Dressed Human Modeling, Detection, and Parts Localization

Liang Zhao

CMU-RI-TR-01-19

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

July 26, 2001

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

This research was partially sponsored by PennDOT Agreement No. 62N111 Project TA-34 executed under the Federal Technical Assistance Grant Program 001-008-701-99-7 and SAIC.

Abstract

This dissertation presents an integrated human shape modeling, detection, and body part localization vision system. It demonstrates that the system can (1) detect pedestrians in various shapes, sizes, postures, partial occlusion, and clothing from a moving vehicle using stereo cameras; (2) locate the joints of a person automatically and accurately without employing any markers around the joints.

The following contributions distinguish this dissertation from previous work:

1. Dressed human modeling and dynamic model assembling: Unlike previous work that employs a fixed human body model or global deformable template to perform human detection, in this dissertation merged body parts are introduced to represent the deformations caused by clothing, segmentation errors, or low image resolution. A dressed human model is dynamically assembled from the model parts in the recognition step; the shapes of the body parts and the size and spatial relationships between them (the contextual information) are represented as invariant under translation, rotation, and scaling. Therefore, the system can detect people in different clothes, positions, sizes, and orientations.

2. Bayesian similarity measure: A probabilistic similarity measure is derived from the human model that combines the local shape and global relationship constraints to guide body part identification and human detection. Thus, the identification of a part does not only depend on its own shape but also the contextual constraints from other parts. In contrast with previous work, the proposed similarity measure enables

efficient shape matching and comparison robust to articulation, partial occlusion, and segmentation errors through coarse-to-fine human model assembling.

3. Recursive context reasoning algorithm: Contour-based human detection depends on reliable contour extraction, but contour extraction is an under-constrained problem without the knowledge about the objects to be detected. Unlike previous work that assumes perfect and complete contours are available, this dissertation proposes a recursive context reasoning (RCR) algorithm to solve the above dilemma. A contour updating procedure is introduced to integrate the human model and the identified body parts to predict the shapes and locations of the parts missed by the contour detector; the refined contours are used to reevaluate the Bayesian similarity measure and to determine if a person is present or not. Therefore, contour extraction, body part localization, and human detection are improved iteratively.

Acknowledgements

First of all I would like to thank my advisor Chuck Thorpe for his advice, encouragement and patience, particularly during the final stages of this dissertation. I would also like to thank the other members of my thesis committee, Steve Seitz, Bob Collins, and Larry Davis for their comments and insights on this work.

I am especially grateful to Dr. Tomaso Poggio and Constantine Papageorgiou at MIT for providing me their pedestrian detection program so that I could conduct a performance comparison between my system and theirs.

In the course of graduate studies, I have benefited from interactions with many people. I would especially like to thank my old officemates, Daniel Huber and Stewart Moorehead for many interesting discussions, both technical and otherwise. Henry Schneiderman, Jianbo Shi, Yanghai Tsin, and Bob Wang have provided helpful feedback on my work. John Kozar, Christoph Mertz, Bob Wang, and Dave Duggins have helped me collect data and conduct experiments outdoor.

Finally, I want to thank my parents for keeping things in perspective by asking me, now and then, whether I had finished my thesis yet.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"The First Law of Robotics: A robot must not injure a human being or, through inaction, allow a human being to come to harm." — Isaac Asimov

This dissertation presents an integrated human shape modeling, detection, and body part localization vision system. It demonstrates that the system can (1) detect pedestrians in various shapes, sizes, postures, partial occlusion, and clothing from a moving vehicle using stereo cameras; and (2) locate the joints of a person automatically and accurately without employing any markers around the joints. This is achieved through TRS[1]-invariant dressed human modeling and dynamic model assembling, a Bayesian similarity measure, and a recursive context reasoning procedure, as shown in Figs. 1.1 and 1.2.

1.1 Motivation and Goal

Automatic human detection and body part localization are important and challenging problems in computer vision. The solution to these problems can be employed in a wide range of applications such as safe robot navigation, visual surveillance, human-computer interface, performance measurement for athletes and patients with disabilities, virtual reality,

[1]TRS is the abbreviation of translation, rotation, and scaling

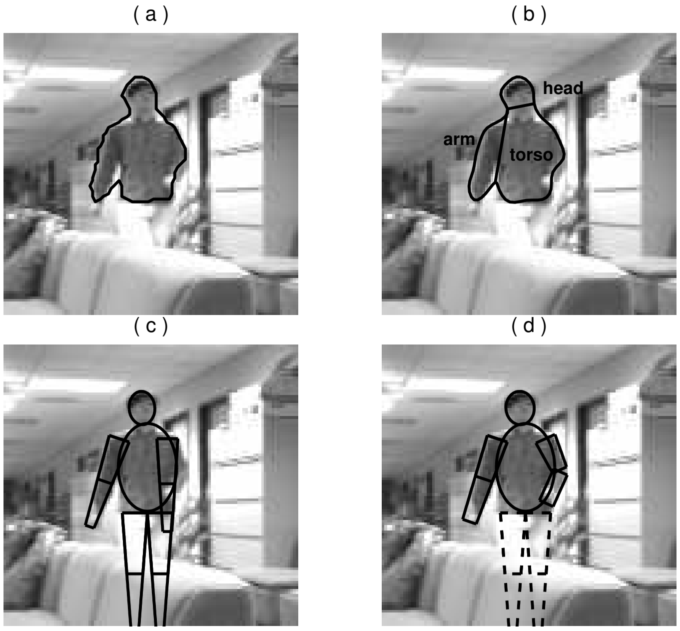

Figure 1.1: Human detection and body part localization: (a) initial contour detection (b) body parts identification (c) contour prediction (d) contour alignment

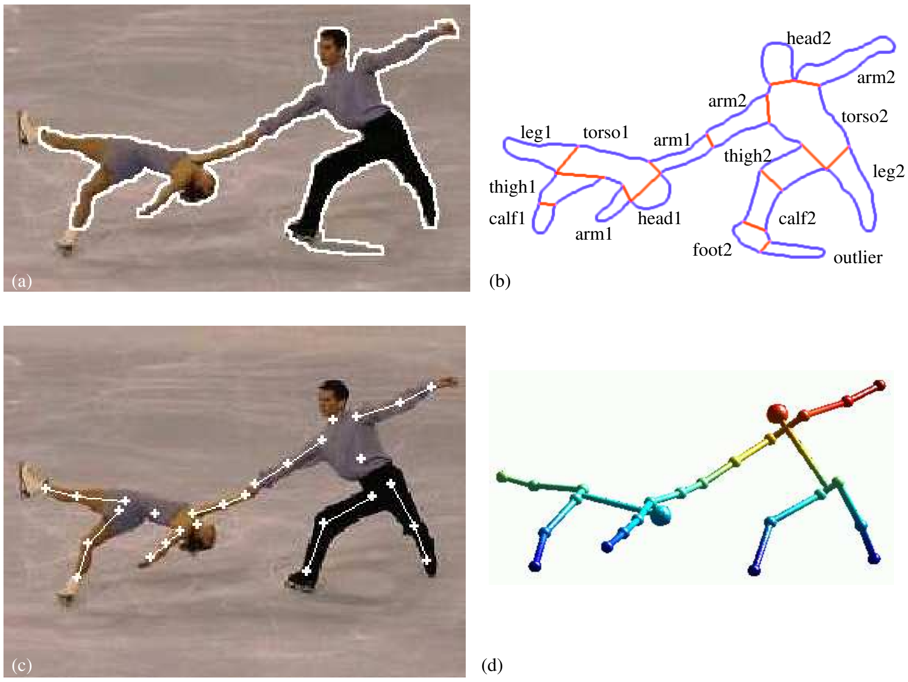

Figure 1.2: Locating the body parts of multiple people

and figure animation. Example areas involving human detection and body part localization include:

- **mobile robot navigation**: A mobile robot with the ability to detect people can work among people more safely and accomplish more tasks, such as following a person or giving a tour without running into people. This follows the first law of robotics created by the popular science fiction writer Isaac Asimov as cited at the beginning of this chapter.
- **visual surveillance**: A computer with the ability to sense people can monitor a security region to check if somebody breaks through and can report this person's actions.
- **human motion capture**: Body part localization is essential for human motion capture, which has applications in figure animation, virtual reality, and human-computer interaction. It can be used to evaluate the performance of an athlete and help him/her to correct inaccurate actions. Furthermore, a user can communicate with the computer more easily by gestures.
- **shape-based image retrieval**: Beyond human detection, the metrics and methods developed here have other uses. A recent survey about cognition aspects of object retrieval shows that users are more interested in retrieval by shape than by color and texture [14]. However, retrieval by shape is still considered one of the most difficult aspects of content-based search. The Bayesian similarity measure proposed in this thesis can be used to perform shape classification for database creation and image retrieval.

The research goal of this dissertation is the design, implementation, and validation of reliable and accurate methods for human detection and body part localization. Two goals are addressed and achieved simultaneously under the same probabilistic framework. Because humans are a class of objects with articulated parts and deformable shapes, research on human detection and body part localization will not only produce a wide range of applications but will also shed light on general object detection and shape analysis. The next section discusses the challenges of human detection and body part localization, and presents my strategies to handle these challenges.

1.2 Challenges and Strategies

1.2.1 Why is it difficult to detect humans

Although humans can easily detect other humans and estimate their body part locations from a single image, these problems are inherently difficult for a computer. The difficulties stem from the number of degrees of freedom in the human body, self-occlusion, appearance variation due to clothing, and the ambiguities in the projection of a 3D human shape onto the image plane. Figures 1.3 and 1.4 show some examples that illustrate appearance changes due to clothing and articulation. Contour features (the silhouettes of objects) are commonly used to overcome variable texture and illumination. However, to extract perfect and complete contours of objects from a cluttered scene is very difficult. Although many contour extraction methods have been proposed [121, 122], they tend to have errors and are distracted by scene clutter, as shown in Fig. 1.5. To locate the joints of a person is even harder, because these are hidden by muscle, skin, and clothing. Segmentation or contour extraction errors also pose a significant challenge to accurate joint localization. Although a great number of object recognition and detection methods have been proposed (see [12]), none of them can handle the above challenges very well.

The context for human detection is the general object detection problem. There are two approaches to object detection. One approach is to search the whole image at multi-scales for objects. This is a time consuming procedure and may result in multiple responses from a single object. Another approach is to first segment foreground objects from the background, then classify each segmented object as human or non-human. In this dissertation, I employ the second approach to human detection. Classifying only segmented objects rather than whole images significantly reduces computational complexity. Several methods such as depth-based segmentation [123], background subtraction [80], or frame differencing can

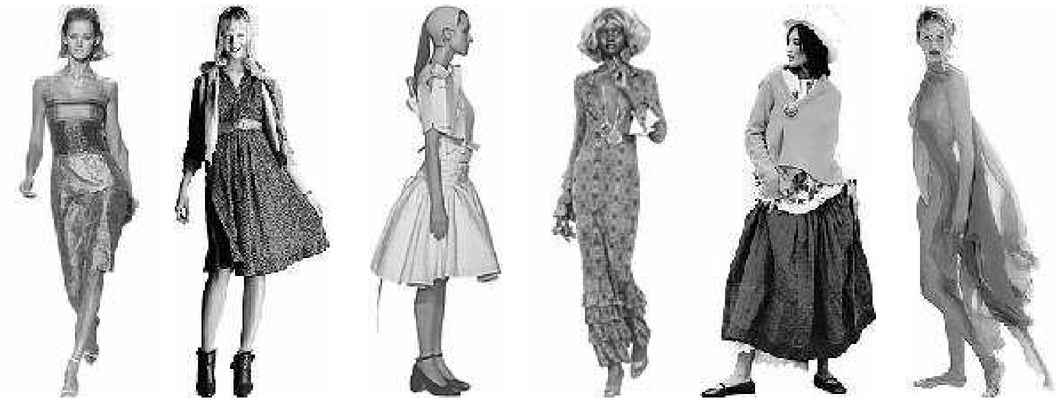

Figure 1.3: Examples of various appearances due to clothing

Figure 1.4: Examples of various appearances due to articulation

Figure 1.5: Examples of silhouette extraction using depth segmentation

be employed to separate foreground objects from the background.

This dissertation focuses on how to classify a previously segmented object as human or non-human. This is an object classification problem. Usually an object classifier consists of an object representation/model and a classifier [96]. A good object model should allow the recognition of objects independent of their positions, orientations, sizes, and articulation for articulated objects. It should also accommodate variations among the instances of an object class and should be insensitive to objects with partially missing parts.

Object classification by shape is difficult mainly because:

- shapes of a class can be distorted by sensor noise or digitization;
- shapes of a class can differ due to varying viewpoint;
- shapes of a class can be deformed non-rigidly. For example, an object such as a human may have moving parts and may be flexible (see Fig. 1.4);
- some parts of a shape may not be visible. These partially occluded shapes need to be classified correctly.

Human detection is even more difficult because of shape variation due to clothing. The texture and shape variances among dresses make even the same person appear significantly different when wearing different dresses (see Fig. 1.3).

1.2.2 How to handle the difficulties

There are many problems to address for an object detection method. First of all, we need to consider what features to use for recognition. Contours (the silhouettes of objects) are commonly used to overcome variable texture. Therefore, I employ the occluding contours or the silhouettes of the human body parts as features to detect humans. Parts-based approaches [21, 25, 31, 58, 78, 80] can handle occlusion and articulated motion effectively. The importance and relevance of significant parts for object recognition has been verified by numerous psychophysical experiments [48]. My recognition scheme is thus based on the shapes of body parts and the relationships between them. Then the questions left are how to decompose a silhouette into parts, and how to represent the shapes and the relationships between the parts.

