to integrate both view-based and structural description-based models into a hybrid body representation to support integrated pose recognition, localization, and segmentation.* Particularly, it can facilitate shape prior guided segmentation, by which bottom-up features can be extracted to drive the top-down inference in a cascade fashion. Additionally, both off-line and online learning are involved to learn general and subject-specific knowledge respectively, including the colors, shapes and spatial structure.

Existing pose recognition, localization, and segmentation methods can be broadly grouped into three major categories according the way how the body is represented: *the representation-free methods*, *the view-based methods*, and *the structural descriptions-based methods*.

The first category mainly contains some bottom-up approaches, in which there is no explicit shape prior representation, [94] and [93]. All the information used is a series of region grouping rules established according to physical constraints such as the body part proximity. In general, these approaches focus on exploiting bottom-up cues.

The second category includes all silhouette-based pose analysis methods. In [128], a specific view-based approach was proposed where pose information is implicitly embodied into a classifier learned from SIFT-like features. In general, no intermediate feature or color is used in these approaches. All view-based approaches normally aim at detecting particular body pose without extracting body parts. Thus it cannot recover anthropometric information.

The pictorial structure model proposed [122] is a typical approach belonging to the third category, in which the human body is described by several parts with their appearances and spatial relationships. This kind of approach usually requires a robust part detector. The edge histogram [116] and other SIFT-like features are widely used to represent parts. Very recently, a region-based deformable model is used to represent parts [129] where segmentation was used to verify the object hypothesis. The method in [129] is similar in spirit to the part-level inference proposed here. However, in our approach, where an image is represented by small building blocks (watershed cells), the coupled shape model is involved in a hypothesis-and-test paradigm where the region prior forms a segmentation given a position hypothesis and the edge prior evaluates the formed segmentation.

As the name suggests, the hybrid human body representation proposed here absorbs recent multifaceted advances in the field. The proposed representation involves shape prior guided segmentation and inference in a multi-stage fashion. Unlike previous methods, we use segmentation to extract bottom-up features to drive the top-down inference. Our contributions in this work include: (1) *a hybrid human body representation* that supports the online color model learning and involves an online learned deformable shape model to segment the whole body and parts, (2) *an effective hypothesis-and-test paradigm* for the part-level inference that involves the coupled region-edge shape priors, (3) *a three-stage cascade computational flow* to integrate pose recognition, localization and segmentation into a "biologically plausible" framework, and (4) *a new watershed-based Graphic-cut segmentation* where both region and edge shape priors are used for *optimal* segmentation.

5.3 Proposed Approach

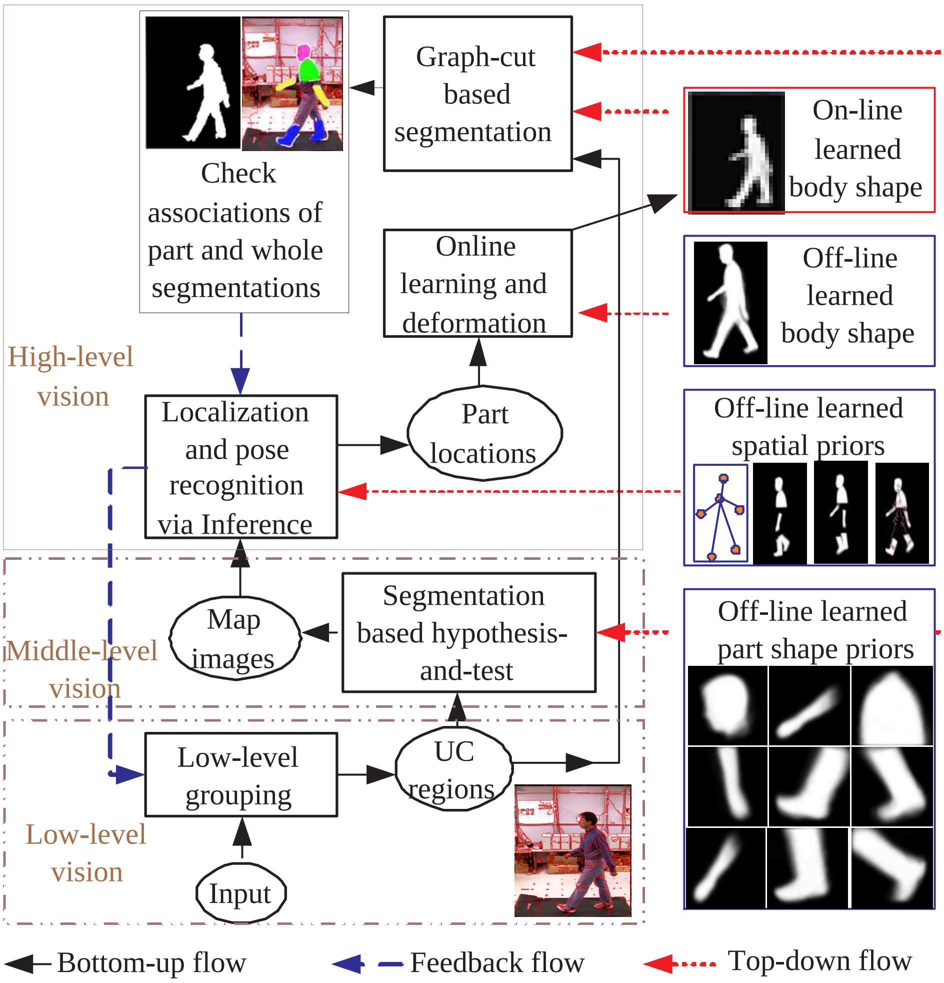

Graph-cut based segmentation
Check associations of part and whole segmentations
On-line learned body shape
Online learning and deformation
Off-line learned body shape
High-level vision
Localization and pose recognition via Inference
Part locations
Off-line learned spatial priors
Segmentation based hypothesis-and-test
Map images
Middle-level vision
Off-line learned part shape priors
Low-level grouping
UC regions
Low-level vision
Input
Bottom-up flow
Feedback flow
Top-down flow

Figure 5.2: Overview of our approach.

The proposed hybrid body representation synergistically integrates pose recognition, localization and segmentation of the whole body as well as body parts in an image, as shown in Fig. 5.2. Several key issues are addressed. **Off-line and online learning:** Off-line and online learning are used to obtain both general and subject-specific information, respectively. The former one acquires the general shape and spatial priors for both body parts and the whole body, and the latter one captures the subject specific information, including colors and shapes. **Part-whole organization:** Parts and the whole are two complementary components for object representation. The part-level inference produces the *map* images that are assembled to localize the whole body as well as body parts. The detected body parts will be further used to create a subject specific shape model for whole body segmentation. **Coupled region-edge shape model:** The coupled region-edge shape representation supports a hypothesis-and-test paradigm, where the region-based prior is used to form a segmentation and the edge-based prior is used to evaluate the formed segmentation. After the online learning of the whole body, both priors are used in a new Graph-cut segmentation framework for an *optimal* segmentation. **Balance of bottom-up and top-down:** Bottom-up and top-down flows are well balanced in a cascade fashion. From weak to strong, the top-down information is incorporated into the bottom-up processes at low-level, middle-level and high-level vision through segmentation and inference. Moreover, two feedback loops make our approach to be a dynamic computational framework.

5.4 Hybrid Human Body Representation

Consider a walking cycle with K typical poses $\mathcal{W} = \{W^{(k)}|k = 1, ..., K\}$, We model each pose $W^{(k)}$ by both part-based and whole-based statistical representations $W^{(k)} = \{V_{1:d}^{(k)}, \mathcal{L}^{(k)}, SW_{off}^{(k)}\}$, where $V_{1:d}^{(k)}$ are shape priors of d part, $\mathcal{L}^{(k)}$ is a set of statistical parameters that encode the spatial relationships between parts in a star graphical model, and $SW_{off}^{(k)}$ is the off-line learned shape prior of the whole body. The shape prior of each part $V_i^{(k)}$ is represented by the region-based shape prior $SP_i^{(k)}$, the edge-based shape prior $\mathcal{M}_i^{(k)}$, and the average orientation $\bar{\theta}_i^{(k)}$, i.e., $V_i^{(k)} = \{SP_i^{(k)}, \mathcal{M}_i^{(k)}, \bar{\theta}_i^{(k)}\}$. Moreover, during the inference processes, the part-based and whole-based color models as well as the subject specific whole shape model will be online learned as part of the hybrid body representation. For clearness, we may omit the pose index (k), in some places below.

5.4.1 Part-based Shape Prior Learning

Following the same learning method as introduced in the provirus chapter section4.5, for each body part V_i, we have obtained a set of training images (pre-segmented binary images with a fixed window size and the measured orientation). Let $\bar{\theta}_i$ be the average orientation of part V_i. All training images have been aligned to the average orientation first. Let $\{\Omega_i^j(p)|j = 1, ..., Q\}$ denote the aligned training images, the shape prior $SP_i(p)$ can be obtained by adding all aligned training images.

$$SP_i(p) = \frac{1}{Q}\sum_{j=1}^{Q}\Omega_i^j(p).$$

$SP_i(p)$ and $1 - SP_i(p)$ reflect the the probability of pixel p belonging to the object and background respectively. Given a threshold ε, an *average* object boundary $\mathcal{M}_i$ can be extracted from the learned region-based shape prior $SP_i(p)$ by a level-set like method,

$$\mathcal{M}_i = \{p|SP_i(p) = \varepsilon\}. \tag{5.1}$$

5.4.2 Part-based Spatial Prior

We use the spatial prior model proposed in [116] to characterize the variability of spatial configuration of body parts. For pose k, we define the part-based spatial prior by a start graphical model as shown in the second figure of Fig. 5.1 that is parameterized by $\mathcal{L}^{(k)} = \{\mu_i^{(k)}, \Sigma_i^{(k)}|i = 1, ..., d, i \neq r\}$. Specifically, $\{\mu_i^{(k)}, \Sigma_i^{(k)}|i \neq r\}$

denote the Gaussian priors for the *relative* locations between the non-reference part i and the reference part r. These statistical parameters can be obtained by a maximum-likelihood estimator (MLE) from labeled training data. Given a particular spatial configuration of d parts, $L = (l_1, ..., l_d)$, the joint distribution of d parts with respect to pose k can be written as the following:

$$p_{\mathcal{L}^{(k)}}(L) = p_{\mathcal{L}^{(k)}}(l_1, ..., l_d) = p_{\mathcal{L}^{(k)}}(l_r) \prod_{i \neq r} p_{\mathcal{L}^{(k)}}(l_i|l_r). \tag{5.2}$$

The same as in [116], we assume that $p_{\mathcal{L}^{(k)}}(L)$ is Gaussian. Therefore, the conditional distribution $p_{\mathcal{L}^{(k)}}(l_i|l_r)$ is still Gaussian. As defined above, $\mu_i^{(k)}$ and $\Sigma_i^{(k)}$ are the mean and covariance for the spatial distribution (relative) of part i in pose k. Then, for each non-reference part i, the conditional distribution of its position with respect to pose k is defined below,

$$p_{\mathcal{L}^{(k)}}(l_i|l_r) = \mathcal{N}(l_i - l_r|\mu_i^{(k)}, \Sigma_i^{(k)}). \tag{5.3}$$

5.4.3 Whole Body Shape Prior

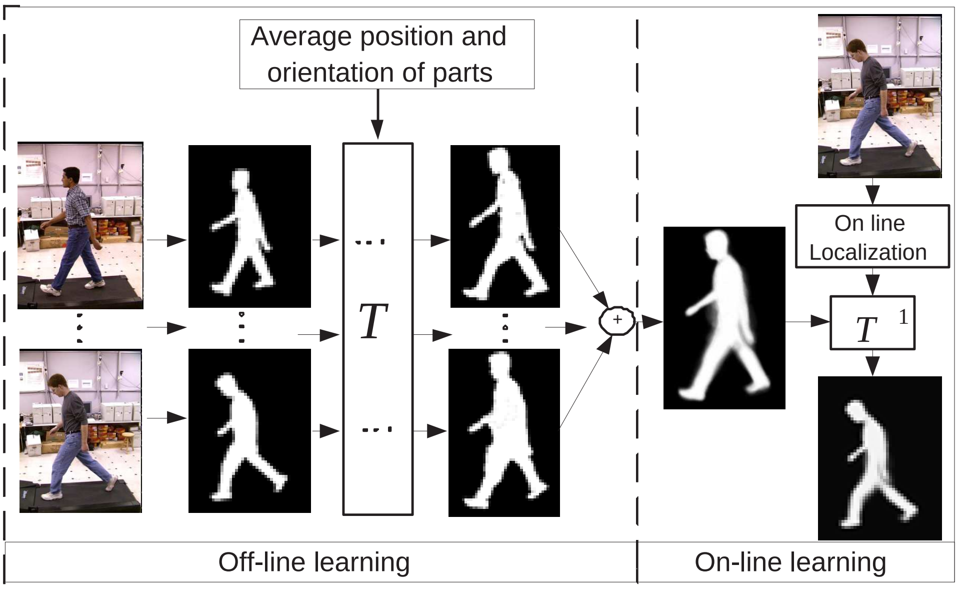


Figure 5.3: The learning of the whole body shape prior.

For each pose, a whole body shape prior is needed for body segmentation after pose recognition and localization. Both off-line and online learning are involved for generating shape models that capture the general representation as well as the subject specific information, as shown in Fig. 5.3.

- **Off-line learning** The off-line learning is similar to that of parts, except that a part-based alignment is needed due to the spatial variability of each pose. For each pose, we can compute the average location and orientation for all parts. For each training image with a segmented body and parts, we want to find a set of control points based on which the training image can be deformed in a way that all parts are transformed to the average location and orientation. To preserve the shape information of parts, we will use the edge points of all parts to be control points which can be transformed to a target image via 2-D rigid transformations obtained from the averaged locations and orientations. After we get control points in both source and target images, the multilevel B-spline method [130] is used to obtain non-linear transformations by which all pixels in the source image are mapped to the target image. Small holes can be filled by simple morphological operations. These aligned biliary images are used construct a whole body shape prior, i.e., $SW_{off}(p)$.

- **Online learning** The online learning is used to create a subject specific shape model $SW_{on}(p)$ after all parts are localized. The goal is to deform $SW_{off}(p)$ in a way that the detected parts are reflected in the shape prior. The similar technique described above for off-line learning is used here to find the non-linear transformation functions for every pixel in $SW_{off}(p)$ by which $SW_{off}(p)$ is converted to $SW_{on}(p)$ that carries a subject specific shape model. It is worth noting that $SW_{off}(p)$, unlike the training image in the off-line learning, is not a binary image and an appropriate interpolation is needed to fill possible holes in $SW_{on}(p)$.