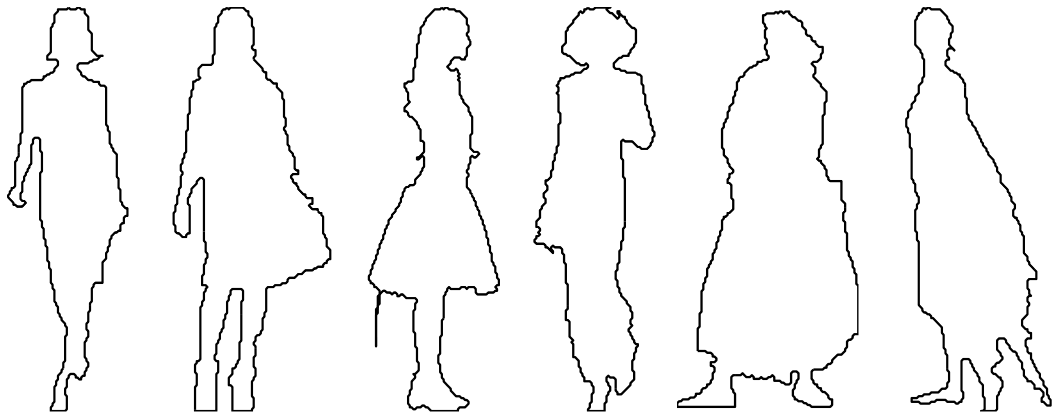

Figure 1.6: Examples of various shapes due to clothing (same as Fig. 1.3 but only contour)

Shape decomposition should result in real/intuitive parts of an object (as shown in Fig. 2.1). However, without prior knowledge of the identification of the object, this goal is hard to achieve. Many shape decomposition approaches tend to produce over-segmented partitions due to noise and local distortions of shapes. Although the multi-scale shape decomposition approach [42] can reduce the effect of noise, the procedures of curve evolution tend to be very slow. In this thesis, I develop an efficient shape decomposition method to yield a fine-to-coarse organization of parts. The over-segmented parts are grouped into coarse level parts. The matching based on this multi-scale shape decomposition starts from the coarse level of the hierarchy and goes down. This enables the significant parts to be matched first, which then constrains the match of insignificant parts. The hierarchical matching of parts is both efficient and robust to over-segmentation of a silhouette.

Many approaches [23, 31] have been proposed to perform matching purely based on the hierarchical structure, but they can not distinguish humans from other animals with a similar shape hierarchy. In order to do so, geometric metrics are needed to constrain the shapes and spatial relationships between the parts. In this dissertation I develop a TRS[2]-invariant representation of the shapes of the body parts and their size and spatial relationships. For the human model, the probability distributions are used to handle the shape variations between different people. This results in a TRS-invariant probabilistic

[2]TRS is the abbreviation of translation, rotation, and scaling

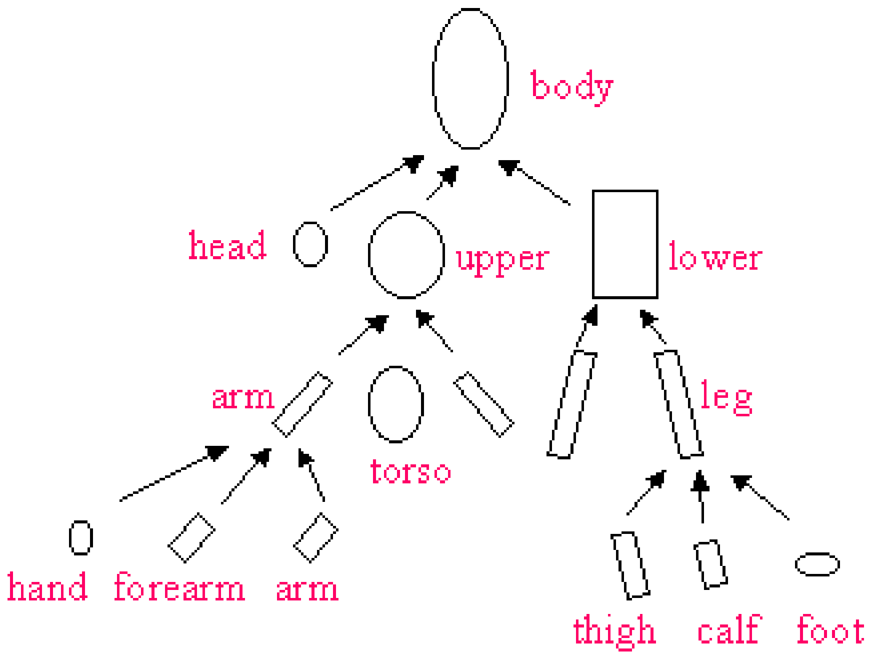

Figure 1.7: Generating merged body parts

human model.

Previous work usually does not model clothes but only the human body. However, clothes may drastically change the shape of a person (see Fig. 1.6). One of the effects of clothes is that they cover some body parts and merge them into a single component. This makes it difficult to distinguish the covered parts. The second effect is that the clothes may generate some spurious body parts along the silhouette that distract from the locations of the real body parts. Although the global deformable templates [32, 41] can model some shape deformations, they have difficulty dealing with large articulated motions and partial occlusion.

To handle these effects, I first introduce merged body parts to model the merging of multiple body parts as shown in Fig. 1.7. Using the merged body parts, various shape configurations can be built as shown in Fig. 1.8, and the locations of the real body parts can be inferred from the merged parts covering them. The models containing the merged body parts are called dressed human models; they can represent the deformations caused

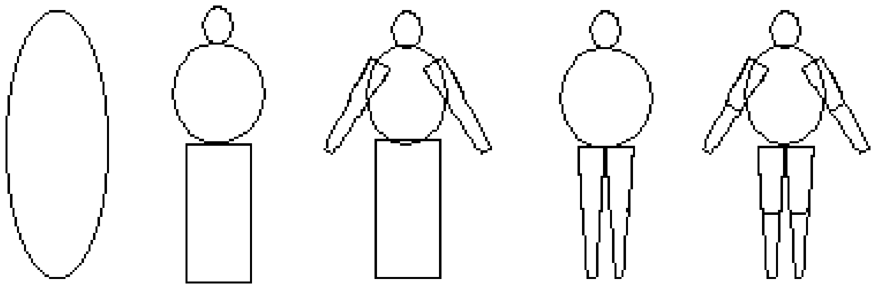

Figure 1.8: Assembling the human models

by clothing, segmentation errors, or low image resolution. A dressed human model is dynamically assembled from the model parts in the body part identification procedure. An evaluation function is developed to select the appropriate model parts and assembling scheme to label the decomposed contour segments. The identification of a part does not only depend on its own shape but also on contextual constraints from other parts. Thus, the labeling is globally optimal and the real body parts can be discriminated from the pseudo parts generated by clothes or other objects held by the person. Multiple labelings are used to handle situations where multiple people are segmented as a single region, as shown in Fig. 1.2.

Second, a Bayesian similarity measure is derived from the human model that combines the local shape and global relationship constraints into a single equation to evaluate the degree of resemblance between a contour and the assembled human model. In contrast with previous work, the Bayesian similarity measure enables efficient shape matching and comparison robust to articulation, partial occlusion, and segmentation errors through coarse-to-fine human model assembling.

Third, a coarse-to-fine procedure is developed to locate the joints between body parts accurately: (1) match the extracted ribbons with the model body parts based on the derived similarity measure; (2) infer the locations of the missed body parts from the identified body parts; and (3) adjust the locations of the joints to achieve consistency with the modeled size and spatial relationships between the body parts.

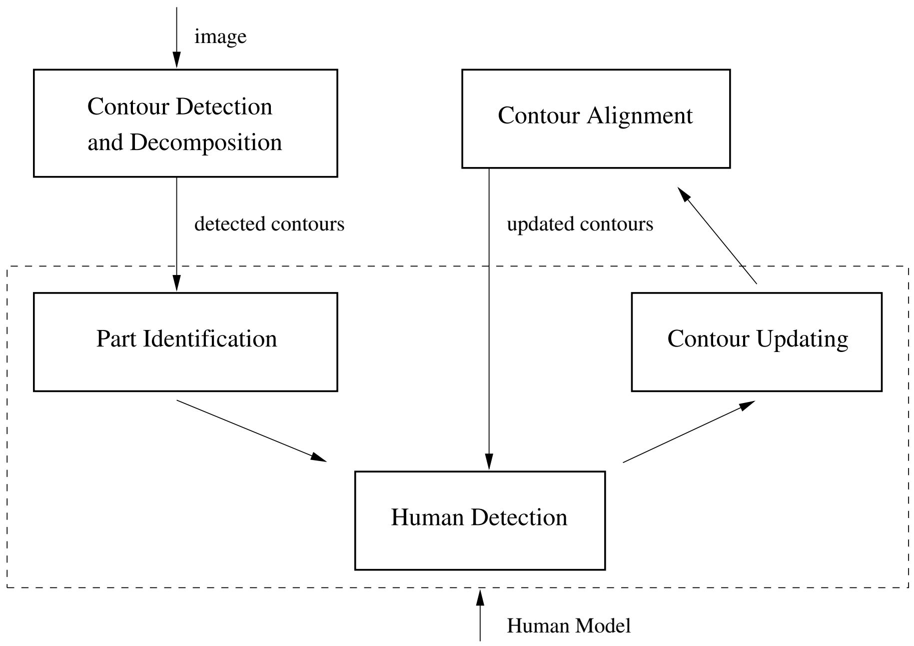

Figure 1.9: Flow chart of the RCR algorithm

Previous work usually assumes that the initial contour extraction and shape decomposition are perfect. However, no robust contour extraction and shape decomposition exist yet to achieve this goal. In fact contour extraction and shape decomposition are underconstrained problems without the knowledge about the objects to be detected. In contrast I develop a recursive context reasoning (RCR) algorithm (see Fig. 1.9) to solve the above dilemma. Specifically, a contour updating procedure is introduced to integrate the detected body parts and the human model using the weighted Least Squares method to predict the contours of the missed parts. The predicted contours are aligned with edge features, and the refined contours are used to reevaluate the likelihood of a person being present in the image. Therefore, both contour extraction, human detection, and body part localization are improved accordingly. A probabilistic model is developed to perform context reasoning and to integrate the above procedures into a single framework. Experiments in cluttered scenes demonstrate that the proposed approach is robust to occlusion and variations in people's appearance.

1.3 Application Context: Driver Collision Warning System

While there are many areas to which human detection and body part localization can be applied, the chosen domain for this dissertation is vehicle safety, and most of the results are presented in the context of a driver warning system. Despite the specific nature of the chosen application area, the underlying ideas and technologies presented in this dissertation are readily applicable to other tasks described in Section 1.1 and to detecting other objects.

The sponsoring program is titled "Development and testing of performance specifications for a next generation side collision warning system", sponsored by USDOT [15, 16]. The goal of the project is to warn bus drivers of potential collisions by focusing on blind spots along the sides of a city bus. There are several parts to the project: studying accident records and determining causes of bus collisions; developing object detection sensors; predicting bus motion; and developing the driver interface. The preliminary functional goals first call for detecting vehicles, fixed objects, bicycles, and pedestrians. Each of these categories of objects can be detected as a 3-D object; but each has different motion characteristics and each should be identified separately. Next, the functional goals call for predicting bus motion, for example incipient motion when the bus is about to pull away from a stop, the sweeping motion of a bus turning through an intersection, or the angled motion of a lane change. Object detection will be especially important in the region the bus is about to traverse. Motion prediction can come directly, from rate gyros and wheel encoders; indirectly, from a route map and positioning system; or by inference, for example noting that the door has just been closed. Finally, the system will generate graded warnings. At the lowest level, the system will generate non-intrusive "situation awareness" information, similar to the driver looking in the rear-view mirrors to check for nearby objects. At the next level, the system will generate "alerts", a gentle notification (audio or video) of a potentially dangerous situation. The system will then progress to one or more levels of "warning", an intrusive alarm designed to notify the driver of impending collision.

Pedestrian detection is especially important for this project. While there are relatively few collisions with pedestrians, they tend to be more serious: the number of fatalities from bus-pedestrian collisions is about the same as the number of fatalities from bus-vehicle collisions, even though there are many more bus-vehicle crashes. A fixed object will usually stay put; pedestrians are more unpredictable. The driver will receive a higher level alarm for a pedestrian than for a sign post in the same location.

At the same time, the state of the art of commercial side-looking systems is not very useful for pedestrian detection. Most side-looking systems are designed to cover blind spots on the side of a semi-trailer traveling along a limited-access highway. The systems are designed to look for cars or trucks, and therefore the sensors are widely spaced and are tuned to find large metal objects. The users are specifically warned that these systems are not designed to detect pedestrians.

Thus, a successful system such as the one described in this dissertation could play a large role in a collisions warning system. While the initial implementation will be for city buses, pedestrian safety is also a large issue for construction equipment, agricultural equipment, and other large moving vehicles that operate in close proximity to people.

1.4 Dissertation Overview

The remainder of the dissertation is organized as follows:

Chapter 2 illustrates the part-based shape representation and the dressed human model. First I give the definition of natural parts and present a hierarchical algorithm to decompose a shape. Then, I describe a TRS-invariant shape representation, and the dressed human models that represent the shape deformations caused by clothing, segmentation errors, and low image resolution.

Chapter 3 describes the similarity measure derived from Bayes rule, and explains how to use this similarity measure to perform body part labeling and to determine if a shape is a person or not.

Chapter 4 presents a recursive context reasoning algorithm to refine the performance of human detection, body part localization, and contour extraction.

Chapter 5 describes two applications of the techniques developed in this thesis: pedestrian detection and human motion capture. Experimental results on pedestrian detection and joint localization are given to demonstrate the effectiveness of the algorithm.

Chapter 6 discusses the contributions of this thesis and suggests future work.

Chapter 2

Dressed Human Modeling

2.1 Requirements for a Good Object Class Model

Human modeling is an essential part of model-based human detection. Although a great number of human models have been proposed in the literature, few of them are appropriate for human detection. Most models are developed for other purposes, such as human tracking [34, 36, 59, 61] or figure animation [35]. These models are either too complicated to be practical for efficient human detection, or can just be used to detect a particular person rather than all instances of humans. The common drawbacks with previous human models are (1) the representations of human shapes are not invariant to similarity transforms, thus, they can only detect people of a fixed size or orientation; (2) the models are usually specific to a particular person, and do not model the statistical variance among individuals; (3) most models only represent the shape of a human body, but cannot handle the shape variation due to clothing; and (4) although some models such as deformable templates can handle certain global shape variance, they have difficulty dealing with large articulated motion and partial occlusion.

The human model proposed in this dissertation overcomes these drawbacks. It satisfies all the following requirements of a good object class model for object classification:

1. It should not depend on scale, orientation, and position of objects;

2. It should handle view-dependent shape variation;

3. It should be robust to shape distortions resulting from digitization noise and foreground/background segmentation errors;

4. It should be robust to partial occlusions of an object;

5. It should allow for articulated moving parts;

6. It should not be influenced by the shape variations allowed within the class; and

7. It should support efficient shape recognition/classification.

The part-based human shape model proposed in this thesis satisfies requirements 1 through 7. Requirements 4 and 5 are satisfied by decomposing a contour into visual parts. The satisfaction of requirements 1,2,3 and 6 follows from the fact that the shapes of the parts are abstracted as ribbons, the representation of body part shapes and their relationships is invariant under translation, rotation, and scaling, and probability distributions are used to accommodate the shape variations among individuals as well as absorbing some variations in shape due to viewpoint. The hierarchical organization of the body parts allows efficient object recognition (requirement 7) in a coarse-to-fine manner.