5.5 Low-level Vision: Watershed Transform

Grouping pixels into small homogenous regions is becoming a popular pre-processing for many computer vision tasks. This is well supported by the cognitive theory proposed by [2] that considers uniform connectedness (UC) regions as the building block for object representation. In this work, we chose the watershed transform [107] because of its many "biologically plausible" properties, such as fast, local computation. More importantly, both boundary and regional information are available for each cell. To overcome the over-segmentation problem, the geodesic reconstruction preprocessing [110] is used to control the watershed size through some morphological parameters, which can be dynamically adjusted according to the feedback from the high-level vision (as shown in Fig. 5.2.

A given an image I, is represented by Z watershed cells $I = \{\mathcal{C}_i | i = 1, 2, ..., Z\}$. Each cell consists of a set of pixels $\mathcal{C}_i = \{p_1^{(i)}, p_2^{(i)}, ..., p_{\eta_i}^{(i)}\}$, where η_i is the number of pixels. Moreover, we use a 3-D Gaussian model $\{\mu_i^{(c)},\Sigma_i^{(c)}\}$ to represent the color distribution in the $L*a*b$ color space for cell $\mathcal{C}_i$. The watershed cells are used as the building blocks in the following processes.

5.6 Mid-level Vision: Part-based Inference

The goal of the mid-level vision is to generate immediate part detection results that will be useful for the high-level vision. What we need here is a *map* image that indicates how likely there is an object (i.e., a body part) at each location. How to obtain this kind of a *map* images has already been introduced in the previous chapter. We will not repeat it here. We use $g_i^{(k)}(I, l_i)$ to represent the *map* image for part i of pose k in image I, and l_i denotes an arbitrary position in the given image I.

5.7 High-level Vision: Recognition/Localization

With the obtained *map* images $g_i^{(k)}(I, l_i)$ from mid-level vision processes, let $I_{maps}^{(k)} = \{g_i^{(k)}(I, l_i), ..., g_d^{(k)}(I, l_d)\}$ denotes the set of d *map* images, part localization and pose recognition are formulated as an inference process guided by the spatial priors of different poses represented by $\{\mathcal{L}^{(k)} | k = 1, ..., K\}$. Using Bayes law, the posterior distribution for pose k can be written in terms of the map images $I_{maps}^{(k)}$ and the spatial prior defined in (6.6) as ,

$$p_{\mathcal{L}^{(k)}}(L|I_{maps}^{(k)}) \propto p_{\mathcal{L}^{(k)}}(I_{maps}^{(k)}|L).p_{\mathcal{L}^{(k)}}(L). \tag{5.4}$$

Let $P_{\mathcal{L}^{(k)}}(I_{maps}^{(k)}|L) = \prod_{i=1}^{i=d} g_i^{(k)}(I, l_i)$, by manipulating the terms in (6.17), we have

$$p_{\mathcal{L}^{(k)}}(L|I_{maps}^{(k)}) \propto p_{\mathcal{L}^{(k)}}(l_r) g_r^{(k)}(I, l_r) \prod_{i \neq r} p_{\mathcal{L}_k}(l_i|l_r) g_i^{(k)}(I, l_i). \tag{5.5}$$

Then pose recognition and part localization can be jointly obtained by the following optimization:

$$\{k^*, L^*\} = \arg\max_{k,L} p_{\mathcal{L}^{(k)}}(L|I_{maps}^{(k)}). \tag{5.6}$$

However, the direct evaluation of (6.16) is computationally prohibitive. We use the efficient inference engine proposed in [116] to obtain the solution here. For any non-reference part i of pose k, the quality of an optimal location can be,

$$\epsilon_{k,i}^*(l_r) = \max_{l_i} p_{\mathcal{L}^{(k)}}(l_i|l_r) g_i^{(k)}(I, l_i). \tag{5.7}$$

Given $p_{\mathcal{L}^{(k)}}(l_i|l_r)$ is Gaussian, $\epsilon_{i,k}^*(l_r)$ can be computed by the generalized distance transform. Then, the posterior probability of an optimal configuration for pose k can be expressed in terms of the reference location l_r and ϵ_i^*. Then the posterior probability in (5.5) will become,

$$p_{\mathcal{L}_k}(L|I_{maps}^{(k)}) \propto p_{\mathcal{L}_k}(l_r) g_r^{(k)}(I, l_r) \prod_{i \neq r} \epsilon_{k,i}^*(l_r), \tag{5.8}$$

which will lead to a new *map* image $G_k(I, l_r)$ that indicates how likely the reference part of pose k is in each location. This new *map* image $G_k(I, l_r)$ is the pooling results of all the *map* images in $I_{maps}^{(k)}$ via the spatial prior model of pose k, i.e., $\mathcal{L}^{(k)}$. Therefore, pose recognition and reference part localization can be efficiently implemented by

$$\{k^*, l_r^*\} = \arg\max_{k,l_r} G_{k=1:K}(I, l_r). \tag{5.9}$$

After the reference part is located, the position of each non-reference part can be obtained by

$$l_i^* = \arg\max_{l_i} p(l_i|l_r^*) g_i^{(k^*)}(I, l_i). \tag{5.10}$$

According to the maximum value of obtained new *map* images $G_{k^*}(I, l_r^*)$, we design a feedback loop to adjust the size of watershed cells in low-level vision, as shown in Fig. 5.2.

5.8 Whole Body Segmentation via Graph-cut

After pose recognition and localization, online learned whole body shape priors, $SW_{on}^{L^*}(p)$, $\mathcal{M}_w^{L^*}$ and pose configuration L^* can be obtained. Given image $I = \{\mathcal{C}_i | i = 1, ..., Z\}$, $\tau = \{\tau_i | i = 1, ..., Z\}$ denotes the set of binary class labels for all watershed cells ($\tau_i = 0$: background and $\tau_i = 1$: object). Following the segmentation energy definition from [114]

$$E(\tau) = \lambda . \sum_{i=1}^{Z} R(\tau_i) + \sum_{\mathcal{C}_i \bigcap \mathcal{C}_j \neq \emptyset} E(H_{i,j}) \delta(\tau_i, \tau_j), \tag{5.11}$$

where $R(\tau_i)$ is the regional term, which relates to the posteriori probability of $\mathcal{C}_i$ belonging to class τ_i; $E(H_{i,j})$ is the boundary term, which represents the consistence between the edge-based shape prior $\mathcal{M}_w^{L^*}$ and local boundary formed by two cells, $H_{i,j} = \mathcal{C}_i \bigcap \mathcal{C}_j$; $\delta(\tau_i, \tau_j) = 1$ when $\tau_i \neq \tau_j$ otherwise $\delta(\tau_i, \tau_j) = 0$; λ specifies a relative importance between two terms.

The calculation of $R(\tau_i)$ involves online learning of figure/ground color models where region-based shape prior $SW_{on}^{L^*}(p)$ is involved for kernel-based density estimation, as discussed in Section 4.4.3. Let $\hat{f}_{ob}^{(w)}(x)$ and $\hat{f}_{bg}^{(w)}(x)$ be the figure/ground color models, and $\alpha_i^{(w)}$ and $\beta_i^{(w)}$ are computed from $SW_{on}(p)$ that denote the prior probabilities of $\mathcal{C}_i$ belonging to the object and background respectively. Therefore, $R(\tau_i)$ is defined as

$$\begin{aligned} R(\tau_i = 1) &= -\ln \alpha_i^{(w)} \hat{f}_{ob}^{(w)}(\mu_i^{(c)}), & (5.12) \\ R(\tau_i = 0) &= -\ln \beta_i^{(w)} \hat{f}_{bg}^{(w)}(\mu_i^{(c)}), & (5.13) \end{aligned}$$

where $\mu_i^{(c)}$ is the mean color vector of $\mathcal{C}_i$. Using the same idea of edge-based shape evaluation defined in (4.14), let $X = H_{i,j}$, and we can define $E(H_{i,j}) = \rho_{\mathcal{M}}(X)$, which evaluates the consistence between $H_{i,j}$ and edge-based shape prior $\mathcal{M}_w^{L^*}$ in terms of the degree of parallelness and the shape similarity.

In a similar way, all body parts can also be segmented. Moreover, the segmentation in the high-level vision stage will help us extract more useful features to prune possible false positives. For example, false positives can be identified by checking the color similarity between the two arms or legs. The feedback loop from segmentation to localization (as shown in in Fig. 5.2) makes our framework a dynamic system that has potential to be further optimized.

5.9 Experimental Results

Here we validate the effectiveness of the proposed approach on the CMU Mobo database [115], which contains 25 individuals walking on a treadmill. The image is reduced to the size of 240 × 320, and each body part is defined in a window of 61 × 61. For each pose, there are totally six parts (the *head*, *torso*, *left-arm*, *right-arm*, *left-leg*, and *right-leg*), and 200 manually segmented biliary images are used for the off-line learning of part-based and whole body shape priors. The number of training images can be greatly reduced if we adopt a distance transform based shape prior learning method [131]. The algorithm was programmed in C++, and the test platform is Pentium 4 3.2GHz and 1GB RAM. The evaluation is conducted in three aspects, i.e., pose recognition, localization and segmentation. We will compare the proposed method with the "1-fan" method in [116] in terms of their performance of pose recognition and localization. Although our approach is a dynamic process with potential for further optimization, we have fixed the watershed transform in all experiments (without the feedback from high-level vision to fine tune the watershed transform as shown in Fig.5.2).

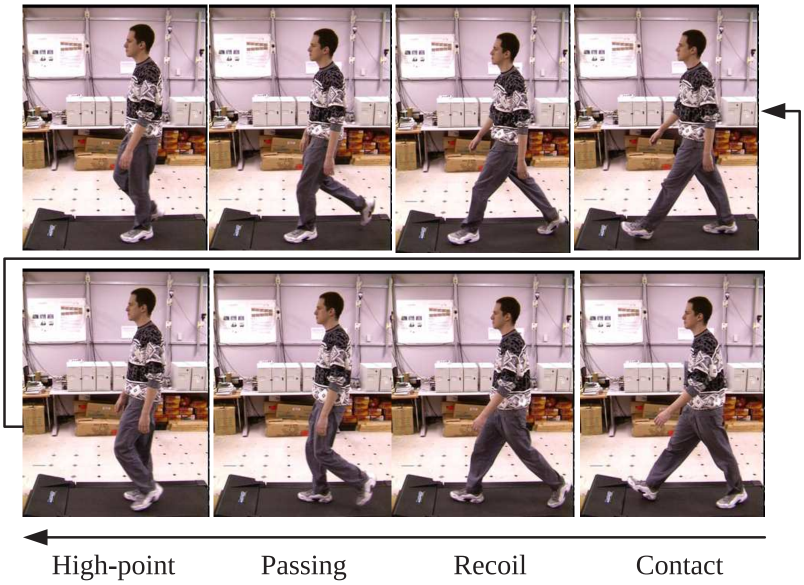


Figure 5.4: Pose definitions [4].

5.9.1 Pose Recognition

In a walking cycle, the human pose is a continuous time-varying variable. Following the pose definition of [4], we describe a walking cycle by four distinct pose couples, *Contact*, *Recoil*, *Passing* and *High-point*, as shown in Fig. 5.4. In our experiments, we combine poses *Recoil* and *Contact* together due to their strong similarity. It is possible to obtain finer pose classes after taking advantages of segmentation results. For each of the three poses, the *torso* is used as the reference part, and 230 labeled training data are collected for learning the part-based spatial prior. It was found that the proposed approach achieves the recognition rate of 98% for the three poses over 144 test images from 21 persons. The mis-classification only occurs for pose *passing* that sometimes is very similar to other two poses. The only way to improve the recognition for this pose is to incorporate the motion information from video sequences. The "1-fan" method in [116] achieved the recognition rate of 93%.

5.9.2 Localization

Based on the same test images used for pose recognition, we have tested two methods on the localization of three poses, i.e., *High-point* (H-point), *Contact* and *Passing.* The comparative results are shown in Tables 6.1 where we have two findings. The two methods are comparable in localizing the *Head* and *Torso*; and the proposed approach shows significant improvements in localizing the *legs* and *arms.*

Table 5.1: The localization comparison for three poses.

Poses	Methods	Head	Torso	Larm	Rarm	Lleg	Rleg
H-point	1-fan	6.7	7.9	10.7	13	16.4	15.4
	hybrid	6.2	6.2	6.4	9.2	6.2	9.3
Contact	1-fan	3.4	6.4	13.9	10.4	8.7	10
	hybrid	5.4	5.6	11.8	9.6	3.8	4.4
Passing	1-fan	5.6	6.2	11.9	9.8	13.1	12.9
	hybrid	6.2	6.3	11.2	7.1	4.5	3.2

The reason for the first findings is because that the relative position between the *head* and *torso* has least variability, and the part-based spatial prior that is shared by the two methods plays the major role for part localization, leading to the similar results. However, there is drastic (relative) spatial variability, both positional and orientational, for the *arms* and *legs*, and the improvements from the proposed method are significant due to the enhanced saliency of the part-based *map* images generated by the segmentation-based hypothesis-and-test paradigm. Overall, the localization accuracy of the *torso*, i.e, the whole body localization, is improved due to the enhancement of each individual part-based *map* image. Some part localization results of three poses are shown in Fig. 5.5 (the first three rows), where the proposed method successfully detects (and segments) all body parts despite the significant variability.

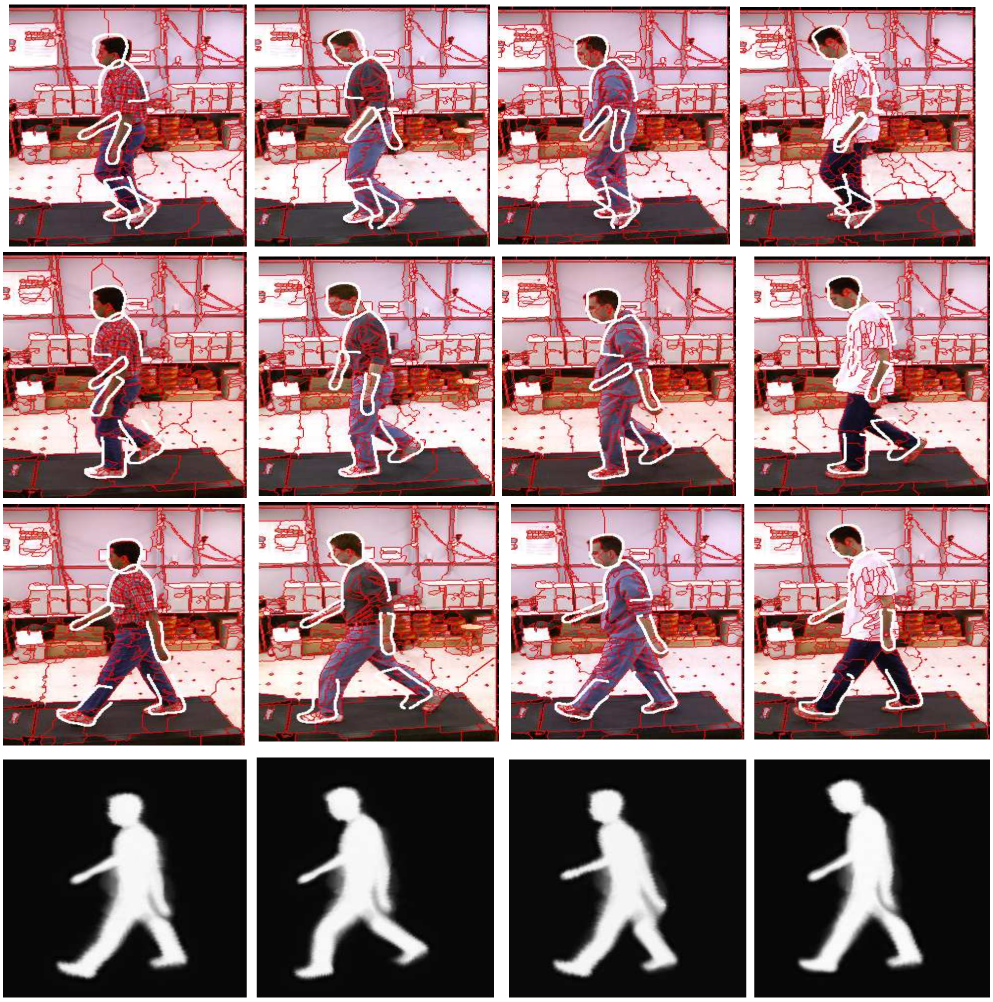


Figure 5.5: Part localization of three poses and online learned whole body shape models that are used for the whole body segmentation.