The remainder of this chapter is organized as follows: Section 2.2 shows that there is a substantial body of work on human modeling, although no technique currently exists that satisfies all of requirements 1 through 7. In Section 2.3, a natural shape decomposition method is presented to extract visual parts from a silhouette. In Section 2.4, a TRS-invariant probabilistic human body model is developed to represent the shapes and relationships of body parts. In Section 2.5, merged parts are introduced to handle the merging of multiple real body parts, and the body parts are organized in a hierarchy to facilitate efficient recognition.

2.2 Previous Work on Human Modeling

There is a large collection of literature on human modeling. Most models employ part-based representations to handle articulation. They vary widely in their level of detail. At one extreme are methods that crudely model the body as a collection of articulated planar patches [34]. At the other extreme are 3D models in which the limb shapes are deformable [35, 36]. For part-based 2D models, the representation of parts varies from planar patches [34] and 2D ribbons [79, 25] to deformable models [77]. The advantage of using 2D models for recognition is that the matching is between 2D and 2D. The disadvantage is that it is hard for 2D models to deal with shape variations due to viewpoint. For 3D models, if 3D data is available we can match the model directly against the data. Gavrila and Davis [36] proposed a complex 3D model of the body that takes into account kinematic constraints, but their method requires searching through a high dimensional pose parameter space for 3D pose recovery. If only 2D data is available, we need to match the 3D model against the extracted 2D data. Assumptions about the viewing conditions vary from scaled orthographic projection [65] to full perspective [66, 67]. To account for large variations in depth, Hogg [58] modeled the body in terms of articulated 3D cylinders viewed under perspective projection. More sophisticated tapered cylinders [67, 68] or superquadrics [70] have been employed. Bowden *et al.* [38] encapsulated the correlation between 2D image data and 3D skeleton pose in a hybrid 2D-3D model trained on real life examples. The model they used allows 3D inference from 2D data, but their method does not generalize easily to new camera positions, because their 2D model is not invariant to viewpoint. The common drawback with the above models is that they do not model the statistical variation among individuals and the effects of clothes on human shape. Thus, they may be used for human tracking or figure animation, but they are not appropriate for detecting people of various shapes and clothing.

Marr and Nishihara [17] proposed a hierarchical 3D human model. At the highest level of the hierarchy, the body is modeled as a large extended cylinder, which is then resolved into small cylinders forming limbs and torso, and so on to fingers and toes. This hierarchical representation is stable in the presence of noise and sensitive to fine-level features, but is impractical because it contains few actual constraints to support human detection.

Contour-based representations have been used to model the 2D human shape. Baumberg *et al.* and Sullivan *et al.* [32, 33] employed a deformable template to handle shape deformation, where the shape model is derived from a set of training shapes. The orthogonal shape parameters are estimated using Principle Component Analysis(PCA). One drawback with this approach is that the model and the extracted contour should be aligned first, which is not a trivial task. Another drawback is that some invalid shapes are produced by the combination of two or more linear deformations. Gavrila *et al.* [41] developed a template hierarchy to capture the variety of human shapes, and the model contains no invalid shapes. The common drawback with the above approaches is that they do not model individual parts, and so they can only handle limited shape variety due to articulation and cannot deal with occlusion very well.

Skeleton-based representations [23] have been used to model the topological structure of the human body, but they do not model the shapes of body parts. These approaches are sensitive to noise and cannot distinguish two classes with the same topological structure but different geometrical structures.

Some models incorporate other cues or features into the model. Pentland [60] introduced a blob-based representation that combines skin color and contour to represent a body part. While the color-blob representation of a person is quite useful, it is not invariant under clothing/lighting changes and so it requires an initial model learning procedure for different subjects and a smoothly changing image background. Papageorgiou *et al.* [37] developed a wavelet-based representation to model pedestrians, but this representation is not invariant under rotation and can not handle large part movements and occlusion very well.

In summary, previous human models only satisfy some of the requirements listed in Section 2.1, but not all of them.

2.3 Shape Decomposition for Part-Based Representation

2.3.1 Definition of Natural Parts

Shape representation is a major problem in computer vision and is the basis for recognition. Here the word shape refers to the geometry of an object's occluding contour in two dimensions. The requirements of a good description that facilitates recognition lead to representations that are segmented and hierarchical. Thus, shape decomposition is a key stage in a part-based shape representation. Fig. 2.1 shows some examples of how to decompose a silhouette into subparts. The boundaries between parts are called *cuts* (shown as thick lines in Fig. 2.1), and a *part* is defined as a region bounded by a portion of the outline of a silhouette and one or more cuts.

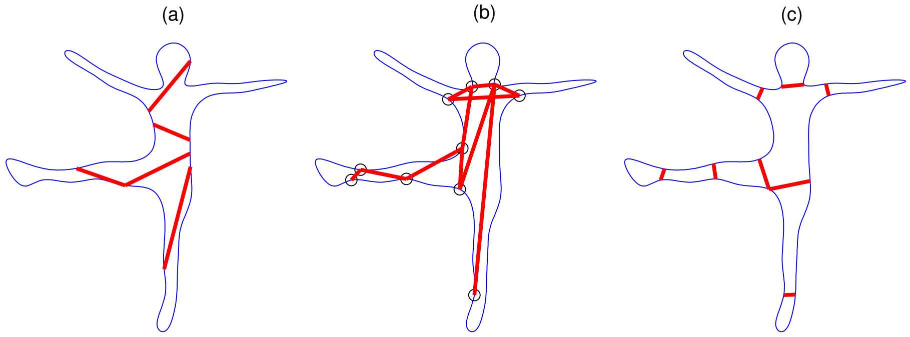

Figure 2.1: Examples of shape decomposition: (a) random decomposition (b) decomposition at negative curvature minima (c) natural decomposition

A silhouette can be decomposed in many different ways as shown in Fig. 2.1, but for the task of recognition not just any partitioning scheme will do. The decomposed parts must satisfy certain requirements for recognition: first, they should correspond to the natural body parts of an object; second, the decomposition should be invariant under translation, rotation, and scaling; third, the decomposition should be computable. These requirements suggest that to break a shape into parts we should use its intrinsic geometry. According to the human intuition about parts, a segmentation into parts occurs at *negative curvature minima* (NCM) as shown with small circles in Fig. 2.1(b). This observation leads to Hoffman and Richards's minima rule [48]:"For any silhouette, all negative minima of curvature of its bounding curve are boundaries between parts." The minima rule constrains cuts to pass through the boundary points it provides, but does not guide the selection of cuts themselves. For example in Fig. 2.1(b), the silhouette of a person is decomposed into parts at NCM, but some segmented parts do not correspond to the natural body parts of a person and one of the legs is not detected at all. This example demonstrates that not every pair of NCM forms a natural part, and some parts such as the limbs of animals may be bounded by a NCM and a non-NCM. Therefore, we need to introduce more constraints to achieve unique and natural shape decomposition.

Singh *et al.* noted that when boundary points can be joined in more than one way to decompose a silhouette, human vision prefers the partitioning scheme which uses the shortest cuts. This leads to the short-cut rule [52] which requires a cut (1) be a straight line, (2) cross an axis of local symmetry, (3) join two points on the outline of a silhouette, such that at least one of the two points has negative curvature, (4) be the shortest one if there are several possible competing cuts. Because only one end of a cut is required to lie on a portion of the boundary with negative curvature, this enables us to decompose a shape such as a leg at the right position as shown in Fig. 2.1(c). Singh *et al.*'s scheme restricts the cut to cross a symmetry axis in order to avoid short but undesirable cuts. However, robust computation of symmetry axes is difficult since from their very definitions [17, 114, 115] most axes are extremely sensitive to noise.

In this thesis, the constraint on the salience of a part is used to replace the second requirement in the short-cut rule in order to avoid the computation of symmetry axes. According to Hoffman and Singh's study [49], there are three factors that affect the salience of a part: the size of the part relative to the whole object, the degree to which the part protrudes, and the strength of its boundaries. Among these three factors, the computation of a part's protrusion (the ratio of the perimeter of the part (excluding the cut) to the length of the cut) is more efficient and robust to noise and partial occlusion of the object. Thus, the protrusion of a part is employed to evaluate its salience; the salience of a part increases as its protrusion increases.

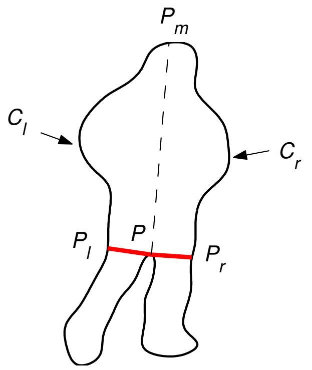

Figure 2.2: Computing the cuts passing through point P

Therefore, the short-cut rule and the salience requirement are combined to constrain the other end of a cut. For example in Fig. 2.2, let S be a silhouette, C be the boundary of S, P be a point on C with NCM, and P_m be a point on C so that P and P_m divide the boundary C into two curves C_l, C_r of equal arc length. Then two cuts are formed passing through point P: $\overline{PP_l}$, $\overline{PP_r}$ such that points P_l and P_r lies on C_l and C_r, respectively. The ends P_l and P_r of the two cuts are located as follows:

$$P_l = arg \min_{P'} ||\overline{PP'}|| \quad \text{s.t.} \quad \frac{||\overset{\frown}{PP'}||}{||\overline{PP'}||} > T_p, P' \in C_l, \overline{PP'} \in S \tag{2.1}$$

$$P_r = arg \min_{P'} ||\overline{PP'}|| \quad \text{s.t.} \quad \frac{||\overset{\frown}{PP'}||}{||\overline{PP'}||} > T_p, P' \in C_r, \overline{PP'} \in S \tag{2.2}$$

where $\overset{\frown}{PP'}$ is the smaller part of boundary C between P and P', $||\overset{\frown}{PP'}||$ is the arc length of $\overset{\frown}{PP'}$, and $\frac{||\overset{\frown}{PP'}||}{||\overline{PP'}||}$ is the salience of the part bounded by curve $\overset{\frown}{PP_l}$ and cut $\overline{PP_l}$.

Eq. (2.1) means that point P_l is located so that the cut $\overline{PP_l}$ is the shortest one among all cuts sharing the same end P, lying within the silhouette with the other end lying on contour C_l, and resulting in significant parts whose salience is above a threshold T_p. The other point P_r is located in the same way using Eq. (2.2).

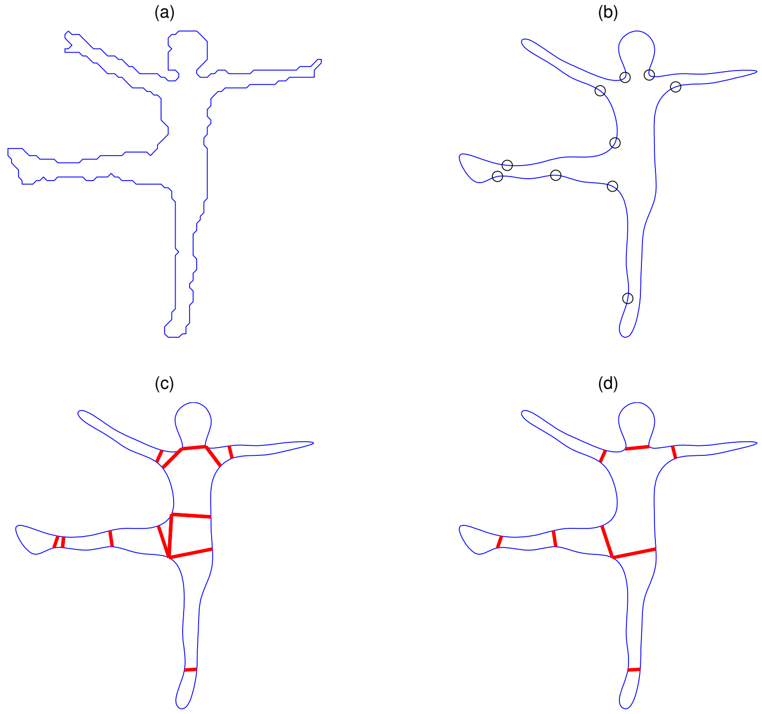

Figure 2.3: Shape decomposition procedure: (a) the original boundary of a silhouette, (b) smoothing the boundary and selecting the significant NCM (illustrated with small circles), (c) computing the cuts of the silhouette using Eqs. (2.1) and (2.2), (d) grouping over-segmented parts.

2.3.2 Computing a Natural Shape Decomposition

The definitions of natural parts and cuts given in subsection 2.3.1 constrain the selection of the best shape partition, but leave open how in practice a partition may be computed from an image despite noise and local distortions of a silhouette. In particular, since negative minima of curvature are obtained by local computation, their computation is not robust in real digital images. One way to handle noise is to smooth the contour of a silhouette. Then we need to select a natural scale to smooth it. If the scale is too small, the description will be affected by noise, and if it is too large, important structures may be lost. Because parts appear at many spatial scales, smaller parts nest within larger ones to form a hierarchy. This observation of shape hierarchy leads to a multi-scale shape decomposition approach based on discrete curve evolution to yield robust computation and hierarchical organization of parts.

Multi-scale shape decomposition is based on curve evolution obtained from different operators. Latecki and Lakmper [54] proposed curve evolution by linearization. A continuous curve is first decomposed into maximal digital line segments. Then the evolution proceeds by substituting two consecutive line segments with a single line segment joining their endpoints. However this approach is very sensitive to the order of point substitution. Malladi and Sethian [55] developed a level-set approach to curve evolution and decomposition. Siddiqi *et al.* [50] proposed a theory of shape based on a reaction-diffusion equation to produce a hierarchical decomposition of shape into parts and protrusions. The main drawback with the multi-scale approach is that the procedure of curve evolution is very slow. For example in [56], a convolution with the Gaussian filter is needed for each scale.

This section presents an efficient and robust shape decomposition algorithm. In contrast with the multi-scale approach, the proposed algorithm takes several computationally efficient strategies to reduce the effects of noise. Fisrt, a B-spline approximation is used to moderately smooth the boundary of a silhouette. Second, the NCM with small magnitude of curvature are removed to avoid parts due to noise or small local deformations. However, curvature is not scale invariant (e.g. its value doubles if the silhouette shrinks by half). One way to transform curvature into a scale-invariant quantity is to first find the chord joining the two closest inflections which bound the point, then multiply the curvature at the point by the length of this chord. The resulting normalized curvature does not change with scale — if the silhouette shrinks to half size, the curvature doubles but the chord halves, so their product is constant.