5.9.3 Segmentation

After localization, an online learned subject specific shape prior (as shown in the last row of Fig. 5.5) is used in the new Graph-cut algorithm where both region and edge priors are involved in the energy function defined in (5.11). For comparison, we also did one experiment where the second term (5.11) is replaced by a standard color similarity-based boundary penalty term [114] without using the edge-based shape prior. For pose *Contact*, we performed Graph-cut segmentation with ($\lambda = 1/30$ in (5.11)) on 60 test images for which we also obtained the ground-truth segmentation masks for objective and quantitative evaluation. Segmentation results are evaluated by the ratio between the falsely detected region size (including both false positives and false negatives) and the ground truth region size. The error rate of the segmentation using both region-based and edge-based priors is 17.2%, while that of the one without using the edge-based shape prior is 38.1%. Therefore, we have obtained more than 50% improvement. These segmentation results may be further improved if more dedicated 2-D deformation methods are involved.

More localization, on-line shape learning, and segmentation results are shown in the Fig 5.6 and Fig 5.7

5.10 Conclusion

In this chapter, we have proposed a hybrid body representation that supports an integrated pose recognition, localization and segmentation framework. Particularly, segmentation, as a bridge between bottom-up cues and top-down information, plays an important role in all three levels of vision. At low-level vision, the watershed-based image representation facilitates the subsequent learning and inference. At mid-level vision, the segmentation-based hypothesis-and-test paradigm enhances the saliency of *map* images obtained from the part-level inference and leads to accurate localization of both parts and the whole body. At high-level vision, a subject specific shape

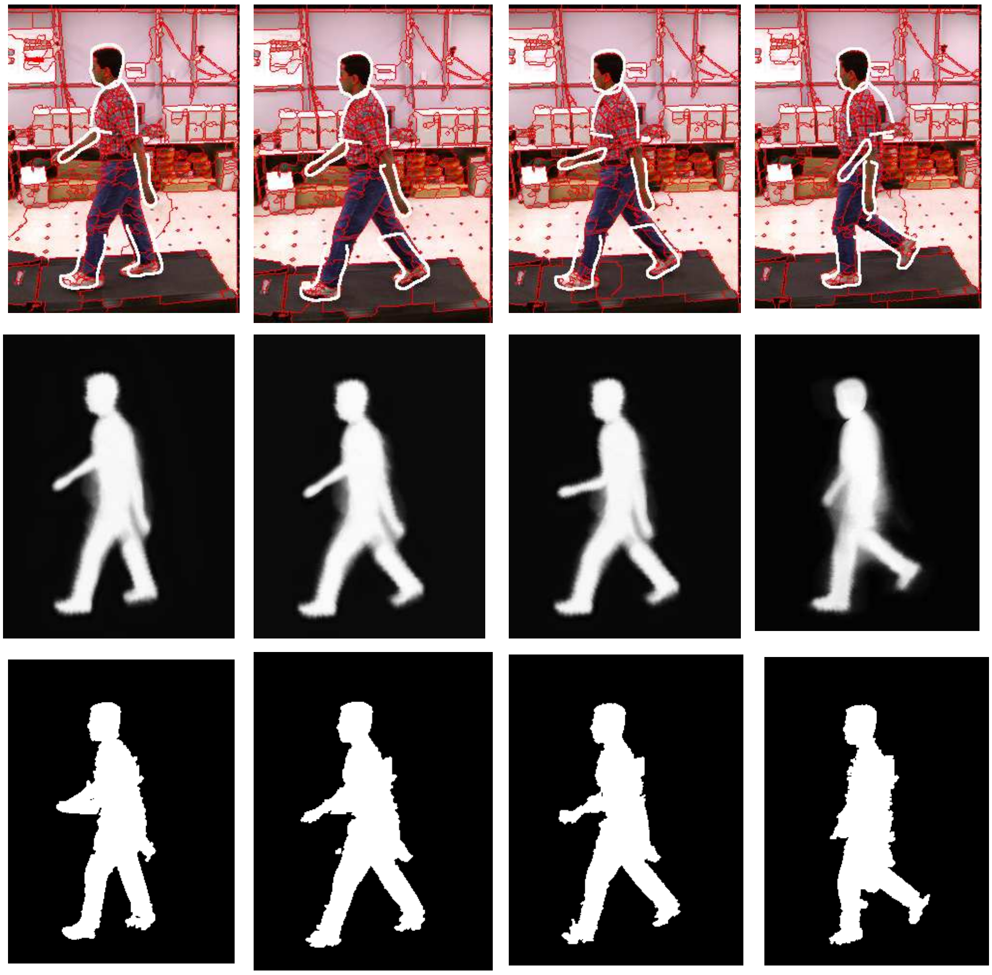


Figure 5.6: More localization, on-line shape learning, and segmentation results

prior is learned online based on part localization results and used in the new Graph-cut algorithm where both region and edge priors are jointly utilized to optimize the segmentation. The proposed framework is essentially a dynamic system with feedback loops that has potential to be further optimized. In the next chapter, research will focus on extending the proposed body representation to be a dynamic human body representation that supports video-based pose recognition, localization, tracking and segmentation.

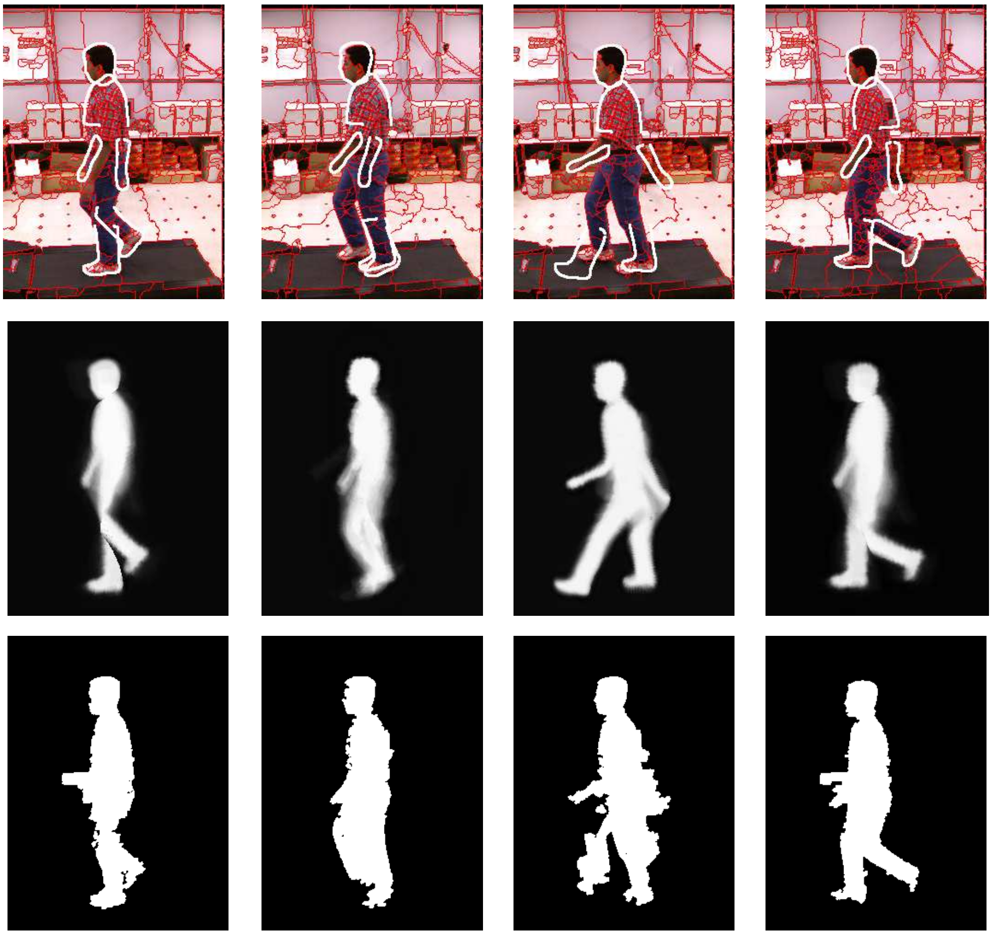


Figure 5.7: More localization, on-line shape learning, and segmentation results

CHAPTER 6

High-level vision: Tracking and Localization

6.1 Overview

In the previous chapter, we proposed a promising human recognition, localization and segmentation algorithm for images. But it uses only appearance and spatial constrains, which are represented in a pictorial structure model. We know that temporal consistency is an essential property of human body appearing in a video sequence, which makes body part detection and tracking in a video sequence different from repeated application of an image-based detection algorithm. Therefore, how to combine both spatial and temporal constrains into human localization and segmentation in a video sequence is the objective of this chapter. In other words, the purpose of this chapter is to exploit the complementary context information in both temporal priors and spatial priors for human tracking.

Human tracking and body part localization are among the most challenging research issues largely due to the ambiguity, complexity and non-linearity in observed video sequences as well as the ill-posed nature of the problem. Using appropriate prior knowledge (such as motion pattern or shape) would make the problem better defined and hopefully easier to tackle. The major advantages of using priors are to reduce the search space by taking advantages of various constraints and to ensure a plausible solution that is consistent with prior knowledge. Two commonly used priors are spatial and temporal priors, both of which play very important roles for human detection and tracking, and have been well studied by many computer vision researchers in different context.

Spatial priors are usually defined on body parts and characterize the spatial configuration of a certain pose [116]. One important question is how to make one spatial prior adaptable to a large number of pose variations. One straightforward extension is to train separate spatial priors for several typical poses [132]. Generally speaking, a spatial prior representation that can only handle a discrete pose variable has difficult to characterize the smooth and continuous pose transition in a video sequence. On the other hand, temporal priors specify certain dynamic constraint of human motion [133], and they can ensure the temporal continuation across adjacent poses. Most temporal models do not impose a strong spatial constraint among body parts or treat each part independently for tracking [134]. How to learn the two priors are also of great interest. There are two kinds of learning strategies. *Off-line learning* normally requires sufficient and/or diverse training samples and usually leads to the learned priors that favor the training data. *Online learning* can learn the priors "on-the-fly" on the testing data. Recent studies show that online learning is more favorable and effective to deal with human motion with significant variability even from different activities [135].

In this work, we propose a new framework for articulated human tracking that integrates both spatial and temporal priors and is supported by online learning. The idea of combining both priors has been well acknowledged and incorporated into most tracking algorithms where both priors are usually learned off-line and one prior often overshadows the other one during inference. In our work, the spatial prior is embedded in the temporal prior, and both priors are learned online from past tracking history in an incremental way. Specifically, the temporal prior can predict the pose for the next frame that induces a pose specific spatial prior. This spatial prior in return is used to evaluate and correct the pose prediction by assembling part-level detection. Our approach distinguishes itself from others in that it incorporates both online learned spatial and temporal priors in one integrated inference framework. The proposed algorithm is able to track subjects with significant shape/color variability, and can also deal with abnormal motion patterns.

6.2 Related Works

The biological vision model proposed in [14, 15] suggested two perception pathways in motion perception, the *appearance pathway* and the *motion pathway.* It was considered as one of the major breakthroughs in recent vision research [55], and motivates researchers to involve both spatial (appearance) and temporal (motion) priors in their tracking algorithms. Broadly speaking, related work can be classified into two groups: the *temporal-prior dominated approaches* and the *spatial-prior dominated approaches.*

In [132], a unified spatial-temporal articulated model was proposed for human tracking, where the pose is a discrete variable and defined as the hidden state of a hidden Markov model (HMM). The temporal prior is incorporated as a state transition matrix, and then the tracking task is formulated as a Bayesian estimation problem. In [136], a single pictorial structure graph model was extended into a dynamic Bayesian network (DBN), where the probabilistic relationships between joints at a given time instant as well as those over time can be learned from motion capture data. Then belief propagation is used as the inference engine to effectively incorporate the top-down spatial prior with bottom-up part detection for articulated human tracking. Along the same venue, a temporal pictorial structure model was developed in [137], which mainly relies on appearance priors for human tracking. Above methods are considered as the *spatial-prior dominated* ones where the spatial prior plays a more important role and only weak temporal priors are involved for dealing with activity variation.

The human pose can be represented in a high dimensional (HD) parameter space where the distribution of plausible human poses is very sparse. Various non-linear dimensionality reduction (DR) techniques were proposed to explore the low-dimensional

(LD) intrinsic structures for a compact pose representation. The Gaussian Process Latent Variable Model (GPLVM) [138] is an effective DR technique that offers a smooth mapping from the LD latent space to the HD kinematic space. Several GPLVM variants were developed for temporal series analysis. For examples, Gaussian Processing Dynamic Models (GPDM) [133] were specifically designed for human motion tracking by introducing a dynamic model on the latent variable that can be used to produce tracking hypothesis in a latent space[134]. Back Constrained-GPLVM (BC-GPLVM) [139] improves the continuity in the latent space by enforcing the local proximities in both the LD and HD spaces. Consequentially, BC-GPLVM produces a smooth motion trajectory in the latent space that can be used as a non-parametric dynamic model for human tracking [140]. All of above DR methods focus on the exploration and exploitation of temporal priors of human motion, and they do not involve spatial (kinematic) priors explicitly. Therefore, we consider them as temporal-prior dominated approaches.

Motivated by previous research, we want to take advantage of the complementary nature of the above two methodologies. On the one hand, our work is similar to [132, 137] in the sense of how the spatial prior is represented. But we involve a *strong* temporal prior that can handle a continuous pose variable. On the other hand, our algorithm inherits some ideas from [140, 141] regarding how the temporal prior is developed for top-down prediction. However, we use a *structured* spatial prior that fuses part detection results to evaluate and correct the prediction. Moreover, we explore the synergy between the two priors in the context of online learning, which is inspired by the local mixed Gaussian process regressors proposed in [135]. To the best of our knowledge, there is no prior research on how to combine spatial and temporal priors in an online learning framework.

6.3 Background Knowledge

We firstly briefly review the two major building blocks regarding the representations of temporal and spatial priors.

6.3.1 Temporal Prior Modeling: GPLVM

The Gaussian process latent variable model (GPLVM) [138] is an effective method to learn $X = \{\mathbf{x}_i\}_{i=1}^N$ in a LD latent space from $Y = \{\mathbf{y}_i\}_{i=1}^N, \mathbf{y}_i \in R^D$ $(D >> q)$ in a HD observation space, and it also provides a probabilistic mapping from X to Y. We refer the readers to [138, 142] for more details. Assuming each observed data point, $\mathbf{y}_i$ is generated through a noisy process from a latent variable $\mathbf{x}_i$,

$$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon, \tag{6.1}$$

where $\epsilon \sim N(0, \beta^{-1}I)$. Assuming a Gaussian distribution over functions $f \sim N(0, k(\mathbf{x}_i, \mathbf{x}_j))$. The covariance $k(\mathbf{x}_i, \mathbf{x}_j)$ characterizes the nature of the functions. One widely used covariance function is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 e^{-\frac{\theta_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2} + \theta_3 + \beta^{-1}\delta_{i,j}, \tag{6.2}$$

where the parameters are given by $\Phi = \{\theta_1, \theta_2, \theta_3, \beta\}$ and $\delta_{i,j}$ is the Kronecker's delta function. The scalar $k(\mathbf{x}_i, \mathbf{x}_j)$ models the proximity between two points $\mathbf{x}_i$ and $\mathbf{x}_j$.

After GPLVM learning, given a new latent variable $\mathbf{x}_*$, the likelihood of the corresponding HD data point $\mathbf{y}_*$ is:

$$p(\mathbf{y}_*|X, \mathbf{x}_*) = N(\mathbf{y}_*|\mu, \sigma^2), \tag{6.3}$$

where

$$\mu = Y^T K_{X,X}^{-1} \mathbf{k}_{X,\mathbf{x}_*}, \tag{6.4}$$

where $K_{X,X} = \{K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)\}$, and $\mathbf{k}_{X,\mathbf{x}_*}$ is a column vector developed from computing the elements of the kernel matrix between the learn latent state data X

and the new point $\mathbf{x}_*$. The variance that is then given below will increase as $\mathbf{x}_*$ deviates from the training data X.

$$\sigma^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{X,\mathbf{x}_*}^T K_{X,X}^{-1} \mathbf{k}_{X,\mathbf{x}_*}. \tag{6.5}$$

To ensure a smooth trajectory in the latent state space for temporal series data, BC-GPLVM was proposed in [139] that enforces local proximities in both the LD and HD spaces. In our work, BC-GPLVM is used to learn a compact LD representation of human motion in the latent space and a probabilistic reverse mapping from the LD latent space to the HD observation space. We adopt the BC-GPLVM to a local online learning strategy [135].

6.3.2 Spatial prior: Star-structured Graphic Model

The pictorial structure based spatial prior representation has become an increasingly compelling approach for articulated human body tracking. Following [116], we represent the spatial prior for a pose by a star-structured graphical model Ψ. Let us regard pose $\mathbf{y} = (l_1, ..., l_d)$ as a vector of 2D configuration (position and orientation) of d body parts. The joint distribution of d part configuration with respect to pose $\mathbf{y}$ can be written as the following:

$$p_\Psi(\mathbf{y}) = p_\Psi(l_1, ..., l_d) = p_\Psi(l_r) \prod_{k \neq r} p_\Psi(l_k | l_r), \tag{6.6}$$

where l_k and l_r are the configuration parameters for non-reference part k and the reference part r respectively. By assuming the conditional probability density functions for $p_\Psi(l_k|l_r)$ following the Gaussian distribution. Then, for each non-reference part k, the conditional distribution of its configuration with respect to pose Ψ is defined below,

$$p_\Psi(l_k|l_r) = \mathcal{N}(l_k - l_r | \mu_k, \Sigma_k). \tag{6.7}$$

We can also assume a Gaussian distribution for $p(l_r)$.

$$p_\Psi(l_r) = \mathcal{N}(l_r | \mu_r, \Sigma_r). \tag{6.8}$$

In an off-line learning framework, the parameters of the start model are often obtained by a maximum-likelihood estimator (MLE) from the labeled training data. For a test image, this spatial prior is used to assembly part detection results, or called *map* images, which indicate the confidence of the existence of each part at every pixel location. Edge histogram-based part detection was used in [116] where a distance transform-based fast inference algorithm is also developed to assemble *map* images for detection and localization. In [5], a segmentation-based hypothesis-and-test method was proposed to produce more salient *map* images for part detection that improves the whole-part localization accuracy. We will make two extensions to the start model-based spatial prior in this work. (1) The spatial prior is *time variant* and is able to handle a continuous pose variable rather than a discrete pose variable in [132, 5]. (2) The spatial prior is embedded in the temporal prior, and can be constructed "on-the-fly" based on the temporal prediction for every incoming frame rather than learned offline [116].

6.4 Proposed Research

6.4.1 Research Overview

Our algorithm is featured by the marriage of two powerful mathematical tools, BC-GPLVM and the star-structured graphical model, which is elaborated in the context of online learning. The synergy between the two priors is explored by embedding the spatial prior into the temporal prior and learning them together. The proposed algorithm involves four major steps as follows.

- *Online learning and Pose Prediction:* From past tracking history, we learn a smooth motion trajectory in the latent space via BC-GPLVM, as shown in Fig. 6.1(a), which can be used as a non-parametric dynamic model to predict the next pose in the latent space by B-spine extrapolation.

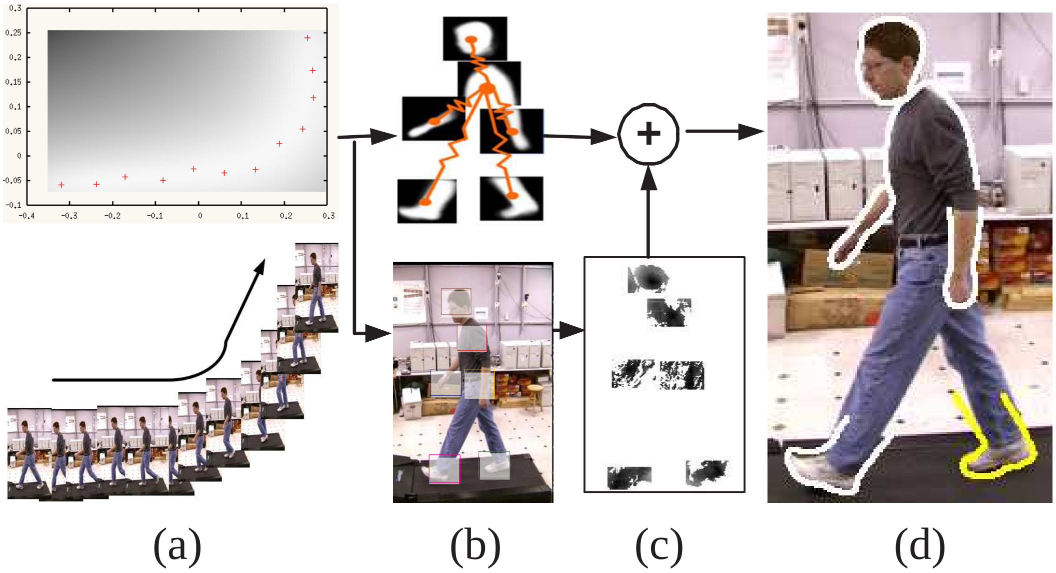


Figure 6.1: The algorithm flow. (a) Online learning and dynamic prediction in the latent space. (b) Pose prediction in the observation space and construction of the start model. (c) Local part detection according to the prediction. (d) Localization by assembling the part detection results via the start model.

- *Spatial Prior Construction:* Based on the prediction in the latent space, we can predict the next pose in the HD observation space via the LD-HD reverse mapping, as shown in Fig. 6.1(b). The predicted pose specifies the possible location of each body part in the next frame that enables the efficient local search of body parts. Also, a star model is constructed accordingly to represent the pose specific spatial prior.

- *Local Part Detection:* Based on the pose prediction, local part detection is performed for d (the number of body parts) body parts that results in d localized *map* images that are shown together in Fig. 6.1(c).

- *Pose Correction:* The pose specific start model is used to assemble the part detection outputs and produce the final localization results for the whole body as well as body parts, as shown Fig. 6.1(d).

6.4.2 Problem Formulation

Given N image frames $I_{i=1:N}$, we want to estimate the pose $\mathbf{y}_i = (l_1^{(i)}, ..., l_d^{(i)})$ for each frame where $\mathbf{y}_i$ is a vector of 2D configuration (position and orientation) of d body parts at frame i. Let $\mathbf{x}_i$ to be the latent state associated with $\mathbf{y}_i$. Let $P_i = \{p_1^{(i)}, ..., p_d^{(i)}\}$ be the appearance models (e.g., an object template) of the d body parts. Given the pose estimation results of N previous frames, i.e., $Y_{1:N} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$, current body part models p_N and next frame I_{N+1}, part localization (tracking) results can be obtained by maximize the posterior probability:

$$\mathbf{y}^*_{N+1} = \arg\max_{\mathbf{y}_{N+1}} P(\mathbf{y}_{N+1}|I_{N+1}, P_N, \mathbf{y}_{1:N}). \tag{6.9}$$

Generally, it is intractable to find the $\mathbf{y}^*_{N+1}$ directly due to its HD nature. Hence we use a *prediction-and-correction* framework to attack this problem, as shown in Fig. 6.2.

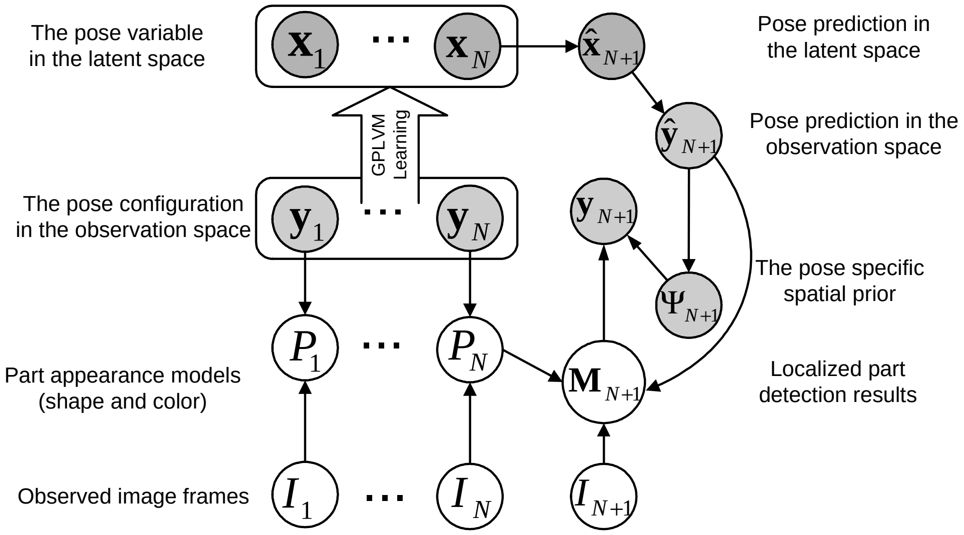

The pose variable in the latent space
$\mathbf{x}_1$
$\mathbf{x}_N$
$\hat{\mathbf{x}}_{N+1}$
Pose prediction in the latent space
GPLVM Learning
$\hat{\mathbf{y}}_{N+1}$
Pose prediction in the observation space
The pose configuration in the observation space
$\mathbf{y}_1$
$\mathbf{y}_N$
$\mathbf{y}_{N+1}$
The pose specific spatial prior
Ψ_{N+1}
P_1
P_N
Part appearance models (shape and color)
$\mathbf{M}_{N+1}$
Localized part detection results
Observed image frames
I_1
I_N
I_{N+1}

Figure 6.2: Problem formulation by a graphical model.

Assume we can learn a smooth motion trajectory $X_{1:N} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ in the latent space via BC-GPLVM based on past tracking history $Y_{1:N}$. We can predict the next pose in the latent space first, $\hat{\mathbf{x}}_{N+1}$, which can be converted to $\hat{\mathbf{y}}_{N+1}$ in the image space. $\hat{\mathbf{y}}_{N+1}$ has two implications. First, it can be used to produce d localized *map* images, $\mathbf{M}_{maps}^{(N+1)} = \{M_1^{(N+1)}, ..., M_d^{(N+1)}\}$. Second, it defines a pose specific spatial prior represented by a start model Ψ_{N+1}. Following the same idea as in [116], these *map* images can be assembled by Ψ_{N+1} in the form of star-structured graphical model. Then the tracking problem can be reformulated as maximizing the posterior probability:

$$\mathbf{y}_{N+1}^{*} = \arg\max_{\mathbf{y}_{N+1}} P(\mathbf{y}_{N+1}|\mathbf{M}_{maps}^{(N+1)}, \Psi_{N+1}), \tag{6.10}$$

where Ψ_{N+1} is the spatial prior represented in (6.6). The optimization problem of (6.10) can be efficiently solved using the fast inference algorithm developed in [116]. Then $\mathbf{y}_{N+1}$ can be used to achieve the updated appearance models P_{N+1} based on I_{N+1}, and will be involved in the next step BC-GPLVM learning to predict $\mathbf{y}_{N+2}$ as the slide window moves forward one frame.

6.5 Learning and Inference

In this section, we detail the four major steps for learning and inference in our tracking algorithm.

6.5.1 Online Learning and Pose Prediction

In general, a pose $\mathbf{y}_i$ can be represented by a HD vector that records joint angles or positions. Here we use a simple body representation with six body parts where each part is specified by the 2D position and orientation in the image domain. Given a pose series $Y_{1:N}$, BC-GPLVM can be used to learn the kernel parameters $\Phi = \{\theta_1, \theta_2, \theta_3, \beta\}$ and the latent variable series $X_{1:N}$. Different from off-line learning, we use a slide window to involve recently estimated poses for online (local) BC-GPLVM learning. The learned model is only used once for pose prediction for the next frame.

As shown in Fig. 6.3, although there is no explicit dynamic model in the latent space after BC-GPLVM learning, the temporal constraint is well-reflected by the smooth motion trajectory in the LD latent space. We can extrapolate this motion trajectory to predict the pose for the next frame.