When using Eqs. (2.1) and (2.2) to compute the cuts of a silhouette, they may result in over-segmented parts as shown in Fig. 2.3(c). Therefore, a post processing step is needed to merge two over-segmented parts that share a cut into a larger one if this larger part cannot be decomposed into significant subparts using Eqs. (2.1) and (2.2). The order of grouping is from the largest to the smallest parts so that the largest one is selected when several

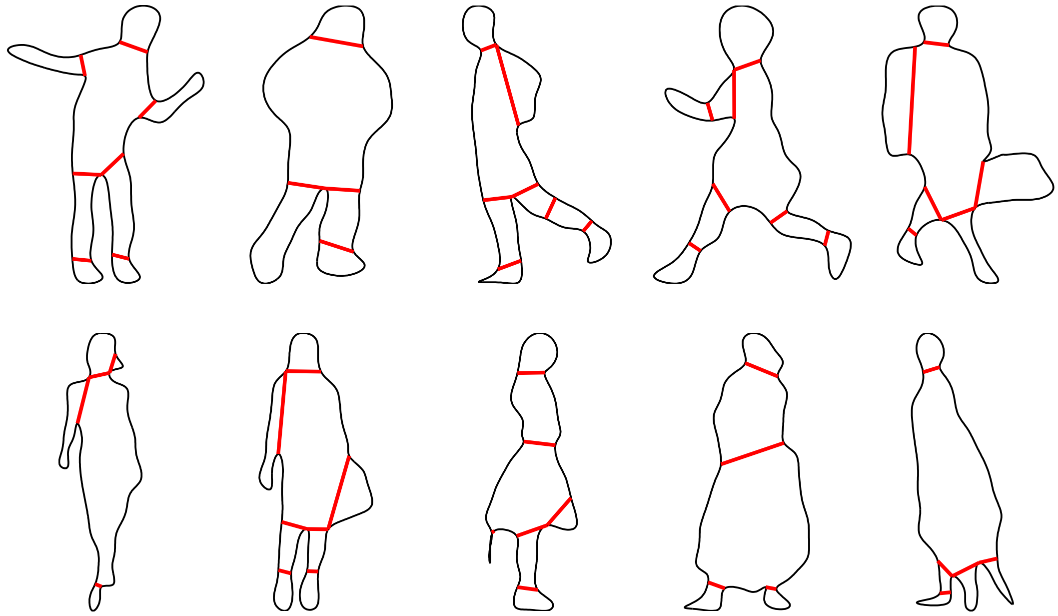

Figure 2.4: Example results of natural shape decomposition

possible competing merges exist. Fig. 2.3 illustrates the whole procedure of the shape decomposition algorithm. The procedure stops when no part can be further decomposed into significant parts and no two parts can be merged into a non-decomposable larger part.

Fig. 2.4 shows several example results from the shape decomposition algorithm. These results demonstrate that the algorithm can produce natural part decompositions that are robust to noise and local deformation. As discussed in the introduction, it is impossible to decompose a shape into a set of perfect subparts without using higher level information. For example, variations in the locations of subparts may occur due to self-occlusion and flexible deformation. There are also missing parts resulting from the inherent difficulty in finding the cut points. All of these will cause non-perfect decomposition, which will be fixed using the model-guided, top-down verification and refinement process presented in Chapter 4.

2.4 Human Body Model

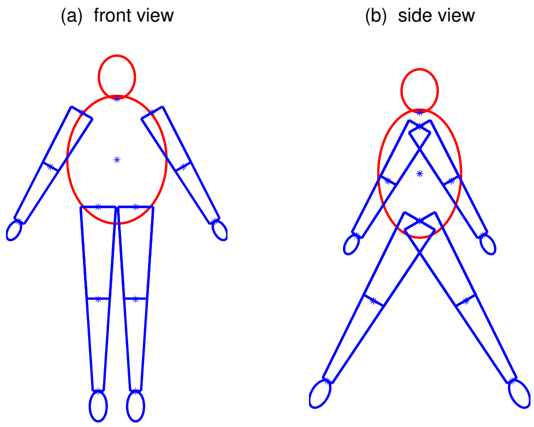

Figure 2.5: Human body model

A human body can be represented directly using a 3D model or indirectly using a collection of 2D models corresponding to different views. Since the goal of this thesis is to detect people in an image, 2D models are preferred to a 3D model because a 2D model can be compared directly with a 2D shape without projecting the 3D model onto the image plane by searching a continuous viewpoint/pose space. The question of how many and which viewpoints to use is an open question and also depends on the application. In the case of pedestrian detection, we found two 2D human body models were sufficient — the front-view and the side-view models as shown in Fig. 2.5. The two models share the same body parts. The main differences are the spatial relationships between the parts and the shape of the torso. The views not modeled by these two models are partially absorbed by the probability distributions of the spatial relationships among the body parts encoded in the human model. For the purpose of human detection and model learning, a TRS-invariant representation of the shapes of parts and the relationships between them is developed. For the purpose of modeling the shape variations between individuals and due to viewpoint changes, probability distributions are employed to encode the variations of the model parameters. The resulting model is called *the TRS-invariant probabilistic model*. The following subsections describes the human body model in detail.

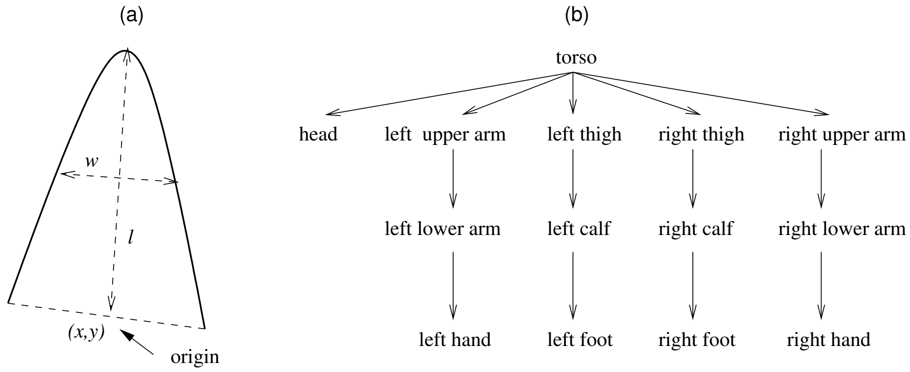

Figure 2.6: (a) Body part model (b) “connect-to” hierarchy

2.4.1 TRS-Invariant Body Model

Each human body model consists of six main body parts (the head, the torso, two arms and two legs) and twelve subparts of the four limbs. The body parts are constrained to connect to each other at the joints. The body parts are modeled with ribbons [1] [114, 115] as shown in Fig. 2.6(a). The width (w) of a ribbon is defined as the average width along the ribbon, and its length (l) and major axis come from the ribbon spine. The aspect ratio $a = w/l$ is invariant under similarity transforms, and it captures the global shape of a ribbon while ignoring small local shape deformations. Thus the aspect ratio is appropriate for the purpose of recognition.

However, the aspect ratio is too ambiguous to be used alone to distinguish different parts. For example, the head and the torso have similar aspect ratios. Therefore, besides the aspect ratios of the body parts, the geometric relationships between them are also modeled. To do so a local coordinate frame is associated to each body part. The origin of each part frame is located at the joint connecting the part to its parent in the “connect-to” hierarchy as shown in Fig. 2.6(b), except that the torso frame origin is located at its geometric center. The advantage of locating the origin at the joint instead of the geometric center of a body

[1] Following Rosenfeld’s definition [114], a “ribbon-like” planar shape is defined by specifying an arc, called the spine or axis, and a geometric figure such as a disk or line segment, called the generator, that “sweeps out” the shape by moving along the spine, changing size as it moves.

part is that the location of the origin becomes invariant to the body part's orientation. In summary, a body part is parameterized with a vector (a, l, x, y, θ), where $a = w/l$ is the aspect ratio that captures the general shape of a ribbon, (x, y) are the coordinates of the origin in the coordinate frame of the parent part, and θ is the intersection angle between the major axes of this part and its parent part.

Assuming that the human body model consists of m body parts that are parameterized with vectors $f_1, f_2, ..., f_m$, where $f_i = (a_i, l_i, x_i, y_i, \theta_i)$. Then the human body model is parameterized with four model matrices: the aspect ratio vector $A = \{a_1, ..., a_m\}$, the length ratio matrix $S = \{s_{ij}\}, i, j = 1, ..., m$, where $s_{ij} = l_i/l_j$, the relative position vector $X = (x_1, y_1, ..., x_m, y_m)$, and the orientation or posture vector $\Theta = \{\theta_1, ..., \theta_m\}$. Obviously, A and S are TRS-invariant.

The coordinates (x_i, y_i) of a subpart f_i (for example the lower arm) in the coordinate frame of its parent f_j (the upper arm) is simply:

$$(x_i, y_i) = (0, l_j). \tag{2.3}$$

Because the lengths of the body parts are constrained by the length ratio matrix S, only the relative positions of the six main body parts needs to be modeled. Let $X = (x_1, y_1, ..., x_6, y_6)$. To make this vector TRS-invariant, the coordinates of the joints are represented in a normalized torso coordinate system with the length of the torso normalized to be 1. Then the TRS-invariant relative positions of the six main body parts are $U = (0, 0, u_2, v_2, ..., u_6, v_6)$, where $(u_i, v_i) = ((x_i, y_i) - (x_1, y_1))/l_1$, (x_1, y_1) are the coordinates of the torso's center, and l_1 is the length of the torso.

Currently, the human model is parameterized with three TRS-invariant matrices: A, S, U, which constrain the aspect ratios, the relative sizes and positions of the body parts, respectively. In the future, the orientation constraints Θ will be incorporated to form stronger constraints on the appearance of a person.

2.4.2 TRS-Invariant Probabilistic Model

No two people are exactly alike due to difference in age, sex, racial membership, and the great range of diversity among individuals poses a problem for vision-based human detection. In this thesis, probability distributions are employed to accommodate shape variation among people and they can also absorb some of the shape variation due to changes of viewpoint. For simplicity, the variations of the model parameters are assumed to be subject to the Gaussian distributions. To simplify the calculation of the joint Gaussian distributions, and the variations of the aspect ratios, the length ratios, and the relative positions of the main body parts in the normalized torso coordinate system are assumed to be statistically independent of each other. Based on the above assumptions, the probability distributions of the three model matrices A, S, U can be estimated separately:

$$A \sim \mathcal{N}(\bar{A}, \Sigma_A) \tag{2.4}$$

$$S \sim \mathcal{N}(\bar{S}, \Sigma_S) \tag{2.5}$$

$$U \sim \mathcal{N}(\bar{U}, \Sigma_U). \tag{2.6}$$

Assuming that the variations of the aspect ratios of the body parts are statistically independent [2], then

$$N(\bar{A}, \Sigma_A) = \prod_i N(\bar{a}_i; \sigma^2_{a_i}) \tag{2.7}$$

Assuming that the variations of the ratios of the lengths between the body parts and a reference body part f_i are independent [3], then

$$N(\bar{S}_i, \Sigma_{S_i}) = \prod_{j \neq i} N(\bar{s}_{ji}; \sigma^2_{s_{ji}}) \tag{2.8}$$

[2]Based on the statistical data provided by Tilley [116], the aspect ratios of the body parts are not necessarily correlated. For instance, a small woman may have a round or thin face and a large or small hip, and so on.

[3]This assumption does not mean that the length of a body part is independent of the lengths of other body parts. In fact, the length ratio matrix S constrains the lengths of the body parts to be proportional. For instance, the mean ratio between the lengths of the left arm and the right arm is 1, and this constrains the lengths of the left arm and the right arm to be equal.

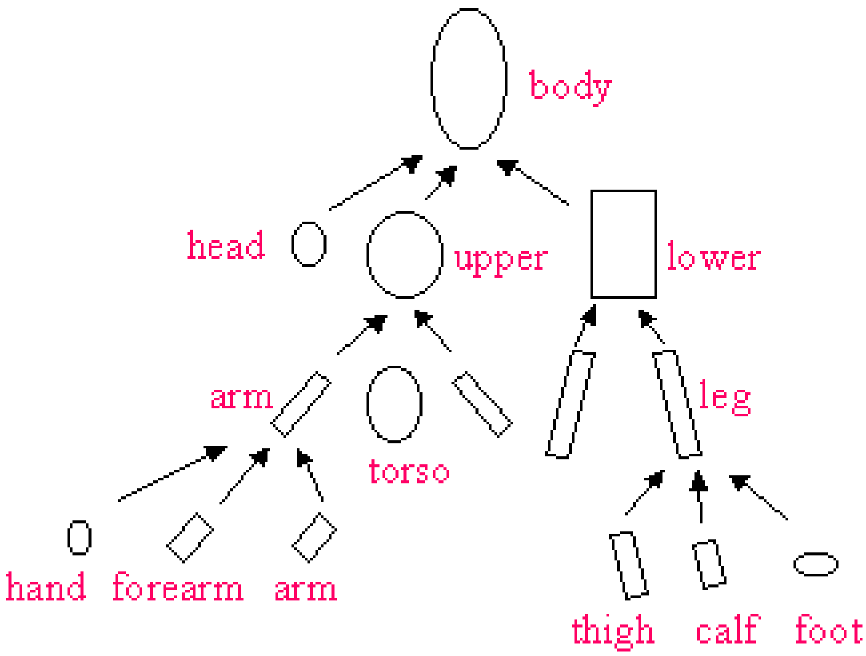

Figure 2.7: Generating merged body parts

where $S_i = \{s_{ji}\}$, $s_{ji} = l_j/l_i$, $j = 1, ..., m$, $j \neq i$. Thus, S_i is used to constrain the relative lengths of the body parts with respect to the length l_i of a reference part f_i.

The above probability distributions provide metrics to evaluate the shape, size relationship, and configuration similarities between the detected contour and the human body model. Their parameters (means and covariances) are estimated from the measurements provided by Tilley[116] (See Appendix).