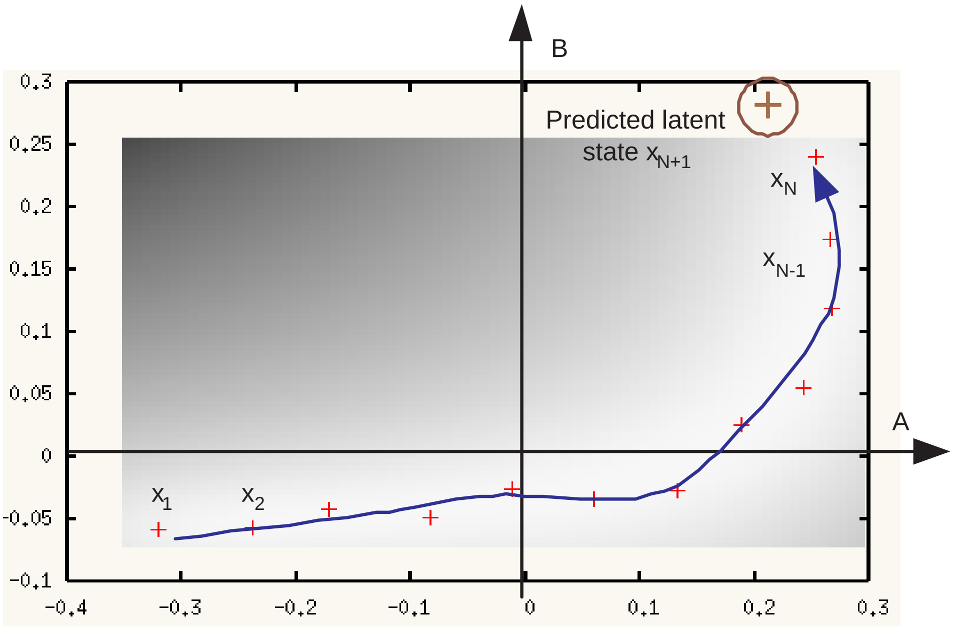


Figure 6.3: An example of BC-GPLVM online learning and dynamic prediction via B-spline extrapolation in the 2D latent space.

Let $\mathbf{x}_i = (a_i, b_i)^T$, we can apply the B-spline regression process on the latent states $X = \{\mathbf{x}_i\}_{i=1}^N$, as shown in Fig. 6.3. The two obtained B-spline functions $\mathcal{A}(\cdot)$ and $\mathcal{B}(\cdot)$ will satisfy the following constraints:

$$\begin{cases} \mathcal{A}(i) \cong a_i, \\ \mathcal{B}(i) \cong b_i, . \end{cases} \tag{6.11}$$

where $i = 1, ..., N$. Then through B-spline extrapolation $a_{N+1} = \mathcal{A}(N+1)$ and $b_{N+1} = \mathcal{B}(N+1)$, we can compute the predicted latent state for the next frame $(N+1)$ as $\mathbf{x}_{N+1} = (a_{N+1}, b_{N+1})^T$, (as indicated by the circled marker in Fig. 6.3). From the predicted latent variable $\mathbf{x}_{N+1}$, the associated pose in the image space $\mathbf{y}_{N+1}$

can be constructed through the reverse LD-HD mapping given in (6.4), and defined as:

$$\hat{\mathbf{y}}_{N+1} = Y^T K_{X,X}^{-1} k_{X,\mathbf{x}_{N+1}}. \tag{6.12}$$

6.5.2 Constructing the Spatial Prior

The uncertainty of $\hat{\mathbf{y}}_{N+1}$ is reflected by the variance defined in (6.5). So far, it is assumed that the configurations of d parts are independent, indicating a weak spatial constraint if only the temporal prior is used. In order to incorporate the spatial constraint among body parts, we need to construct the pose specific spatial prior represented by the start model from $\hat{\mathbf{y}}_{N+1}$. That means we need to estimate the conditional distributions between each non-reference part and the reference part, which can be derived straightforwardly from the Gaussian assumption of $\hat{\mathbf{y}}_{N+1}$. Therefore, the conditional distribution defined in (6.7) will become:

$$p_\Psi(l_k|l_r) = \mathcal{N}(l_k - l_r|\mathbf{y}^k_{(N+1)} - \mathbf{y}^r_{(N+1)}, 2\sigma^2 \cdot \mathbf{I}), \tag{6.13}$$

where σ^2 is given (6.5), $\mathbf{y}^r_{(N+1)}$ is the configuration of the reference part, and $\mathbf{y}^k_{(N+1)}$ is the relative configuration of non-reference part k with respect to the reference part. Similar to (6.8), the distribution of the reference part will become:

$$p_\Psi(l_r) = \mathcal{N}(l_r|\mathbf{y}^r_{(N+1)}, \sigma^2 \cdot \mathbf{I})). \tag{6.14}$$

Strictly speaking, the covariance matrices of (6.13) and (6.14) need to be added with the additional error terms to accommodate the prediction error in the latent space. In this work, the value of this prediction error term is set by experiment, and it is found that the tracking performance is not very sensitive to the choice of this value.

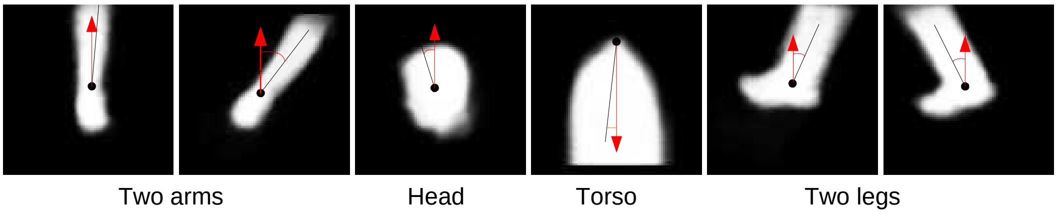


Figure 6.4: Off-line leaned part shape models where the average orientation of each part is also given.

6.5.3 Part Detection

The predicted pose $\hat{\mathbf{y}}_{N+1}$ specifies the possible locations of d body parts that could define a search region for each part. Ideally, each search region should be isotropic and determined by (6.5). For simplicity, we use a square region of 21×21 for local part detection which is centered with the part position encoded in $\hat{\mathbf{y}}_{N+1}$. Similar to [5], we resort to a segmentation-based hypothesis-and-test method for part detection where off-line learned shape models are used, as shown in Fig. 6.4. These shape models can be further represented by the coupled region-edge shape priors which are used to compute *map* images given an image represented by watershed cells. At each hypothesized location, the region prior is used to form a segmentation by merging watershed cells, while the edge prior is applied to evaluate the formed segmentation in terms of *shape similarity* and *boundary smoothness*.

To take advantage of localization results in the previous frame, we modify the evaluation criterion for computing the *map* images by replacing the *boundary smoothness* with the *template matching score*. Given a segmentation Z formed in a location, we represent the boundary of Z by $\Gamma(Z)$, and the new evaluation score for Z is given by

$$\rho_{\mathcal{M}}(Z) = -d(\Gamma(Z), \mathcal{M})) + \zeta SAD(I_{(N+1)}, P_N), \tag{6.15}$$

where the first term is the chamfer distance indicating the shape similarity between $\Gamma(Z)$ and the off-line learned edge prior $\mathcal{M}$, and the second term is the SAD (Sum of absolute differences) that reflects the degree of match between the online learned template P_N and $I_{(N+1)}$. ζ balances the relative importance between the off-line and online learned part priors. Some part detection examples are shown in Fig. 6.5(b) where a dark pixel value indicates a high possibility of the existence of a part. The computation of these *map* images can be very efficient due to local part detection constrained by $\hat{\mathbf{y}}_{N+1}$, instead of the full search used in [116, 5].

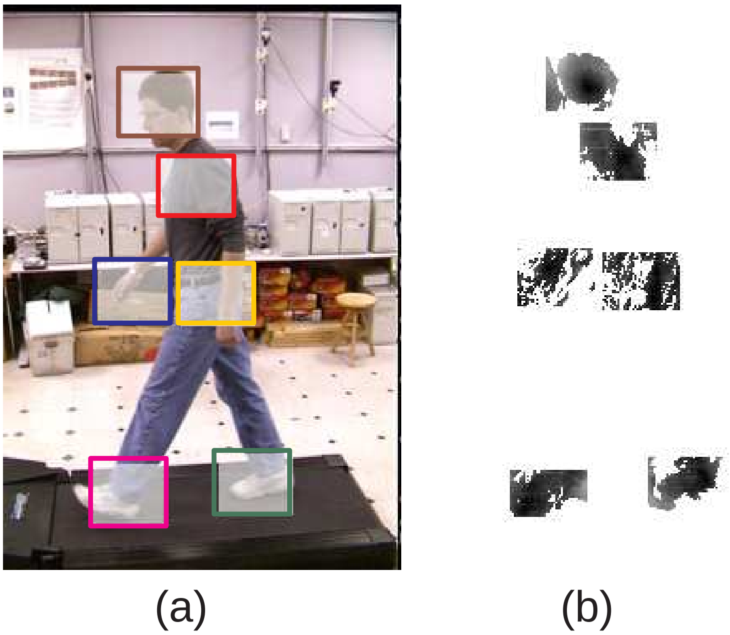


Figure 6.5: Part detection results. (a) An image with predicted local search regions. (b) The localized *map* images for six body parts.

6.5.4 Pose Correction

The part-based *map* images will be assembled by the online learned spatial prior through the star-structured graphical model. The same as in [116], given the set of *map* images $\mathbf{M}_{map}^{(N+1)} = \{M_k^{(N+1)} | k = 1, ..., d\}$, the optimal $\mathbf{y}_{(N+1)}$ can be obtained by re-writing (6.10) as,

$$\mathbf{y}^*_{(N+1)} = \arg \max_{\mathbf{y}_{(N+1)}} p_\Psi(\mathbf{y}_{(N+1)} | \mathbf{M}_{maps}^{(N+1)}). \tag{6.16}$$

Using Bayes law,

$$p_\Psi(\mathbf{y}_{(N+1)}|\mathbf{M}_{maps}^{(N+1)}) \propto p_\Psi(\mathbf{M}_{maps}^{(N+1)}|\mathbf{y}_{(N+1)}), p_\Psi(\mathbf{y}_{(N+1)}). \tag{6.17}$$

where

$$P_\Psi(\mathbf{M}_{maps}^{(N+1)}|\mathbf{y}_{(N+1)}) = \prod_{k=1}^{k=d} M_k^{(N+1)}\left(\mathbf{y}_{(N+1)}^k\right),$$

where $M_k^{(N+1)}\left(\mathbf{y}_{(N+1)}^k\right)$ is the value of the *map* image of part k in location $\mathbf{y}_{(N+1)}^k$. Also $p_\Psi(\mathbf{y}_{(N+1)})$ can be evaluated through the learned star-structured graphical model defined in (6.6). A fast distance transform based inference method can be used to solve this problem efficiently [116].

6.6 Occlusion Handling

Occlusion handling is an important issue for articulated human tracking where some body parts may be invisible for some poses. In this work, we are interested self-occlusion, and our method is inspired by the *multi-object tracking theory* proposed in [60], where the notion of "*object files*" was developed to store episodic representations for real-world objects. Each object file contains the joint spatio-temporal information (such as appearance and motion) about a particular object in a scene. An "object file" is established for each body part being tracked that plays an important role for occlusion handling. The algorithm flow is shown in Fig. 6.6 where three occlusion-related issues are addressed.

Occlusion detection: The *map* images can be directly used for occlusion detection. Given a detection threshold S, if $\min M_o(\cdot) > S$, we can declare that part o is occluded. The position estimation of part o only depends on the prior knowledge encoded in the graphical model Ψ.

Prediction of an occluded part: Given part o that is declared in I_N, its "object file" can be used for predicting its position in I_{N+1} as follows:

$$\hat{\mathbf{y}}_{(N+1)}^o = \hat{\mathbf{y}}_{(N+1)}^r + \Delta l_o, \tag{6.18}$$

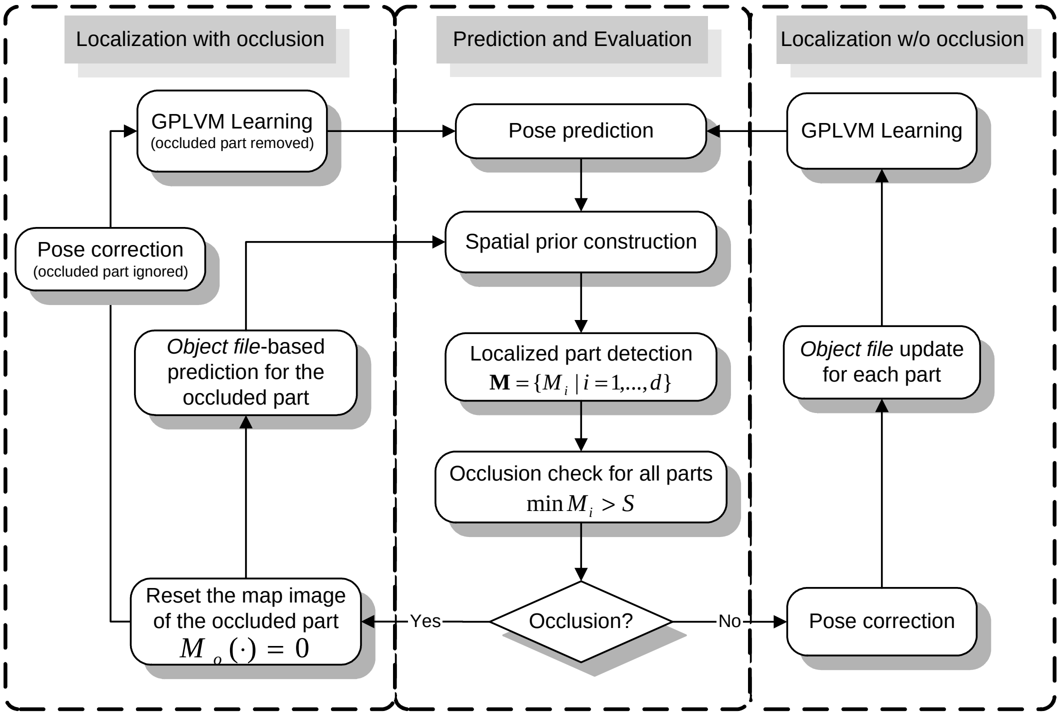

Localization with occlusion
Prediction and Evaluation
Localization w/o occlusion
GPLVM Learning
(occluded part removed)
Pose prediction
GPLVM Learning
Pose correction
(occluded part ignored)
Spatial prior construction
Object file-based prediction for the occluded part
Localized part detection
$\mathbf{M} = \{M_i \mid i = 1,\ldots,d\}$
Object file update for each part
Occlusion check for all parts
$\min M_i > S$
Reset the map image of the occluded part
$M_o(\cdot) = 0$
Yes
Occlusion?
No
Pose correction

Figure 6.6: The flow of tracking and occlusion handling.

where $\hat{\mathbf{y}}^r_{(N+1)}$ is the predicted position of the reference part (assuming the reference part is never occluded), and Δl_o is the relative configuration between part o and the reference part r that can be retrieved from the "object file" of part o.