2.5 Dressed Human Modeling

In contrast with a model of a particular object, a model of an object class should represent the generic shapes of the objects belonging to the class well and emphasize shape differences between classes, while the shape variations allowed within classes should not influence the description. Modeling human shapes is a challenging problem considering

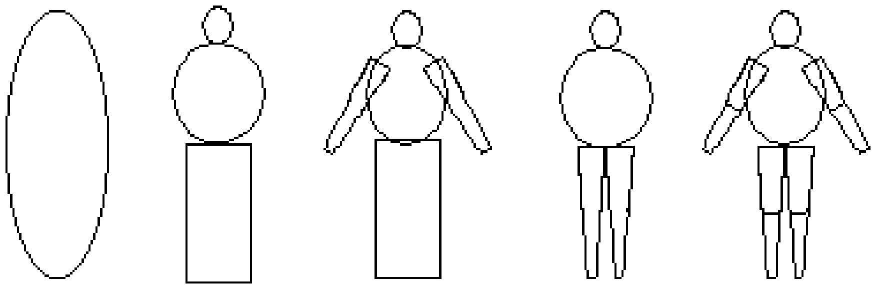

Figure 2.8: Assembling the human models

the high degree of shape variation as illustrated in Chapter 1 due to articulation, occlusion, and clothing. Articulation and occlusion can be handled by part-based representation very effectively, however, shape variation from clothing is much more difficult to deal with.

As shown in Fig. 1.3 clothes may drastically change the appearance of a person. One of the effects of clothes is that they cover some body parts and merge them into a single component. This makes it difficult to identify the covered parts and as a result we cannot recognize a dressed person. To handle this effect, merged body parts are introduced to model the merging of multiple body parts covered by clothes as shown in Fig. 2.7. The merged body parts are generated in a hierarchical manner as shown in Fig. 2.7 by computing the union of its subparts (the subparts are assembled based on the "connect-to" hierarchy shown in Fig. 2.6(b)). This hierarchy provides a compact representation of the coarse-to-fine modeling of human shapes (see some examples given in Fig. 2.8), and the generated models are called the dressed human models. The hierarchical organization of parts resembles the one proposed by Marr and Nishihara [17], but it includes more levels between the whole body and the major body parts to include the merged body parts. Thus, this hierarchy has more shape modeling power as demonstrated in Fig. 2.8 (it not only models the real body parts, but also the parts generated due to clothing and resolution degeneration), while sharing the same advantage as a normal hierarchical representation — stability and sensitivity (see section 2.2 for explanation).

The merged parts are also modeled with ribbons connected with each other at joints,

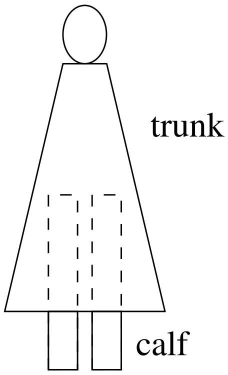

Figure 2.9: Adjustable length of the trunk

and the TRS-invariant probabilistic representation explained in subsection 2.4 is used to encode the shapes of the merged parts and their relationships. The coordinates of the origin of each part are represented in the local coordinate systems of both the upper part in the "connect-to" hierarchy (see Fig. 2.6(b)) and the parent part in Fig. 2.7. There are two kinds of relationships between the parts: one is "connect-to"; another is "part-of". Therefore, the location of a part can be inferred from two sources: either from the connected parts or from the merged parts covering it.

One special merged body part called the *trunk* (see Fig. 2.9) is introduced to cover the torso and some of the other body parts, and it occupies the same position as the torso in Fig. 2.7. The length and the width of the trunk is adjustable so that it can handle different self occlusion situations and dresses of various lengths (as shown in Fig. 1.3). Thus, the length constraint is not on the length of the trunk but on the sum of the lengths of the trunk and the attached legs. For example in Fig. 2.9, the sum of the lengths of the trunk and the calf should equal the difference between the lengths of the whole body and the head.

The deformations of the merged parts are represented by normal distributions. The parameters of the distributions (means and covariances) are learned from a set of training data collected from some web-sites and catalogs. The training data are first decomposed using the approach presented in subsection 2.3. The decomposed subparts are labeled manually and transformed into the TRS-invariant representations. Parts with the same label are grouped into the same subset to estimate the parameters of the corresponding model part (see Appendix).

In summary, this chapter presents a TRS-invariant dressed human model that can be used to detect people in an image independent of their sizes, poses, articulation, and clothing. Part-based representation is used to model the occluding contour of a person; merged parts are introduced to represent the merging of multiple real body parts. Furthermore, all parts are organized in a hierarchy to facilitate coarse-to-fine shape decomposition and classification. The next chapter will explain how to assemble a dressed human model dynamically to match an extracted contour and how to evaluate the resemblance between the assembled model and the contour to guide body part identification and human detection.

Chapter 3

Bayesian Similarity Measure

3.1 Requirements for a Good Shape Similarity Measure

A shape similarity measure designed for shape classification is useful in many applications such as object detection, image retrieval at the class level, and hierarchical database construction. Here an object's shape refers to its silhouette. Although a large number of shape similarity measures have been proposed, most of them are for the purpose of image retrieval and object recognition at the individual object level. This chapter focuses on designing a shape similarity measure for deformable shape classification, especially for articulated objects such as humans. A good shape similarity measure should satisfy the following requirements:

1. It should give large similarity measurements within the class while giving small ones between classes.

2. It should not depend on the position, size, and orientation of an object.

3. It should support articulation and partial occlusion.

4. It should be robust to noise, deformation, and blur resulting from image digitization and poor segmentation.

5. It should be efficient to compute.

This chapter presents a Bayesian similarity measure that satisfies all the above requirements. This similarity measure involves accumulation of the similarity between matched primitives. Because the comparison is between a shape and a shape class, the probability distributions described in Chapter 2 are used to represent the variations allowed within the class, and a Bayesian similarity measure is derived from the human model to satisfy the first requirement. Second, body parts are employed as primitives for shape representation and comparison. Part-based shape comparison contributes in a significant way to the fact that the similarity measure satisfies requirements 3 and 5. According to Siddiqi *et al.* [51], part-based representations allow for robust object recognition and play an important role in the theories of object categorization and classification. There is also strong evidence for part-based representations in human vision [51]. Furthermore, there are many fewer natural parts in a shape than there are other primitives such as points, line segments, convex arcs, etc. So a part-based representation enables fast shape matching. Third, the TRS-invariant shape representation described in Section 2.4 facilitates requirement 2. Fourth, the body parts are organized in a coarse-to-fine hierarchy and body part matching is performed in a coarse-to-fine manner to reduce the effects of noise and contour over-segmentation and to enhance fast shape matching.

The organization of the rest of the chapter is as follows: Section 3.2 discusses the related work on shape similarity measure. Section 3.3 presents the definition of the Bayesian similarity measure and illustrates how to calculate it based on the best match between the parts of a contour and a human model guided by coarse-to-fine model assembling and a goodness function. Section 3.4 demonstrates that the Bayesian similarity measure can be used to distinguish humans from other objects and therefore can be used for human detection. A discussion of the decision rule for human detection is given in Section 3.5.

3.2 Related Work on Shape Similarity Measure

The problem of determining the similarity of two shapes has been well-studied in several fields. The design of a similarity measure depends on how a shape is represented. Many global shape descriptors (see reviews in [96, 97]) such as Fourier Transform, moments, and eigen shapes have been used to compare two shapes, but they cannot handle occlusion and local deformation such as articulations very well. Therefore, this section does not discuss the similarity measures based on global shape but concentrates instead on those that use local shape primitives, such as points, key points, lines, arcs, axes, or parts. Broad overviews of shape similarity measures can be found in [98, 99, 101].

A point-based similarity measure such as the Hausdorff distance is commonly used to compare two shapes, but it is very sensitive to noise and occlusion. A similar measure that is not as sensitive is the partial Hausdorff distance [100]. This measure can deal with occlusion and clutter very effectively; the measure itself is used to guide the search for an alignment transform in the discrete space. Technically speaking, different transformations such as similarity, affine, or non-rigid transforms can be used for shape alignment. However, the dimension of a non-rigid transformation space is too high to be searched efficiently. A match-based method has been proposed to avoid the search for the transformation in a high dimensional space. An alignment transform is calculated from the matched points, then the similarity measure is calculated as the sum of the residual distances between the corresponding primitives. The common drawback with the above measures is that they have to transform one shape to another before shape comparison because the distance metric is not invariant under similarity transform.

Various cost functions have been proposed to evaluate the dissimilarity between two contours without aligning them. A cost function weights the similarity of the matched points on the basis of their local properties, such as the difference in the tangent or curvature of the contours at those points. The cost function itself is used to guide the search for the best match. Basri *et al.* [101] defined the cost function as "elastic energy" needed to deform (stretch or bend) one curve to another. However, the computation of elastic energy (which is defined in terms of curvature) is very sensitive to noise. Other alternatives also exist such as turning functions [104], arch height functions [102], size functions [103], or functions combining multiple local properties [105, 106]. Given a choice of cost functions, several methods such as dynamic programming, gradient descent, or the shortest path algorithm have been employed to find the correspondence between contours that minimize the cost. The main drawback of these methods is their high computational complexity due to searching for correspondences at the point level. Furthermore, none of these cost functions is invariant under scaling and/or rotation of the point data.

Other features such as key points and lines have been used to reduce the computational cost because a digital contour usually consists of much fewer features than of points. Pope and Lowe [109] modeled an object with a graph whose nodes represent the feature values and whose edges represent the spatial arrangement (symmetric, parallel) of the features. Objects are considered similar if their graphs are isomorphic; a similarity metric based on a probability density estimator is used to identify if a shape is an instance of a modeled object. To handle occlusion, partial matching is allowed and the largest mutually compatible matches are found by constructing an association graph to search for the maximal clique [110]. The main drawback of these methods is that they cannot handle articulated motion because the spatial relationships between features are assumed to be fixed. Moreover, it is difficult to find a coherent set of features that is shared by all possible shapes in a class and that can be extracted reliably.

Part-based representations have been proven to handle articulation and occlusion effectively. They have several advantages over other representations such as points, lines, and arcs. First, articulation usually happens at part boundaries, thus, a part-based representation is a more natural and coherent description of articulated shapes. Second, a shape contains fewer aggregate parts than other features. Third, part-based methods find strong support from human vision [51]. The main concerns of a part-based similarity measure are how to decompose a shape into stable parts and how to set up correspondence among them. Parts generally are defined to be convex or nearly convex shapes separated from the rest of the object at concavity extrema [49], or at inflections [42]. One type of approach is to represent shapes as skeletons or graphs and then to use graph matching or qualitative properties such as topology to compare shapes. The main drawback of these approaches to part-based shape analysis is that the shape decomposition is not stable. Since only qualitative properties are used for shape classification, they cannot distinguish two shapes with the same body part structure but different body part shapes and geometric relationships. Zhu and Yuille [23] developed a similarity measure to compare silhouettes based on both the local shapes of parts and the topology but the method can not handle shape degeneration or resolution changes very well. Several curve evolution approaches [50, 54, 55] have been proposed to model shapes of an object at different scales, but the related similarity measure is sensitive to occlusion and is not invariant under scaling.

Leung *et al.* [27, 28, 29] have proposed a method which combines the intensity pattern and the spatial relationships between the facial features to detect faces from the cluttered environment. However, they do not use the spatial relationship to help detect face features, and no size relationship and recursive procedure is involved in face detection. The main reason is that the facial features have very distinctive patterns and can be detected based on their intensity patterns. In human detection, we rely heavily on the spatial and size relationships to identify the human body parts, because the body parts such as arms and legs do not present very distinctive texture patterns.

In summary, none of the above similarity measures satisfies all the requirements listed in Section 3.1. Point-based approaches are time consuming, while feature-based approaches are not stable and can not handle articulation appropriately. In contrast, the part-based approach is a more promising direction, however, current methods can not handle shape decomposition errors and shape degeneration very effectively. Above all, the above shape similarity measures can not deal with large shape variations within a class. They are not appropriate for the purpose of classifying shapes such as those of humans.

3.3 Bayesian Similarity Measure and Body Part Identification

A part-based similarity measure evaluates the resemblance between a contour and a model based on the best match between their body parts. This section presents the definition of a Bayesian Similarity Measure derived from the probabilistic human model, and illustrates how to use this measure to guide the matching between a set of decomposed ribbons and a set of model body parts through coarse-to-fine model assembling.

3.3.1 Problem Formulation

The body part identification problem is to match a set of ribbons decomposed from a contour against a set of model body parts described in Chapter 2. This problem is formulated as an optimal hypothesis selection problem. Assume that n ribbons are extracted from a contour $C = \{c_1, c_2, ..., c_n\}$, and that the human model consists of m model body parts: $F = \{f_1, f_2, ..., f_m\}$. Let $H = (h_1, h_2, ..., h_m, view)$ represent a match hypothesis, where $view \in \{front/back, side\}$ and

$$h_i = \begin{cases} j, & \text{if } c_j \text{ corresponds to } f_i \\ 0, & \text{if no ribbon corresponds to } f_i \end{cases}$$

This is not a one to one mapping; it allows some body parts to be occluded and also allows some ribbons to not correspond to any body parts. The maximum *a posteriori* (MAP) hypothesis H^* is selected from the hypothesis space $\mathcal{H}$ such that, given a person is present in the image,

$$H^* = arg \max_H \; P(H, C|person).$$

According to Bayes' rule,

$$\begin{aligned} H^* &= arg \max_H \; P(C|H, person)P(H|person) && (3.1) \\ &= arg \max_H \; P(C|H, person)P(person|H)\frac{P(H)}{P(person)}. && (3.2) \end{aligned}$$

Assume that all hypotheses have the same prior, then $P(H)/P(person)$ is a constant and can be dropped from (3.2). Thus, the best hypothesis can be selected as

$$H^* = arg \max_{H} \ P(C|H, person)P(person|H). \tag{3.3}$$

Accordingly, the *goodness function* that rates the hypothesis is defined as follows:

$$G(H) \triangleq P(C|H, person)P(person|H) \tag{3.4}$$

$P(C|H, person)$ evaluates the degree of resemblance between the matched pairs, while $P(person|H)$ is proportional to the number of identified body parts — the more body parts being identified, the more likely the extracted contour is a person. Thus, the MAP hypothesis H^* selected using (3.3) maximizes the resemblance between the matched pairs and the number of identified body parts. Consequently, the similarity measure that evaluates the resemblance between the contour C and the human model is defined as

$$BSM(C) \triangleq G(H^*), \tag{3.5}$$

where H^* is the best hypothesis selected using (3.3).

3.3.2 Estimation of the Goodness Function

In order to calculate the goodness function $G(H)$, two terms need to be estimated: the likelihood $P(C|H, person)$ and the posterior probability $P(person|H)$.

Let $\hat{A}', \hat{S}_i', \hat{U}'$ be the aspect ratios, relative sizes, and relative positions of the identified body parts, respectively. They are estimated with the parameters of the corresponding ribbons, where $\hat{S}_i' = \hat{s}_{ji}, j \neq i$. The reference body part f_i is selected so that its aspect ratio has the highest probability: $N(\hat{a}_i; \bar{a}_i, \sigma_{a_i}^2) > N(\hat{a}_j; \bar{a}_j, \sigma_{a_j}^2), \forall j \neq i$. Let $(\bar{A}', \Sigma_A')$, $(\bar{S}_i', \Sigma_{S_i'})$, $(\bar{U}', \Sigma_U')$ be the expectations and covariances of the model parameter matrices assembled from the identified model parts. Because the matrices A', S', U' are assumed to be mutually statistically independent, the likelihood $P(C|H, person)$ can be estimated as

$$P(C|H, person) = N(\hat{A}'|\bar{A}', \Sigma_A')N(\hat{S}_i'|\bar{S}_i', \Sigma_{S_i'})N(\hat{U}|\bar{U}, \Sigma_U) \tag{3.6}$$

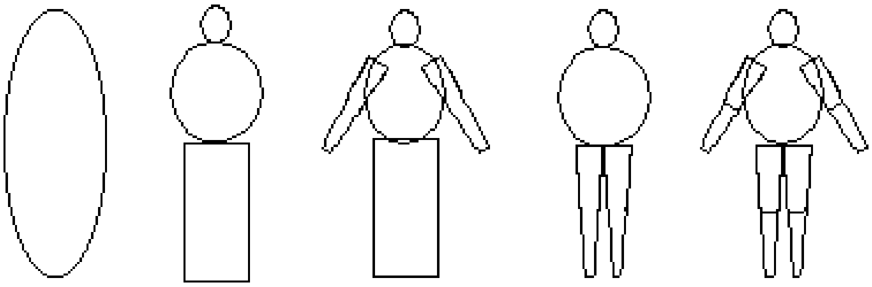

Figure 3.1: Examples of the calculated $P(person|H)$ given the degeneration factor $\alpha = .9$ and the contour parts matching the model parts exactly

The likelihood $P(C|H, person)$ derived above constrains the best match between the set of decomposed ribbons and the model body parts to be of similar relative sizes and positions as well as similar aspect ratios. Note that the model parameter matrices $(\bar{A}', \Sigma'_A)$, $(\bar{S}'_i, \Sigma_{S'_i})$, $(\bar{U}', \Sigma'_U)$ are not fixed but are dynamically assembled based on the match hypothesis. This means that the matching procedure selects not only the best match pairs but also the best human model to describe the extracted contour C.

The conditional probability $P(person|H)$ is estimated based on the number and types of the identified body parts. The more body parts being identified, the more likely the extracted contour is a person. And the presence of a fine level body part indicates a higher likelihood of the presence of a person than that of a coarse level body part does. According to the above observation, the following formula is designed to estimate $P(person|H)$:

$$P(person|H) = \sum d_i w_i \tag{3.7}$$

where $d_i = 0$, if $h_i = 0$ (the model body part f_i is not identified), and $d_i = 1$, if $h_i \neq 0$. w_i is a weight used to evaluate the contribution of the identification of the body part f_i to the presence of a person, and it is estimated as $w_i = n_i/n$, where $n = \sum n_j, \forall f_j$ that has no subpart, and $n_i = 1$ if f_i does not have subparts (see Fig. 2.7), except for the head and the torso. Although the head and the torso do not have subparts, their appearance imply a high likelihood of the presence of a person than the appearance of other body parts such as two arms. Therefore, n_i is set so that $P(person|H) > 0.5$ when $d_i \neq 0$

and f_i is a head or a torso. For the body parts who have subparts, n_i is defined recursively:

$$n_i = \alpha \sum n_j, \forall f_j \text{ being the subpart of} f_i, \tag{3.8}$$

where $\alpha < 1$ is a *degeneration factor* used to punish the ambiguity with the coarse level body parts so that a contour that only matches a low resolution model well does not get a high similarity measure (see Fig. 3.1). Thus, the degeneration factor helps to reduce the false alarms caused by the generality of the human model.

In summary, the following function is designed to evaluate the goodness of a hypothesis H:

$$G(H) = N(\hat{A}'|\bar{A}', \Sigma_{A'})N(\hat{S}_i'|\bar{S}_i', \Sigma_{S_i'})N(\hat{U}'|\bar{U}', \Sigma_{U'})P(person|H) \tag{3.9}$$

Unlike previous work [23, 58] that identifies each body part independently, the goodness function proposed in this thesis couples the local and global constraints to guarantee the best match between the extracted ribbons and the human body parts to be of consistent global spatial and size relationships as well as having consistent local shapes.

3.3.3 Selecting the Optimal Hypothesis through Dynamic Model Assembling

Most previous methods [78, 79, 86] detect body parts sequentially without backtracking. The disadvantage is that if one of the body parts is located incorrectly, the locations of the rest of the body parts would be wrong. In contrast, in this thesis work every extracted ribbon is considered as a candidate for every body part. The exhaustive search for the best hypothesis H^* is to check every possible hypothesis and find the best one based on the goodness function $G(H)$ given in Eq. (3.9). However, the hypothesis space is generally very large due to combinatorial explosion. For example, if there are m model body parts and n extracted ribbons, then the number of match hypotheses would be: $\binom{m}{k}\binom{n}{k}k! + \binom{m}{k-1}\binom{n}{k-1}(k-1)! + ... + \binom{m}{1}\binom{n}{1}$, where $k = \min(m, n)$.

This section presents a coarse-to-fine search procedure to identify and locate the body parts. First, derive a coarse-level decomposition of the contour C (as shown in Fig. 3.2(b))

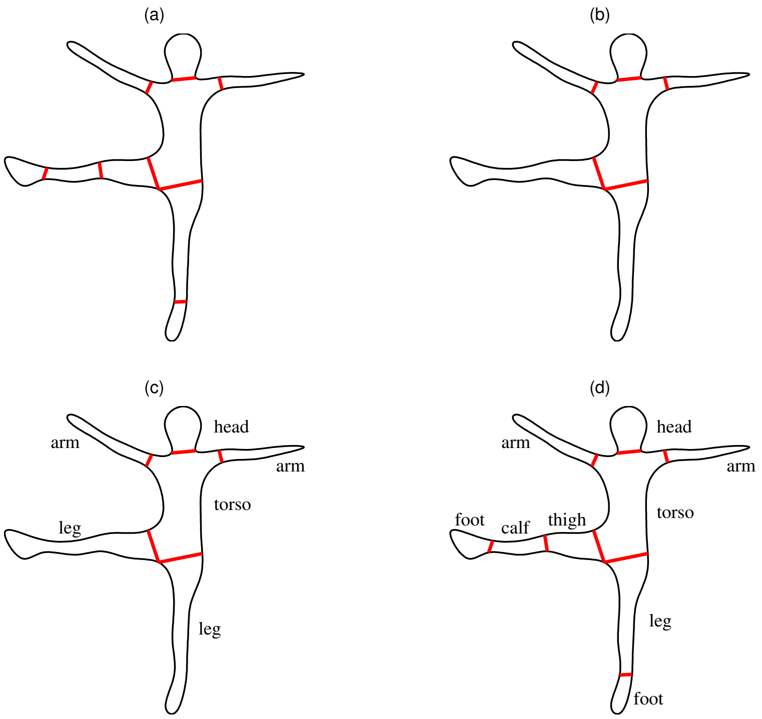

Figure 3.2: Body part identification: (a) fine-level contour decomposition (b) coarse-level contour decomposition (c) coarse-level part identification (d) fine-level part identification

be grouping the ribbons whose major axes sharing an endpoint and their widths are similar (their width ratio is within the range [.5, 1.5]). Second, identify the coarse-level body parts based on the goodness function (as shown in Fig. 3.2(c)). Third, if a part at a coarse-level is decomposed into subparts, identify its subparts (as shown in Fig. 3.2(d)). Fourth, derive the locations of the subparts of the parts not decomposed by the contour decompostion procedure. This step will be explained in the body part updating procedure in Section 4.3. Because there are fewer body parts at a coarse-level than at a fine-level, the hypothesis space is significantly reduced. To guarantee that the procedure converges to the global optimal solution, a coarse level hypothesis will result in a lower posterior $P(person|H)$ than a fine-level hypothesis will, because of the degeneration factor (α) used in Eq. (3.8). Therefore, by combining the constraints on the aspect ratios, relative sizes and positions, and the posterior $P(person|H)$, the goodness function proposed in this thesis selects the

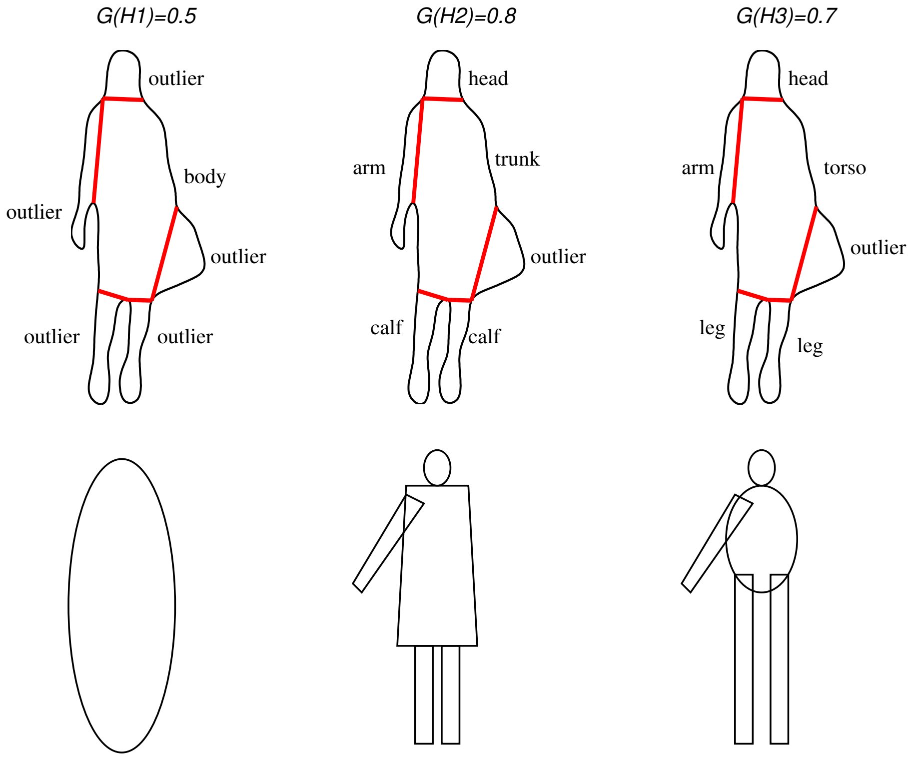

Figure 3.3: Coarse-to-fine hypothesis selection and dynamic model assembling (hypothesis H_2 gets the highest score and is selected as the best hypothesis to match the decomposed parts against the model body parts.)

human model and model parts at the right resolution to label the decomposed contour segments as shown in Fig. 3.3.

The efficiency of the search procedure can be further improved by selecting the candidates of each model body part through gating on the aspect ratios of the decomposed ribbons:

$$D_i = \{c_j | \; c_j \in C, N(\hat{a}_j; \bar{a}_i, \sigma_i^2) > \lambda\} \tag{3.10}$$

where D_i is the set of candidates for the model body part f_i, c_j is a ribbon decomposed from C, a_j is the aspect ratio of c_j, $\bar{a}_i$ and σ_i are the mean and the standard deviation of the aspect ratio of the model part f_i, and λ is the threshold which should not be set too high to achieve a global optimal hypothesis.

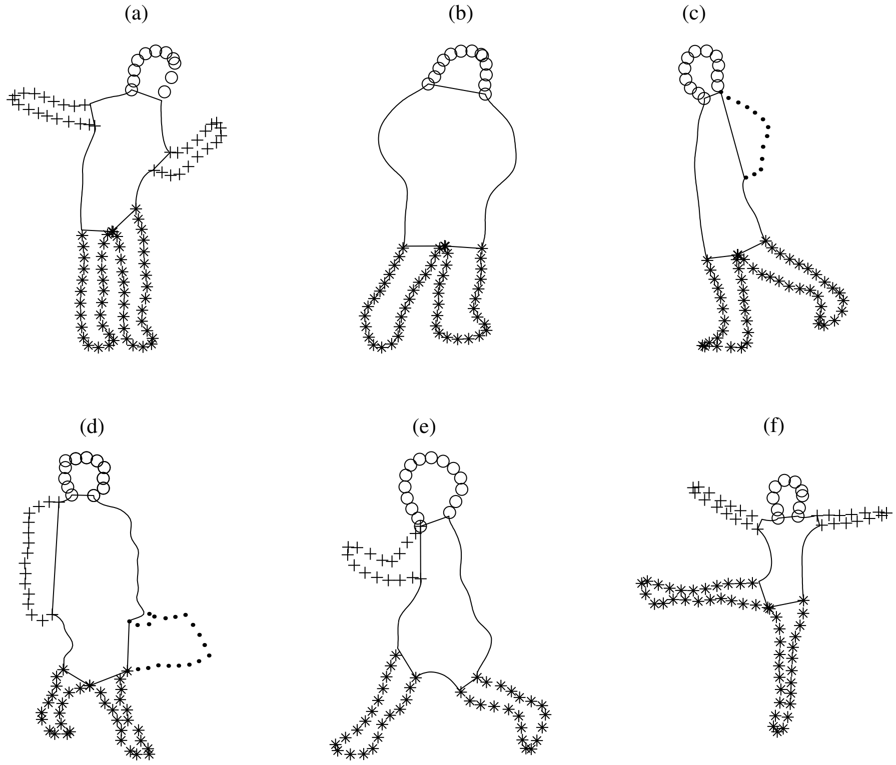

Figure 3.4: Identify the main body parts indicated by ooo (head) — (torso) +++ (arm) *** (leg)

3.3.4 Experimental Results

Figs. 3.4 and 3.5 show some results of the main and extended body part identification, respectively. Here the contour is first decomposed into natural body parts using the algorithm described in Section 3.3. Then the decomposed parts are matched against the model body parts through the coarse-to-fine optimal hypothesis selection procedure described in subsection 3.3.3. The results demonstrate that the proposed algorithm can identify the human body parts at different resolution levels correctly. Even if some parts are occluded, it does not affect the identification of other parts. The algorithm can also detect the outliers simultaneously. The outliers may be due to the objects carried by a person, clothes, shadows or the background. Thus, through the body part identification procedure described in Section 3.3.3, certain contour extraction errors can be corrected. In Fig. 3.4(c), the arm is incor-