Learning and inference for occlusion: When part o is occluded in I_N, we disable its *map* image by setting $M_o(.) = 0$. Its configuration (given by the spatial prior) will be ignored in the online BC-GPLVM learning for I_{N+1}.

The above occlusion handling technique is simple yet effective, and could be extended to handle more sophisticated cases. The synergistic use of spatial and temporal priors allows the tracking algorithm to have more flexibility and capability of handling occlusion.

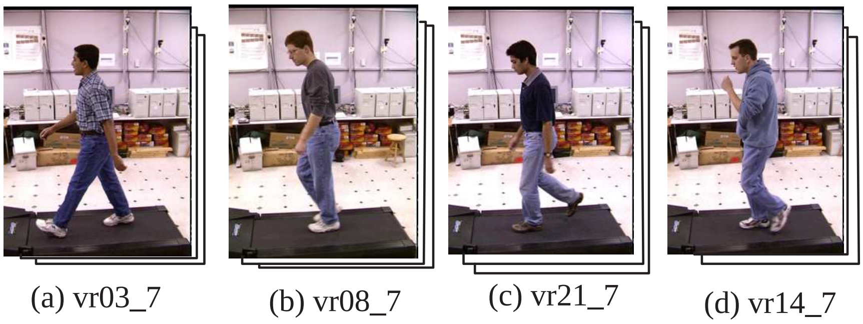
6.7 Experiment Results and Conclusion:

We used the CMU Mobo database for algorithm validation [115], which includes video sequences captured from 25 individuals walking on a treadmill. Our algorithm was tested on four side-view sequences: vr03_7, vr08_7, vr21_7 and vr14_7 as shown in Fig.6.7, and we selected a walking cycle of 33 frames for each sequence. The first three have a normal walking style with different appearances (body shapes and colors), and the last one has an abnormal pattern where the subject touched his nose during walking. We have three specific experiments. The first one evaluates the tracking accuracy that is measured by the localization accuracy of each body part over a complete cycle. The second shows the capability of handling an unseen activity with an occluded body part. The last one presents body part segmentation that is part of online appearance learning.

The proposed method was implemented in C++, and the testbed is a PC computer with a Core 2 Duo E4700/2.6GHz CPU and 2GB RAM. One feature of this work is its capability of online simultaneous learning of temporal and spatial priors, where the size of the slide window for training sample selection has to be determined. We found that a number between 5-12 frames is acceptable. It takes about 200 ms for BC-GPLVM training (100 iterations) over 12 frames. Part-based evaluation is about 200 ms per frame that include the localization and segmentation of each body part. After the initialization on the first 5-12 frames, the proposed tracking algorithm can run at about 2 fps for the following frames.

6.7.1 Part Localization

To evaluate the accuracy of articulated human tracking, we have manually obtained the ground-truth (position and orientation) of six body parts in all test video frames, i.e., the *Head*, *Torso*, *Right_arm* (R_arm), *Left_arm* (L_arm), *Right_leg* (R_leg), *Left_leg* (L_leg). Similar to [136], we evaluate the tracking accuracy by comparing the esti-



(a) vr03_7 (b) vr08_7 (c) vr21_7 (d) vr14_7

Figure 6.7: Four test video sequences (320 × 240, 33 frames).

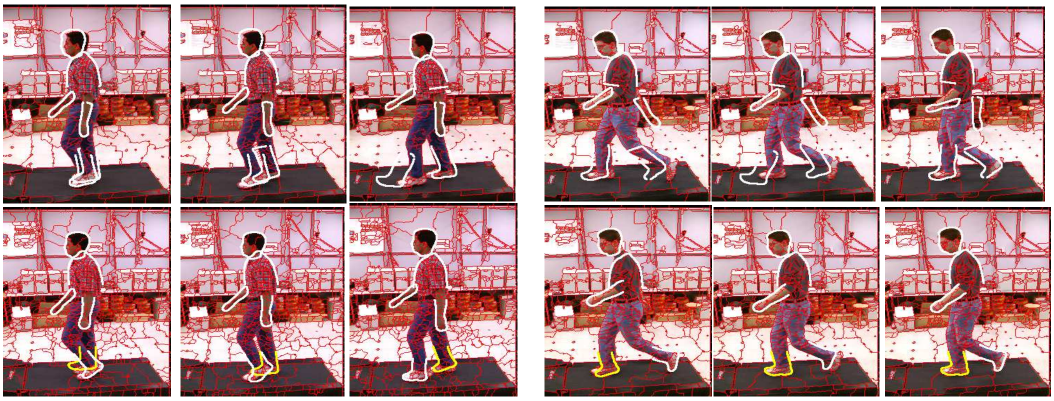


Figure 6.8: The comparison of part localization between the *hybrid* approach [5] (top) and the proposed one (bottom) for three continuous frames from two test videos.

mated position/orientation with the ground-truth ones. There are two competing algorithms both of which only involve the same part-bard spatial prior that has to been trained off-line for each typical pose (i.e., *High-point*, *Contact* and *Passing*). One is the *1-fan* method [116] that involves the edge histogram for part detection, and the other is the *hybrid* approach [5] where coupled edge-region shape priors are used for part detection. Both algorithms require several pose specific spatial priors that are learned off line, and they do not involve any temporal prior by treating each frame independently. Although the hybrid approach improves the localization accuracy for six body parts compared with the 1-fan method, it is time-consuming due to the fact that segmentation is involved in part detection. The proposed algorithm is much more efficient and effective because the combined spatio-temporal priors are online learnt for each pose (a continuous variable) and dramatically narrows the local search for part detection.

The comparative results are shown in Table 6.1, where the results from the 1-fan and hybrid approaches are the averages over three typical poses, while that of ours is the average over a complete walking cycle. Our algorithm demonstrates significant improvements over the two competing algorithms in tracking accuracy and efficiency. The localization performance is quite consistent over all three test videos. Also, we report the results of orientation estimation for each body part in Table 6.2. Some visual comparisons for two test videos can be seen from Fig. 6.8, where we can see obvious advantages of our method even under occlusion.

The comparison above shows the advantage of using the combined spatio-temporal priors over one with the spatial prior alone. One may wonder how about the tracking results of using the temporal prior only. It was shown in our experiments, when the background is clean and no occlusion, the temporal prior alone could be sufficient given reasonable part-based appearance models. However, when the background is cluttered (with many false alarms) or occlusion occurs, the contribution from the spatial prior cannot be neglected. Or in the other words, although there is some *redundancy* by using the two priors together, this redundancy is essential to the tracker's robustness and adaptability.

6.7.2 Special Case Handling

One major advantage of online learning is the ability to handle unseen motion patterns and even occlusion. Sequence vr14_ 7 shows an abnormal walking pattern that is hard to cope with for a tracker that relies on off-line learning. Moreover, one arm is occluded most of time during the walking cycle. The two competing algorithms fails in this case since the spatial prior learned off line is not able to deal with this

Table 6.1: The comparison of localization error (in pixel).

Methods	Head	Torso	R_arm	L_arm	R_leg	L_leg
1-fan [116]	5.3	6.8	12.2	11.1	12.7	12.8
hybrid [5]	5.9	6.0	9.8	8.6	4.8	5.6
vr03_7	0.6	0.9	1.3	1.8	1.3	1.0
vr08_7	0.6	1.8	0.9	4.8	1.7	1.1
vr21_7	0.9	1.0	1.5	7.4	2.7	2.7
Average	0.7	1.3	1.2	4.7	1.9	1.6

Table 6.2: The orientation estimation error (°) of proposed method.

Videos	Head	Torso	R_arm	L_arm	R_leg	L_leg
vr03_7	7.1	1.8	1.0	1.1	3.1	2.5
vr08_7	2.3	2.4	1.1	1.0	1.3	2.2
vr21_7	2.9	2.2	1.3	3.8	3.8	4.3

abnormality. However, the proposed algorithm can accurately detect all visible body parts, regardless of the unusual motion pattern and one occluded arm. Some tracking results are shown in Fig. 6.9. It is shown that the proposed tracker can effectively localize the arm that deviates from its normal motion pattern, proving the usefulness of online learning. Although the majority of one arm is occluded in most frames, the tracking results of other body parts are not affected, showing the effectiveness of occlusion handling.

6.7.3 Part Segmentation

The proposed algorithm can also online learn part-based appearance models from frame to frame during tracking. Part detection in this work is similar to the one used in [5]. After we localize the whole body (we use the Torso as the reference part) in

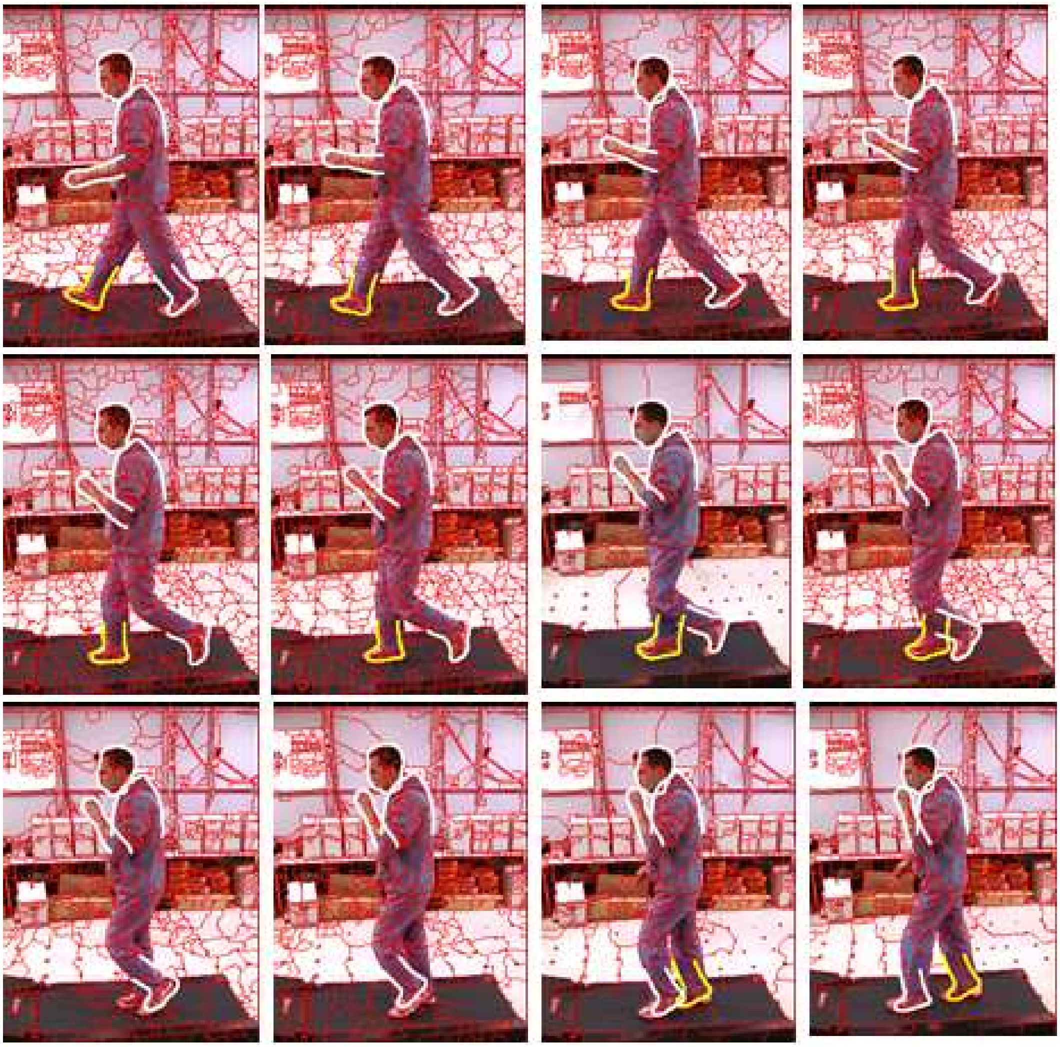


Figure 6.9: Tracking results for an abnormal activity.

the current frame, we can localize and segment each body part correspondingly. The segmented body parts can be used to speed-up part detection by providing an object template that can be updated sequentially by the tracker. Some examples of online learned part appearances are shown in Fig. 6.10. Although each body parts exhibit significant shape/color variability among four test sequences, the segmentation results are quite accurate and robust.

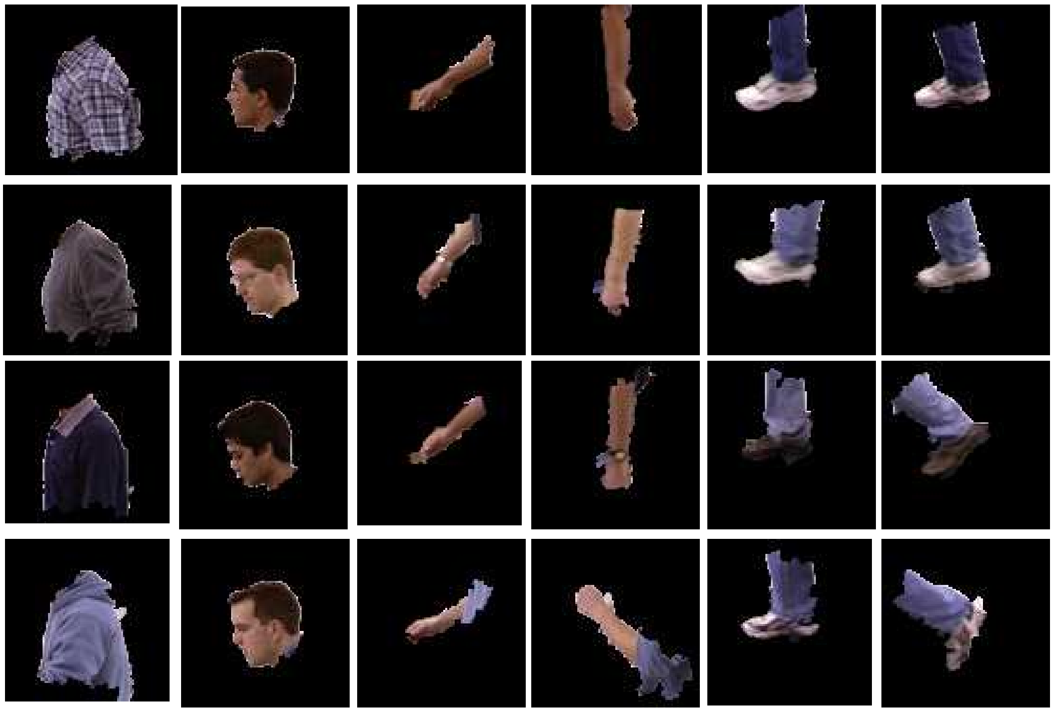


Figure 6.10: Online learned part appearance.

6.8 Conclusion

In this chapter, based on our low-level and middle-level vision study results, we have proposed a new algorithm for articulated human tracking that combines both the spatial and temporal priors in an online learning framework. Compared with prior efforts, we want to fully take advantage of both the spatial and temporal priors in a balanced way in order to optimize the tracking performance. Although there might be certain redundancy between the two priors, the synergistic use of them greatly enhances the robustness and adaptability of the tracker, especially in a challenging environment with complex background or occluded pats. The online learning mechanism makes the proposed algorithm effective to track subjects with significant appearance and motion variability. All of these makes our algorithm a promising tool to support video-based human motion analysis in a general setting.

CHAPTER 7

Conclusions and Future Works

7.1 Conclusions

In this dissertation, we have studied human detection, tracking and segmentation, where the research issues range from low-level vision to high-level vision. We first build a biologically plausible computational model for our algorithm based on our functional understanding of the HVS. Guided by this model, we investigate some low-level vision problems, such as joint spatial-temporal grouping, short/middle-range motion feature extraction and two kinds of non-convex classification problems. Then we investigate some mid-level vision problems, such as the usage of the complimentary information of region and edge features in a combined bottom-up and top-down framework. At the end, our research focuses on high-level vision problems. First, we propose a hybrid body representation for integrated pose recognition, localization and segmentation of the whole body, as well as body parts, in a single image. Then, we extend our success from image-based to video-based processing by exploiting the complementary context information in both temporal priors and spatial priors. Our simulation results show that our current algorithm can successfully detect and segment all body parts despite the significant variability. We conclude this dissertation as follows.

- We show that a biologically plausible comprehensive computational model can guide computer vision algorithm designing to achieve significant performance improvement.

- For low-level and middle-level bottom-up processing, we advocate UC-region(or called super-pixel) based compact image/video representation. We suggest a cascaded multi-stage classification architecture which can combine the merits of both statistical modeling and graph theory approaches. Using a split-and-merge paradigm, extracted mid-level region-based motion and color features can be used to deal with the no-convex classification problems, more meaningful segmentation results can be obtained with less over-segmentation and under-segmentation in some complex and realistic scenarios. This cascaded multi-stage classification frame work is also computationally effective.

- For middle-level vision, we suggest an effective segmentation based hypothesis-and-test paradigm based on coupled region-edge shape prior which unifies the representation of both region-based and edge-based shape prior. Given a configuration hypothesis, the region-based shape prior is used to guide a bottom-up segmentation. The edge-based shape prior is used to evaluate the obtained segmentation result as well as a configuration hypothesis. In this way, a correct localization will facilitate object segmentation, and a good segmentation will enhance the confidence of a localization hypotheses. The optimal segmentation and the spatial configuration can be obtained simultaneously. The obtained segmentation result can be further refined through an improved Graph-cut based method, in which both region-based and edge-based shape priors are jointly involved. Our experiments demonstrate that this framework leads to significant localization and segmentation performance improvements over some state-of-the-art approaches.

- For high-level vision image based processing, we advocate a hybrid body representation for integrated pose recognition, localization and segmentation of the whole body as well as body parts in a single image. A typical pose is represented by both template-like view information and part-based structural information. Specifically, each body part as well as the whole body are represented by an off-line learned shape model where both region-based and edge-based priors are combined in a coupled shape representation. Part-based spatial priors are represented by a "star" graphical model. This hybrid body representation can synergistically integrate pose recognition, localization and segmentation into one computational flow.

- For high-level vision video based processing, we point out that integrate spatial and temporal priors is essential to robust tracking an articulated human body from a monocular video sequence. The spatial prior represented by a star-structured graphical model is embedded in the temporal prior. Both temporal and spatial priors can be online learned in a seamless fashion through the Back Constrained Gaussian Process Latent Variable Model (BC-GPLVM) that involves a moving window for training sample selection. Experimental results show that the new algorithm can achieve accurate tracking and localization results for different walking subjects with significant appearance and motion variability.

Our research provides practical tools for human motion analysis from the image based initialization step to video-based analysis. This research will lead to a significant progress for human-oriented video analysis technologies that are playing an increasingly important role in homeland security and many other human-based video applications. However, human tracking and segmentation problem is far away from solved. Our algorithms still leave many aspects that can be further improved in the future.

7.2 Future Research

In order to achieve fast, robust tracking and segmentation, the future research will focus on the following issues:

7.2.1 Combining Both Appearance and Motion Information

According to the "two-pathway" theory [14, 15], we know that that both appearance and motion information are involved in motion perception in the HVS. However, our current work do not explicitly exploit motion information. The main reason is that the robust motion feature extraction is a nontrivial task. The current pixel-based motion feature extraction methods, such as optical-flow, have difficulty providing reliable motion features. Some studies [143] [144] advocate region-based motion estimation approaches, which can extract more reliable and robust motion features. Since our approach can provide body part segmentations, so motion vector of these body parts might be easily estimated. Therefore, our framework has the potential to combine both appearance and motion information. Fully exploiting this potential may greatly boost the performance of our algorithm, which is an interesting topic for further study.

7.2.2 Building a Joint Spatio-temporal Inference Framework

Our current framework for human detection, tracking and segmentation is still an image based approach. Pose configuration is estimated and body parts are segmented frame by frame instead of in a spatio-temporal blob. Considering this point, our current approach is only a partially "biologically plausible" approach. For a joint spatio-temporal processing, pose configuration estimation and body part segmentation should be performed in a spatio-temporal blob. How to develop an effective inference and prior learning framework for this spatio-temporal processing is still an open issue.

BIBLIOGRAPHY

[1] T. Webmaster, Feb 2008, http://www.healthsystem.virginia.edu.

[2] S. E. Palmer and I. Rock, “Rethinking perceptual organization: The role of uniform connectedness,” *Psychonomic Bulletin and Review*, vol. 1, pp. 29–55, 1994.

[3] S. Akaho and O. Michel, “Gaussian mixture model em algorithm,” http://diwww.epfl.ch/mantra/tutorial/english/Gaussian/html, 2004.

[4] Dermot, 2007, http://www.idleworm.com/how/anm/02w/walk1.shtml.

[5] C. Chen and G. Fan, “Hybrid body represenation for integrated pose recognition, localization and segmentation,” in *Proc. IEEE CVPR*, 2008.

[6] T. Webmaster, Feb 2008, http://www.researchandmarkets.com/reports/435324.

[7] S. Sarkar and K. L. Boyer, “Perceptual organization in computer vision: A review and a proposal for a classificatory structure,” *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 23, pp. 382–399, 1993.

[8] T. Pavlidis, “Why progress in machine vision is so slow,” *Pattern Recogn. Lett.*, vol. 13, no. 4, pp. 221–225, 1992.

[9] S. Palmer, *Vision Science: Photons to Phenomenology.* Bradford Books MIT Press, 1999.

[10] S. Gepshtein and M. Kubovy, “The emergence of visual objects in space-time,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, pp. 8186–8191, 2000.

[11] M. Kubovy and S. Gepshtein, “Gestalt: from phenomena to laws,” in *Perceptual Organization for Artificial Vision Systems.* Academic Publishers, Boston, 2000, pp. 41–71.

[12] J. L. McClelland, “On the time relations of mental processes : An examination of systems of processes in cascade,” *Psychological Review*, vol. 86, pp. 287–330, 1979.

[13] M. Minsky and S. A. Papert, *Perceptrons - Expanded Edition.* The MIT Press, 1987.

[14] L. G. Ungerleider and M. Mishkin, *Two cortical visual systems.* Cambridge, MA, USA: MIT Press, 1982, pp. 549 – 586.

[15] F. J. elleman and V. E. D.C., “Distributed hierarchical processing in primate cerebral cortex,” *Cerebral Cortex*, vol. 1, pp. 1–47, 1991.

[16] M. Kubovy and S. Gepshtein, “Grouping in space and in space-time: An exercise in phenomenological psychophysics,” in *Perceptual Organization in Vision: Behavioral and Neural perspectives*, M. Behrmann, R. Kimchi, and C. Olson, Eds. Lawrence Erlbaum Association, Mahwah, N.J., 2003, pp. 45–85.

[17] C.N.Olivers and G. Humphreys, “Spatiotemporal segregation in visual search:evidence from parietal lesions,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, pp. 667–688, 2004.

[18] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, MA: The MIT Press, 1979.

[19] D. Marr, *Vision: a Computational Investigation into the Human Representation andProcessing of Visual Information.* W.H. Freeman and Company, NY, 1982.

[20] R. A. Rensink and J. T. Enns, “Preemption effects in visual search: Evidence for low-level grouping,” *Psychological Review*, vol. 102, pp. 101–130, 1995.

[21] S. Han, G. W. Humphreys, and L. Chen, “Uniform connectedness and classical gestalt principles of perceptual grouping,” *Perception & Psychophysics*, vol. 61, pp. 661–674, 1999.

[22] N. A. Stillings, S. E. Weisler, C. H. Chase, M. H. Feinstein, J. L. Garfield, and E. L. Rissland, *Cognitive Science: An Introduction.* Cambridge, MA: The MIT Press, 1995.

[23] S. E. Palmer, J. L. Brooks, and R. Nelson, “When does grouping happen?” *Acta Psychologica*, vol. 114, pp. 311–330, 2003.

[24] S. Ullman, *High-level vision : object recognition and visual cognition.* Cambridge, MA: The MIT Press, 1996.

[25] O. Braddick, “A short-range process in apparent motion,” *Vision Res.*, vol. 14, pp. 519–527, 1974.

[26] I. M. Thornton, J. Pinto, and M. Shiffrar, “The visual perception of human locomotion,” *Cognitive Neuropsychology*, vol. 15, pp. 535–552, 1998.

[27] C.W.Clifford, J.Freedman, and LM.Vaina, “First- and second-order motion perception in gabor micropattern stimuli: Psychophysical and computational modelling,” *Cogn Brain Res*, vol. 6, pp. 263–271, 1998.

[28] S. L. Franconeri, J. Halberda, L. Feigenson, and G. A. Alvarez, “Common fate can define objects in multiple object tracking,” *Journal of Vision*, vol. 4, no. 8, p. 365a, 2004.

[29] A. Witkin and J. Tenenbaum, *On the role of structure in vision.* W.H. Freeman and Company, NY, 1982.

[30] D. Lowe, *Perceptual Organization and Visual Recognition.* Boston, MA: Kluwer, 1985.

[31] K. L. Boyer and S. Sarkar, "Perceptual organization in computer vision:status, challenges, and potential," *Guest Editorial in Computer Vision and Image Understanding*, vol. 76, pp. 1–5, 1999.

[32] K. L. Boyer and S. Sarkar, Eds., *Perceptual Organization for Artificial Vision Systems.* Kluwer Academic Publishers, 2000.

[33] S. Sarkar, D. Majchrzak, and K. Korimilli, "Perceptual organization based computational model for robust segmentation of moving objects," *Comput. Vis. Image Underst*, vol. 86, no. 3, pp. 141–170, 2002.

[34] D. Mumford, *Neuronal architecture for pattern-theoretic problems.* Cambridge, MA, USA: MIT Press, 1993.

[35] J. Bullier, "Integrated model of visual processing," *Brain Reseach Reviews*, vol. 36, pp. 96–107, 2001.

[36] D. Kersten, P.Mamassian, and A.Yuille, "Object perception as bayesian inference," *Annual Review of Psychology*, vol. 55, pp. 271–304, 2004.

[37] D.C.Knill and W.Richards, *Perception as Bayesian Inference.* UK: Cambridge Univ. Press, 1996.

[38] R.P.N.Rao, B. Olshausen, and M. Lewicki, *Probabilistic Models of the Brain: Perception and Neural Function.* Cambridge, MA: MIT Press, 2002.

[39] T. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *Journal of Optical Society of America*, vol. 20, no. 7, pp. 1434–1448, 2003.

[40] S. P. Vecera and R. C. O'Reilly, "Figure-ground organization and object recognition processes: An interactive account," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, pp. 441–462, 1998.

[41] J. Shi and J. Malik, "Motion segmentation and tracking using Normalized cuts," in *Proc. IEEE International Conference on Computer Vision*, 1998, pp. 1151–60.

[42] C. Fowlkes, S. Belongie, and J. Malik, "Efficient spatiotemporal grouping using the Nystrom method," in *Proc. IEEE CVPR*, vol. 1, 2001, pp. 231–238.

[43] D. DeMenthon and R. Megret, "Spatio-temporal segmentation of video by hierarchical mean shift analysis," *Technical Report: LAMP-TR-090/CAR-TR-978/CS-TR-4388/UMIACS-TR-2002-68*, 2002.

[44] H. Greenspan, J. Goldberger, and A. Mayer, "A probabilistic framework for spatio-temporal video representation and indexing," in *Proc. European Conference on Computer Vision*, vol. 4, Berlin, Germany, 2002, pp. 461–475.

[45] R. Megret and D. DeMenthon, "A survey of spatio-temporal grouping techniques," University of Maryland, College Park, Tech. Rep., March 2002, http://www.umiacs.umd.edu/lamp/pubs/TechReports/.

[46] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 897–915, 1998.

[47] V.Mezaris, I.Kompatsiaris, and M.G.Strintzis, "Video object segmentation using bayes-based temporal tracking and trajectory-based region merging," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, pp. 782–795, 2004.

[48] M. Gelgon and P. Bouthemy, “A region-level motion-based graph representation and labeling for tracking a spatial image partition.” *Pattern Recognition*, vol. 33, no. 4, pp. 725–740, 2000.

[49] D. Wang, “Unsupervised video segmentation based on watersheds and temporal tracking,” *IEEE Trans. on Circuits and Systems for Video Technology*, pp. 539–546, 1998.

[50] F. Porikli and Y. Wang, “Automatic video object segmentation using volume growing and hierarchical clustering,” *Journal on Applied Signal Processin*, vol. 3, pp. 442–453, March 2004.

[51] Y. Tsai, C.Lai, Y.Hung, and Z.Shih, “A bayesian approach to video object segmentation,” *IEEE Trans. Circuits syst. video Technology*, vol. 15, pp. 175–180, 2005.

[52] S. Hochstein and M. Ahissar, “View from the top: Herarchies and reverse hierarchies in the visual system,” *Neuron*, vol. 36, pp. 791–804, 2002.

[53] E. Borenstein, E. Sharon, and S. Ullman, “Combining top-down and bottom-up segmentation,” in *Proc. of IEEE CVPR*, 2004.

[54] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, “Image parsing: Unifying segmentation, detection, and recognition,” *Intenational Journal of Computer Vision*, vol. 63, pp. 113 – 140, 2005.

[55] D. Burr and J. Ross, “Vision: The world through picket fences,” *Current Biology*, vol. 14, pp. 381–382, 2004.

[56] M. A. Giese and T. Poggio, “Neural mechanisms for the recognition of biological movement,” *Nature Neuroscience Review*, vol. 4, pp. 179–192, 2003.

[57] M. W. Oram and D. I. Perrett, “Integration of form and motion in the anterior part of the superior temporal polysensory area (stpa) of the macaque monkey,” *Journal of neurophysiology*, vol. 76, pp. 109–129, 1996.