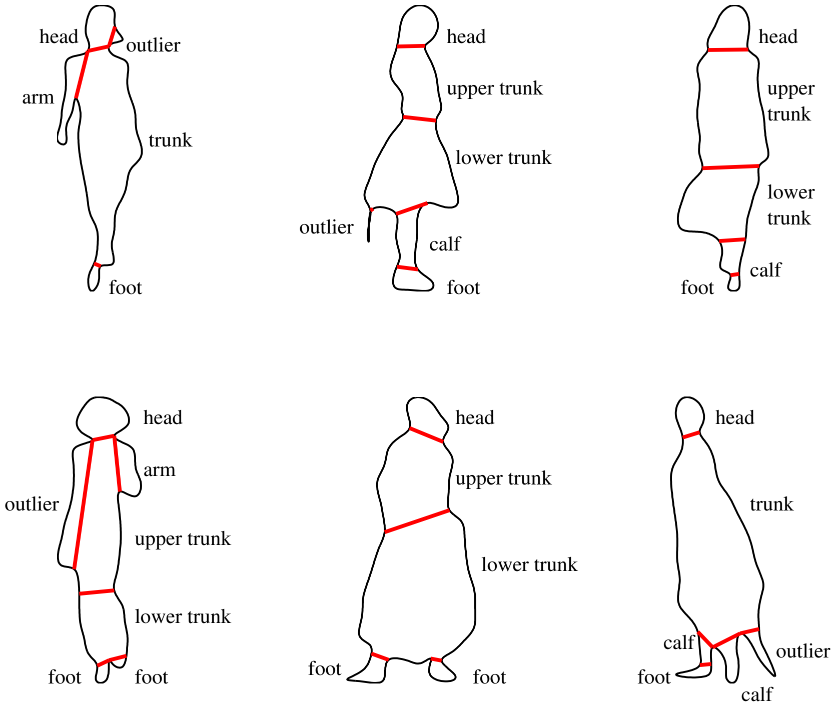

Figure 3.5: Identify the merged body parts

rectly labeled as an outlier, but as you will see in Chapter 4, through the recursive context reasoning algorithm the missed arm can be located correctly (as shown in Fig. 4.3(3)).

3.3.5 Multiple Hypotheses for Analyzing Multiple People

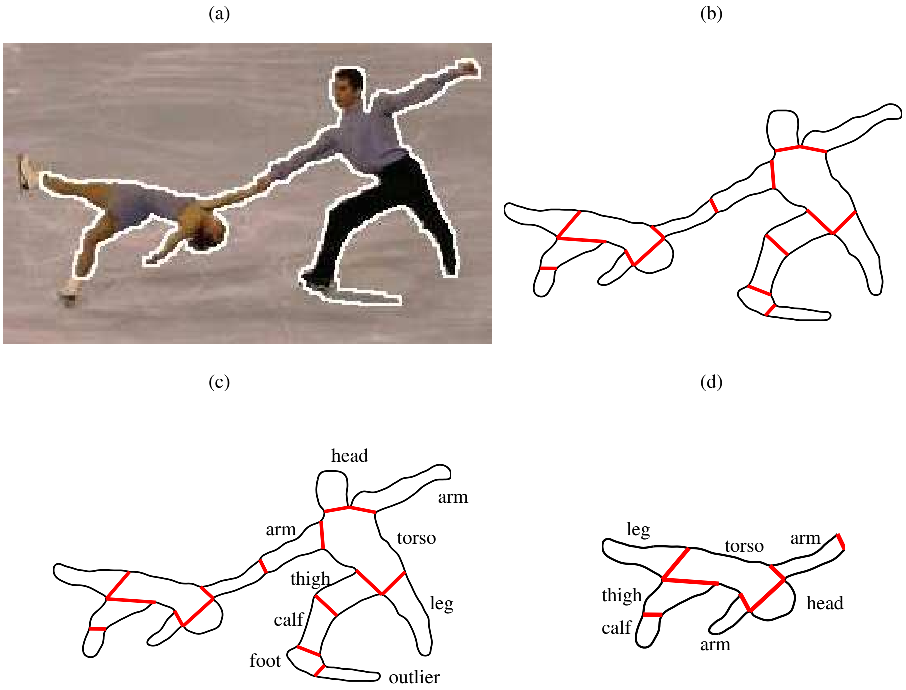

Figure 3.6: Locating the body parts of multiple people: (a) raw image (b) contour decomposition (c) the identified body parts of the first person (d) the identified body parts of the second person

When the extracted contour corresponds to a single person, the best hypothesis can be selected based on the goodness function given in Eq. (3.9) to label the decomposed ribbons. However, if the contour corresponds to multiple people as shown in Fig. 3.6(a), then multiple hypotheses need to be selected. Unlike previous work that depends on motion information [62] or upright-standing constraint [80] to separate overlapping people, this thesis proposes an iterative procedure to extract people one by one from the silhouette. First, select the best hypothesis to identify the body parts of a person contained in the contour. Second, remove the identified body parts from the contour. Third, go to the first step to analyze the remaining silhouette until no contour segment is left. In this procedure, it is possible simultaneously to label the contour and to separate people as shown in Fig. 3.6. Shadows can cause serious problem for people detection methods based on background subtraction [80]. However, using the goodness function proposed in this thesis, the shadow can be correctly identified as an outlier as shown in Fig. 3.6(c). This is because the shadow does not form a valid human body part.

3.4 Bayesian Similarity Measure for Human Detection

3.4.1 Decision Rule

The human detection problem is: given a detected contour C, determine whether this contour represents the silhouette of a person or not.

A simple decision rule to perform human detection is: the contour C corresponds to a person if

$$P(person|C) > threshold. \tag{3.11}$$

Otherwise, C is not a person. The threshold controls the receiver operating characteristics (ROC) of the detector. According to Bayes' rule,

$$P(person|C) = \frac{P(C|person)P(person)}{P(C)}. \tag{3.12}$$

Assume that all contours have the same prior, then $P(person)/P(C)$ can be considered as a constant, and the decision rule becomes: the contour C corresponds to a person if

$$P(C|person) > threshold. \tag{3.13}$$

The likelihood $P(C|person)$ can be calculated by conditioning on the hypotheses for body part identification:

$$P(C|person) = \sum_{H \in \mathcal{H}} P(C|H, person)P(H|person), \tag{3.14}$$

where $\mathcal{H}$ is the hypotheses space.

Because it is not efficient to explore all hypotheses in $\mathcal{H}$, a winner-take-all strategy is employed to approximate $P(C|person)$:

$$P(C|person) \approx P(C|H^*, person)P(H^*|person) \tag{3.15}$$

where H^* is the optimal hypothesis selected in the part identification procedure described in subsection 3.3.3. This approximation works well in the situation of low noise and unambiguous data. Because the recursive context reasoning algorithm explained in Chapter

4 will improve the accuracy of contour extraction iteratively, the above approximation will get more accurate accordingly.

By applying Bayes's rule, Eq. (3.15) becomes

$$P(C|person) \approx P(C|H^*, person)P(person|H^*)\frac{P(H^*)}{P(person)} \tag{3.16}$$

According to the assumption made in section 3.3.1, $P(H^*)/P(person)$ is a constant. Thus, the decision rule can be further simplified as: the contour C corresponds to a person if

$$P(C|H^*, person)P(person|H^*) > threshold \tag{3.17}$$

where the left hand side of Eq. (3.17) is just the similarity measure defined in Eq. (3.5). This means that the similarity measure proposed in this thesis can be used to perform human detection: the contour C corresponds to a person if

$$BSM(C) > threshold. \tag{3.18}$$

Note that the thresholds in (3.11), (3.13), (3.18) are different.

3.4.2 Experimental Results

Figs. 3.8 and 3.9 give some examples of the estimated degrees of similarity between the extracted contours and the human model using Eq. (3.5). The results demonstrate that the probabilistic similarity measure proposed in this thesis generates a higher similarity between the humans' silhouette and the human model than between the animals' silhouettes and the model, although the silhouettes are in different resolutions, clothes and postures, and many animals have configurations similar to that of a human. This demonstrates that the proposed similarity measure can be used to distinguish humans from other objects.

3.5 Discussion

In summary, this chapter presents a probabilistic similarity measure that satisfies requirements 1 through 6. This is demonstrated by theoretical considerations and experimental

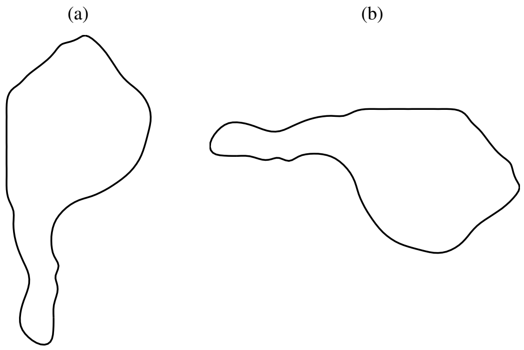

Figure 3.7: An ambiguous contour: the right figure is the same as the left one except for the different orientation

results. The estimation of the similarity measure is based on a coarse-to-fine body part identification procedure. Therefore, the proposed similarity measure can be employed to detect humans in various sizes, positions, orientations, clothes, postures, and partially occluded situations. Because the detected contours are not always complete and perfect, there are situations when the information provided by the detected contours is not enough to make the final decision. Fig. 3.7 gives an example of an ambiguous contour: the contour in Fig. 3.7(a) looks like a partially occluded person, but if rotated 90^o, it looks like a finger as shown in Fig. 3.7(b). To avoid false alarms, the decision is made only when the similarity measurement is sufficiently high or sufficiently low. Otherwise if the similarity measure falls within a defined uncertainty region, then go to find a more distinguished contour which will be described in the next chapter. Thus, the new decision rule becomes:

$$\begin{cases} C \text{ is a person}, & \text{if } BSM(C) > \lambda_2 \\ \text{no decision}, & \text{if } BSM(C) \in [\lambda_1, \lambda_2] \\ C \text{ is not a person}, & \text{if } BSM(C) < \lambda_1 \end{cases} \tag{3.19}$$

There are cognitive experiments [51] to prove that we humans do not immediately recognize those objects for which the figure/image provides an insufficient amount of information. The geometrical descriptions must exceed minimum level of richness or we will fail to recognize the object. The next chapter will show a recursive context reasoning algorithm that searches for more information in an image to support reliable human detection and body part localization.

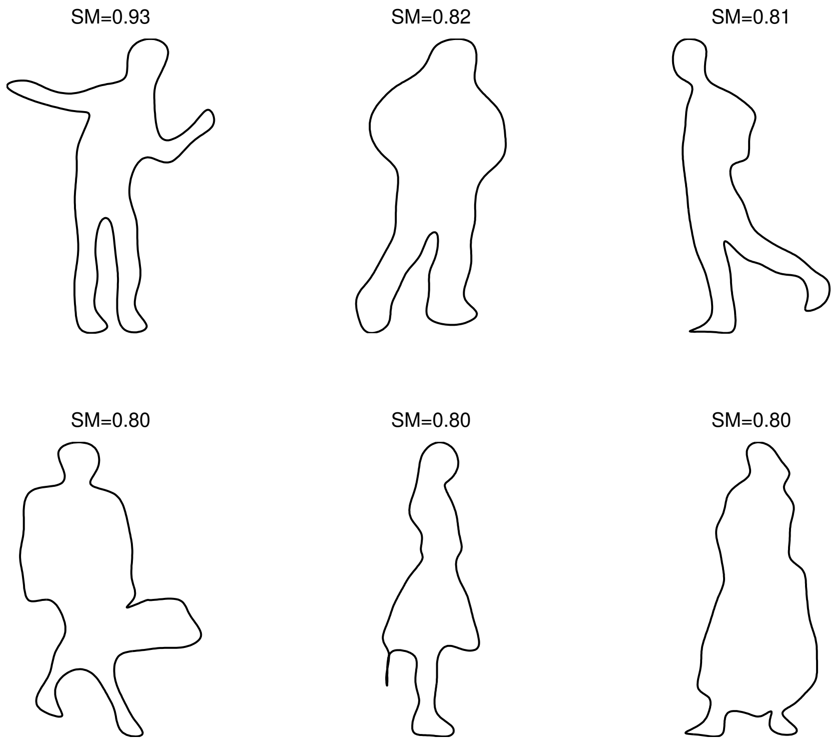

Figure 3.8: Similarity measures between the shapes and the human model

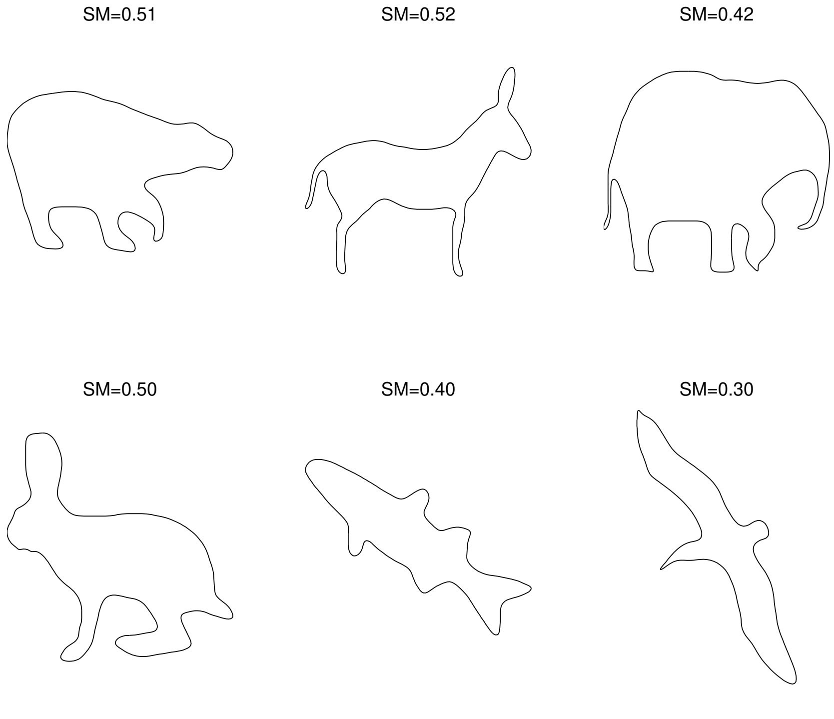

Figure 3.9: Similarity measures between other animals and the human model

Chapter 4

Recursive Context Reasoning

4.1 Why Do We Need Contextual Information ?

High performance object detection depends on reliable contour extraction, but contour extraction is an under-constrained problem without knowledge about the objects to be detected. The experimental results given in Chapters 2 and 3 also illustrate that higher level knowledge is needed to guide the search for more details of the objects in the image to resolve ambiguity with the initial extracted contour.