[58] P. Sajda and K. Baek, “Integration of form and motion within a generative model of visual cortex,” *Neural Networks*, vol. 17, pp. 809–821, 2004.

[59] J. Bullier, “Integrated model of visual processing,” *Brain research review*, vol. 36, pp. 96–107, 2001.

[60] D.Kahneman, A. Terisman, and B. J. Gibbs, “The reviewing of object files: object specific integration of information,” *Cognitive Psychology*, vol. 24, pp. 175–219, 1992.

[61] Z. Pylyshyn and R. Storm, “Tracking multiple independent target: Evidence for a parallel tracking mechanism,” *Spatial Vision*, vol. 3, pp. 1–19, 1988.

[62] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” in *Proc. IEEE ICCV*, 2003.

[63] H. Lim, V. Morariu, O. I. Camps, and M. Sznaier, “Dynamic appearance modeling for human tracking,” in *Proc. IEEE CVPR*, 2006.

[64] Z. Tu and S.-C. Zhu, “Image segmentation by data-driven Markov chain monte carlo,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 657–673, 2002.

[65] A. Micilotta and R. Bowden, “View-based location and tracking of body parts for visual interaction,” in *Proc. of British Machine Vision Conference*, 2004, pp. 849–858.

[66] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. on Image Processing*, vol. 13, no. 11, pp. 1491–1506, 2004.

[67] R. Koenen, "Overview of the mpeg-4 standard (v. 21)," ISO/IEC JTC1/SC29/WG11 N4668, March 2002.

[68] J. M. Martmnez, "Mpeg-7 overview (ver.8)," ISO/IEC JTC1/SC29/WG11 N4980, July 2002.

[69] C.Fowlkes, S. Belongie, F.Chung, and J.Malik, "Spectral grouping using the nystrom method," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 214 – 225, 2004.

[70] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifier," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[71] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 835–850, 2005.

[72] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, vol. 11, no. 2, pp. 417–431, 1983.

[73] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 384–396, 2003.

[74] L. Liu and G. Fan, "Combined key-frame extraction and object-based video segmentation," *IEEE Trans. Circuits and System for Video Technology*, vol. 15, pp. 869–884, July 2005.

[75] X. Song and G. Fan, “Joint key-frame extraction and object segmentation for content-based video analysis,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 16, no. 7, pp. 904–914, 2006.

[76] ——, “Selecting salient frames for spatiotemporal video modeling and segmentation,” *IEEE Trans. Image Processing*, vol. 16, pp. 1–12, 2007.

[77] S. K. Pal and P. Mitra, “Multispectral image segemntation using the rough-set-initialized em algorithm,” *IEEE Trans. Geoscience and remote sensing*, vol. 40, pp. 2495–2501, 2002.

[78] S. Zhong and J. Ghosh, “A unified framework for model-based clustering,” *Journal of Machine Learning Research*, vol. 4, pp. 1001–1037, 2003.

[79] E. C. Hildreth and S. Ullman, *The Computational Study of Vision.* Cambridge, MA, USA: MIT Press, 1989, ch. 15, pp. 581–630.

[80] M. H. Law, M. A. T. Figueiredo, and A. K. Jain, “Simultaneous feature selection and clustering using mixture models,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1154– 1166, September 2004.

[81] P. Felzenszwalb and D. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[82] C. A. Bouman, “Cluster: An unsupervised algorithm for modeling Gaussian mixtures,” Purdue University, Tech. Rep., Oct. 2001, http://dynamo.ecn.purdue.edu/∼bouman/software/cluster/.

[83] M. Vlachos, G. Kollios, and D. Gunopulos, “Discovering similar multidimensional trajectories,” in *Proc. of International Conf. on Data Engineering*, 2002, p. 673 684.

[84] A. Desolneux, L. Moisan, and J.-M.Morel, “A grouping principle and four applications,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 508–513, 2003.

[85] S. Chien, S. Ma, and L. Cheng, “Efficient moving object segmentation algorithm using background registration technique,” *IEEE Trans. Circuits syst. video Technology*, vol. 12, pp. 577–586, 2002.

[86] X. Munoz, J. Freixenet, X. Cufi, and J. Marti, “Strategies for image segmentation combining region and boundary information,” *Pattern recognition letters*, vol. 24, pp. 375–392, 2003.

[87] L. Fisher and J. W. V. Ness, “Admissible clustering procedures,” *Biometrika*, vol. 58, pp. 91–104, 1971.

[88] S. X. Yu and J. Shi, “Object-specific figure-ground segregation,” in *Proc. IEEE CVPR*, 2003.

[89] D. Freedman and T. Zhang, “Interactive graph cut based segmentation with shape priors,” in *Proc. IEEE CVPR*, 2005.

[90] X. Li and G. Hamarneh, “Modeling prior shape and appearance knowledge in watershed segmentation,” in *The 2nd Canadian Conference on Computer and Robot Vision*, 2005, pp. 27–33.

[91] T. Chan and W. Zhu, “Level set based shape prior segmentation,” in *Proc. IEEE CVPR*, 2005.

[92] D. Cremers, T. Kohlberger, and C. Schnorr, “Shape statistics in kernel space for variational image segmentation,” *Pattern Recognition*, vol. 36, pp. 1929–1943, 2003.

[93] P. Srinivasan and J. Shi, "Bottom-up recognition and parsing of the human body," in *Proc. of IEEE CVPR*, 2007.

[94] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: combining segmentation and recognition." in *Proc. IEEE CVPR*, 2004.

[95] J. Wang, E. Gu, and M. Betke, "Mosaicshape: Stochastic region grouping with shape prior," in *Proc. IEEE CVPR*, 2005.

[96] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," in *Proc. ECCV*, 2002.

[97] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV'04 Workshop on Statistical Learning in Computer Vision*, 2004.

[98] G. Mori and J. Malik, "Recovering 3d human body configurations using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1052–1062, 2006.

[99] A. K. Jain, Y. Zhong, and M.-P. Dubuisson-Jolly, "Deformable template models: A review," *Signal Processing*, vol. 71, pp. 109–129, 1998.

[100] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321–331, 1988.

[101] G. Tsechpenakis and D. N. Metaxas, "CRF-driven implicit deformable model," in *Proc. IEEE CVPR*, 2007.

[102] X. Ren, C. C. Fowlkes, and J. Malik, "Cue integration in figure/ground labeling," in *Advances in Neural Information Processing Systems 18*, 2005.

[103] D. Ramanan, "Using segmentation to verify object hypotheses," in *Proc. of IEEE CVPR*, 2007.

[104] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[105] U. Hillenbrand and G. Hirzinger, "Probabilistic search for object segmentation and recognition," in *Proc. ECCV*, 2002.

[106] S. Morishita, "Computing optimal hypotheses efficiently for boosting," in *Progress in Discovery Science*, 2002, pp. 471–481. [Online]. Available: citeseer.ist.psu.edu/492998.html

[107] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based onimmersion simulations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 583–598, 1991.

[108] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[109] P. D. Smet, R. Pires, D. D. Vleeschauwer, and I. Bruyland, "Activity driven non-linear diffusion for color image watershed segmentation," *Journal of Electronic Imaging*, vol. 8, pp. 270–278, 1999.

[110] Y.-C. Lin, Y.-P. Tsai, Y.-P. Hung, and Z.-C. Shih, "Comparison between immersion-based and toboggan-based watershed image segmentation," *IEEE Trans. on Image Processing*, vol. 15, pp. 632–640, 2006.

[111] L. Vincent, "Morphological grayscale reconstruction in image analysis: efficient algorithms and applications," *IEEE Trans. on Image Processing*, vol. 2, pp. 176–201, 1993.

[112] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla, "Shape context and chamfer matching in cluttered scenes," in *Proc. IEEE CVPR*, 2003.

[113] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," in *Cornell Computing and Information Science Technical Report TR2004-1963*, 2004.

[114] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *International Journal of Computer Vision*, vol. 70, pp. 109–131, 2006.

[115] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-01-18, 2001.

[116] D. Crandall, P. Felzenszwalb, and D. Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *Proc. IEEE CVPR*, 2005.

[117] L. Greengard and J. Strain, "The fast Gauss tranform," *SIAM J. Scientific Computing*, vol. 2, pp. 79–94, 1991.

[118] A. Elgammal, R. Duraiswami, and L. S. Davis, "Efficient non-parametric adaptive color modeling using fast Gauss transform with applications to color modeling and tracking," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, 2003, pp. 1499–1504.

[119] D. Lang, M. Klaas, F. Hamze, and A. Lee, "N-body methods code and matlab binaries," 2006, *http : //www.cs.ubc.ca/ awll/nbody_methods.htm.*

[120] N. Sprague and J. Luo, "Clothed people detection in still images," vol. 03, 2002.

[121] S. Ioffe and D. A. Forsyth, "Probabilistic methods for finding people," *International Journal of Computer Vision*, vol. 43, pp. 45–68, 2001.

[122] P. Felzenszwalb and D. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61(1), pp. 55–79, January 2005.

[123] X. Ren, A. C. Berg, and J. Malik, “Recovering human body configurations using pairwise constraints between parts,” in *Proc. IEEE ICCV*, 2005.

[124] C. McIntosh, G. Hamarneh, , and G. Mori, “Human limb delineation and joint position recovery using localized boundary models,” in *IEEE Workshop on Motion and Video Computing*, 2007.

[125] D. Ramanan and C. Sminchisescu, “Training deformable models for localization,” in *Proc. IEEE CVPR*, 2006.

[126] J. E. Hummel and B. J. Stankiewicz, “Two roles for attention in shape perception: A structural description model of visual scrutiny,” *Visual Congnition*, vol. 5, pp. 49–79, 1998.

[127] R. L. Goldstone, “Object, please remain composed,” *Behavioral and brain sciences*, vol. 21, pp. 472–473, 1998.

[128] A. Bissacco, M. H. Yang, and S. Soatto, “Detecting humans via their pose,” in *Neural Information Processing Systems (NIPS)*, 2006.

[129] D. Ramanan, “Learning to parse images of articulated bodies,” in *Proc. of Neural Info. Proc. Systems (NIPS)*, 2006.

[130] S. Lee, G. Wolberg, and S. Y. Shin, “Scattered data interpolation with multilevel b-splines,” *IEEE Trans. on Visualization and Computer Graphics*, vol. 3, pp. 229–244, 1997.

[131] T. Zoller and J. M. Buhmann, “Shape constrained image segmentation by parametric distributional clustering,” in *Proc. IEEE CVPR*, 2005.

[132] X. Lan and D. P. Huttenlocher, “A unified spatio-temporal articulated model for tracking,” *Proc. of IEEE CVPR*, vol. 1, pp. 722–729, 2004.

[133] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian process dynamical models for human motion,” *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, pp. 283–298, 2008.

[134] R. Urtasun, D. J. Fleet, and P. Fua, “3d people tracking with Gaussian process dynamical models,” in *Proc. IEEE CVPR*, 2006.

[135] R. Urtasun and T. Darrell, “Sparse probabilistic regression for activity-independent human pose inference,” in *Proc. IEEE CVPR*, 2008.

[136] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isardy, “Tracking loose-limbed people,” in *Proc. IEEE CVPR*, 2004.

[137] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Tracking people by learning their appearance,” *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, pp. 65–81, 2007.

[138] N. Lawrence, “Probabilistic non-linear principal component analysis with Gaussian process latent variable models,” *Journal of Machine Learning Research*, vol. 6, p. 1783C1816, 2005.

[139] N. D. Lawrence and J. Quinonero-Candela, “Local distance preservation in the gp-lvm through back constraints,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 513 – 520.

[140] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley, “Real-time body tracking using a Gaussian process latent variable model,” in *Computer Vision, ICCV 2007. IEEE International Conference on*, 2007.

[141] L. Raskin, E. Rivlin, and M. Rudzsky, "Using Gaussian process annealing particle filter for 3d human tracking," *EURASIP Journal on Advances in Signal Processing*, 2008.

[142] C. H. Ek, P. H. Torr, and N. D. Lawrence, "Gaussian process latent variable models for human pose estimation," in *Machine Learning for Multimodal Interaction*, 2007.

[143] J. B. Kim and H. J. Kim, "Efficient region-based motion segmentation for a video monitoring system," *Pattern Recognition Letters*, vol. 24, pp. 113–128, 2003.

[144] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by map labeling of watershed segments," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 23, pp. 326–332, 2001.

VITA

Cheng Chen

Candidate for the Degree of

Doctor of Philosophy

Dissertation: HUMAN DETECTION, TRACKING AND SEGMENTATION FROM LOW-LEVEL TO HIGH-LEVEL VISION

Major Field: Electrical Engineering

Biographical:

Education:

Received the B.S. degree from Shenyang Institute of Technology, Shenyang, Liaoning, P.R.China, 1992, in Major Mechanical Engineering
Received the M.S. degree from University of Electronic Science and Technology of China, Chengdu, Sichuan, P.R.China, 1997, in Mechatronics
Completed the requirements for the degree of Doctor of Philosophy with a major in Electrical Engineering, Oklahoma State University in December, 2008.

Name: Cheng Chen Date of Degree: December, 2008

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: HUMAN DETECTION, TRACKING AND SEGMENTATION FROM LOW-LEVEL TO HIGH-LEVEL VISION

Pages in Study: 152 Candidate for the Degree of Doctor of Philosophy

Major Field: Electrical Engineering

The goal of this research is to detect, segment and track a human body as well as estimate its limb configuration from cluttered background. These are fundamental research issues that have attracted intensive attention in the computer vision community because of their wide applications. Meanwhile they also remain to be ones of the most challenging research issues largely due to the ubiquitous visual ambiguities in images/videos. The other challenging factor is the ill-posed nature of the problems. Inspired by the recent findings in cognitive psychology, we adopt several biologically plausible approaches to attack these challenging problems. This dissertation provides a comprehensive study of human detection, tracking and segmentation that covers several research issues ranging from low, middle, and high-level vision.

In low-level vision, we investigate video segmentation where the main challenge is the non-convex classification problem, and we develop a cascaded multi-layer segmentation framework where no-convex classification problems are addressed in a split-and-merge paradigm by combining merits of both statistical modeling and graph theory.

In middle-level vision, we propose a segmentation based hypothesis-and-test paradigm to achieve joint localization and segmentation that exploits the complementary nature of region-based and edge-based shape priors. In addition, we integrate both priors into a Graph-cut framework to improve the segmentation results.

In high-level vision, our research has two related parts. First, we propose a hybrid body representation that embraces part-whole shape priors and part-based spatial prior for integrated pose recognition, localization and segmentation in a given image. Second, we further combine spatial and temporal priors in an integrated online learning and inference framework, where body parts can be detected, localized and segmented simultaneously from a video sequence. Both of them are supported by previous low-level and mid-level vision tasks.

Experimental results show that the proposed algorithms can achieve accurate and robust tracking, localization and segmentation results for different walking subjects with significant appearance and motion variability and under cluttered background.

ADVISOR'S APPROVAL: ______________________________