This thesis proposes a recursive context reasoning (RCR) algorithm to solve the above dilemma. A TRS-invariant probabilistic model is designed to encode the shapes of the body parts and the context information — the size and spatial relationships between body parts. A Bayesian framework is developed to perform human detection and part identification under partial occlusion. A contour updating procedure is introduced to integrate the human model and the identified body parts to predict the shapes and locations of the parts missed by the contour detector; the refined contours are used to reevaluate the Bayesian similarity measure and determine if the detected contour is a person or not. Therefore, contour extraction, body part localization, and human detection are improved by combining the context constraints from the identified body parts and the human model.

The combination of top-down and bottom-up paradigms to improve the performance of the vision system has been demonstrated in previous work. The ACRONYM system [1] developed by Brooks is an influential and very general framework for object recognition. Brooks used a general symbolic constraint solver to calculate bounds on the view point and model parameters from image measurements. The bounds were then used to check the consistency of all potential matches of ribbons to object components. However, the actual calculation of bounds for such general constraints was mathematically difficult and approximations had to be used that did not lead to exact solutions for viewpoints. The alignment method [6] proposed by Huttenlocher and Ullman utilized minimal sets of features which suffice to establish a unique transformation between a model and its hypothesized instance in the image. For each match a corresponding transformation is computed, and the set of model edges is transformed to the image to verify the candidate transformation. The time complexity of the approach is high, because an exhaustive enumeration is applied over all the possible pairings of minimal sets of model and image features. Lowe [2] used the initial matches between some model and image features to constrain the locations of other features of the model and thereby generate new matches that can be used to confirm or reject the initial match and to refine the pose estimate. But Lowe's approach cannot efficiently handle objects with a large number of variable parameters or articulated parts, because it requires an exact description of the object's geometry.

The context reasoning approach proposed in this thesis overcomes the limitations of the above work. The following sections will describe the algorithm in detail.

4.2 Outline of the RCR Algorithm

The outline of the RCR algorithm is as follows.

Step 1: Contour extraction: the algorithm is independent of how the initial contours are extracted from the image. Methods such as depth segmentation or background subtraction can be used for this purpose.

For each extracted contour, run Steps 2 to 7:

Step 2: Contour decomposition (described in Section 2.2): decompose the contour C into natural parts.

Step 3: Body part identification (described in Section 3.3): generate a hypothesis H^* to match the decomposed parts against the model body parts, s.t.

$$H^* = arg \max_H G(H), \tag{4.1}$$

where $G(H)$ is the goodness function calculated using Eq. (3.9).

Step 4: Human detection (described in Section 3.4): calculate the similarity measure

$$BSM(C) = G(H^*) \tag{4.2}$$

to determine if the contour C is a person or not based on the decision rule given in (3.19). If $BSM(C) \in [\lambda_1, \lambda_2]$, then go to Step 3 to search for more body parts to verify the hypothesis; if $BSM(C) > \lambda_2$, then declare that the contour C is not a person; if $BSM(C) < \lambda_1$, then declare that the contour C is not a person.

Step 5: Update the locations and outlines of the body parts from the identified body parts and the human model.

Step 6: Align the predicted outlines of the missed body parts to the edge features in the image.

Step 7: Recalculate the similarity measure and determine if a person is present or not.

Note that the RCR algorithm is different from the EM algorithm [111], a widely used approach to learning in the presence of unobserved variables. In the EM algorithm, if some variable is sometimes observable and sometimes not, then we can use the cases for which it has been observed to learn to predict its values when it is not. However, in the RCR algorithm we use the detected body parts and the geometrical relationships between the detected parts and undetected parts to predict the parameters of undetected parts.

Steps 2 to 4 have been described in Chapters 2 and 3, respectively. Thus, the rest of the chapter elaborates Steps 5 and 6, then gives some example runs of the RCR algorithm.

4.3 Update the Shapes and Locations of the Body Parts

The goal here is to demonstrate that the identified human body parts and the human model can be combined to refine the shapes and locations of the identified body parts and to predict the shapes and locations of the missed body parts.

4.3.1 Update the Parameters of the Identified Parts

The first step is to integrate the identified body parts and the human model to refine the locations of the identified body parts. This is performed using the weighted Least Squares method (LSM) [113]. Similar schemes have been employed by Hel-Or *et al.* [85, 86] to locate the articulated parts sequentially. This thesis extends the perfect object model employed in [85] to a probabilistic model and extends the work to update the shapes and sizes besides the positions of the body parts of an articulated object.

A body part is parameterized with a vector (a, l, x, y, θ) as described in Chapter 2 to represent the aspect ratio, length, position, and orientation of the body part. The Least Squares method provides a simple and efficient way to integrate the parameters estimated from the labeled contour segments and from the corresponding model body parts. Assume that the estimates for a parameter x are $z_1, z_2, ..., z_k$, and the uncertainties with these estimates are $v_1, v_2, ..., v_k$, respectively. Then $z_i = x + v$ and $Z = Hx + V$, where $Z = (z_1, z_2, ..., z_k)^T$, $V = (v_1, v_2, ..., v_k)^T$, and $H = (1, 1, ..., 1)^T_{k\times 1}$. The weighted least squares estimate of x is

$$\hat{x} = [H^T \Sigma_v^{-1} H]^{-1} H^T \Sigma_v^{-1} Z, \tag{4.3}$$

where Σ_v is the covariance of the random vector V.

Take updating the location of the joint between the left arm and the torso as an example. The joint is initially located at the middle point $P_1 : (\hat{x}_1, \hat{y}_1)$ of the cut between the arm and the torso (see Fig. 4.1(a)). The second estimate $P_2 : (\hat{x}_2, \hat{y}_2)$ (see Fig. 4.1(b)) is derived from the locations of the head $(\hat{x}_h, \hat{y}_h)$ and torso $(\hat{x}_t, \hat{y}_t)$. Let the coordinates of the head and the left arm in the model coordinate frame (the normalized torso coordinate frame) be $(\bar{u}_h, \bar{v}_h)$, $(\bar{u}_{la}, \bar{v}_{la})$, respectively. Assuming similarity transformation between the model

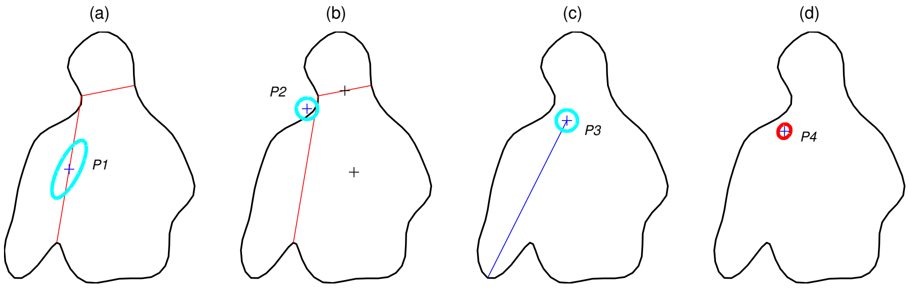

Figure 4.1: Updating the location of the joint between the arm and the torso: (a) the initial estimate based on body part identification, (b) the estimate based on the locations of the torso and head, (c) the estimate based on the major axis of the arm, and (d) the weighted least squares estimate integrating estimates P_1, P_2, and P_3.

and the image coordinate frames, then

$$(\hat{x}_2, \hat{y}_2)^T = sR(\bar{u}_{la}, \bar{v}_{la})^T + (\hat{x}_t, \hat{y}_t)^T, \tag{4.4}$$

where $s = \|(\hat{x}_h, \hat{y}_h) - (\hat{x}_t, \hat{y}_t)\|$, $R = \left(\begin{smallmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{smallmatrix}\right)$, and $\theta = arctan2(\hat{x}_h - \hat{x}_t, \hat{y}_h - \hat{y}_t)$.

The third estimate $P_3 : (\hat{x}_3, \hat{y}_3)$ (see Fig. 4.1(c)) is derived from the major axis of the left arm, the hand location $(\hat{x}_{hn}, \hat{y}_{hn})$ and the expected length of the arm $\hat{l}_a = \hat{l}_h\bar{r}$, where $\hat{l}_h$ is the length of the head, and $\bar{r}$ is the expected ratio of the lengths of the arm and the head. Assume that the orientation of the major axis of the left arm is θ. Then

$$(\hat{x}_3, \hat{y}_3) = (\hat{x}_{hn}, \hat{y}_{hn}) - \hat{l}_a(cos\theta, sin\theta). \tag{4.5}$$

Then the least squares estimate of the joint location (see Fig. 4.1(d)) can be calculated using Eq. (4.3). The aspect ratio and the length of the body part can be updated accordingly.

The remaining question is how to estimate the uncertainties associated with the measurements from the identified body parts. This is a nontrivial task. For example, the measurement of the location of a body part is contaminated by the errors with contour detection, contour decomposition, partial occlusion, and clothing (thus, it is reasonable to assume Gaussian noise, considering the numerous sources of noise). Among these factors, contour decomposition error is the main cause of the error in joint localization. If the boundaries between the body parts are not located accurately, then the joints cannot be located correctly. On the other hand, the aspect ratio of a body part is a good indication of the reliability of the joint location. Therefore, (1) when both ends of a cut are negative curvature minima, the error with contour decomposition is small, and the uncertainty of the joint location is assumed to be half of the contour extraction error; (2) when only one end of the cut is a negative curvature minimum, the uncertainty along the major axis of the body is assumed significantly larger than that along the minor axis, and the standard deviation along the major axis is estimated as $|l' - l|$, where $l' = w/\bar{a}$ is the expected length. The uncertainty along the minor axis is assumed to be the same as the contour extraction error.

As demonstrated in Figs. 4.2(c) and 4.3(a), the integrated estimation improves the accuracy of the locations and shapes of the body parts.

4.3.2 Predict the Parameters of the Missed Parts

The second step is to predict the outlines of the unidentified body parts from the human model and the identified body parts. This is done by estimating a body part's parameter vector $(\hat{a}_j, \hat{l}_j, \hat{x}_j, \hat{y}_j)$. Assume that the aspect ratios of the body parts are independent of each other, i.e., $\hat{a}_j|\hat{a}_i = \hat{a}_j$, the MAP estimation of a_j is simply its mean value $\hat{a}_j = \bar{a}_j$ and its variance is $\Sigma_{\hat{a}_j} = \Sigma_{\bar{a}_j}$.

The length of a body part $\hat{l}_j$ can be estimated from any of the identified body parts: $\hat{l}_j|\hat{l}_i = \bar{s}_{ji}\hat{l}_i$, $\Sigma_{\hat{l}_j|\hat{l}_i} \approx \bar{s}_{ji}^2\Sigma_{\hat{l}_i} + \hat{l}_i^2\Sigma_{\bar{s}_{ji}}$, where $\bar{s}_{ji} = \bar{l}_j/\bar{l}_i$. If more than one part have been identified, then the MAP estimate of $\hat{l}_j$ is the weighted summation:

$$\hat{l}_j = \Sigma_{\hat{l}_j} \sum_{i \in I} \frac{1}{\Sigma_{\hat{l}_j|\hat{l}_i}} \bar{s}_{ji}\hat{l}_i \tag{4.6}$$

$$\Sigma_{\hat{l}_j} = \frac{1}{\sum_{i \in I}(1/\Sigma_{\hat{l}_j|\hat{l}_i})} \tag{4.7}$$

where I is the set of index of the identified parts. For efficiency, the length $\hat{l}_j$ is approximated using the length of the body part with the least uncertainty (the smallest variance).

The location of the main body part f_j is estimated from the locations of the other main body parts. If more than two main body parts have been identified, then the transformation

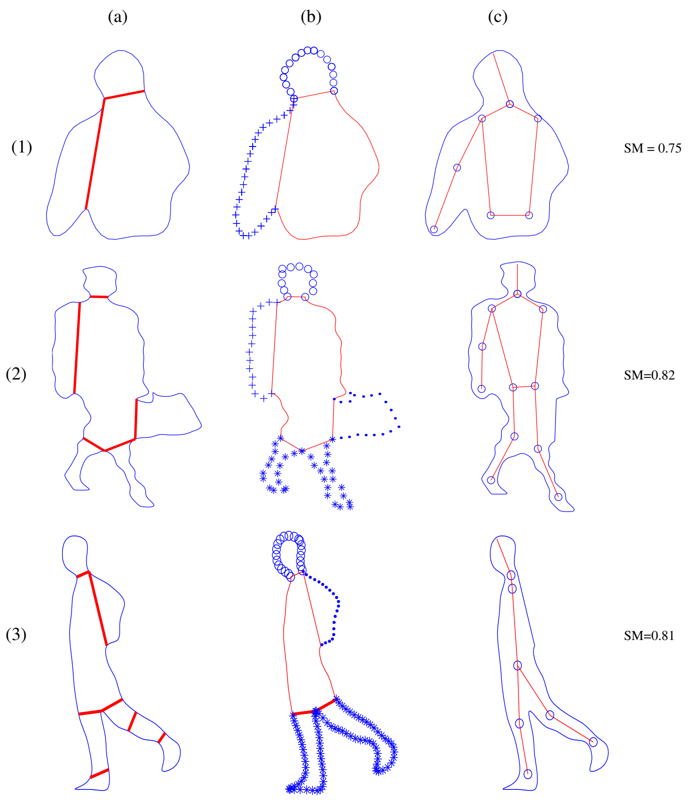

Figure 4.2: The first iteration: (a) contour partition (b) main body part identification indicated by ooo (head) — (torso) +++ (arm) *** (leg) (c) the updated locations of the identified body parts.

T that transforms the coordinates from the human model coordinate system to the image coordinate system can be estimated by the Least Squares Method. Let $\hat{X}_I = (p_1, p_2, ..., p_i)$ be the coordinates of the identified parts in the image coordinate system, and let $\bar{U}_I = (q_1, q_2, ..., q_i)$ be the coordinates of the identified parts in the model coordinate system, where $p_i = (\hat{x}_i, \hat{y}_i, 1)^T$, $q_i = (\bar{u}_i, \bar{v}_i, 1)^T$. Then

$$\hat{X}_I = T\bar{U}_I \tag{4.8}$$

$$\hat{T} = \hat{X}_I \bar{U}_I^T (\bar{U}_I \bar{U}_I^T)^{-1} \tag{4.9}$$

And the position of the unidentified body part f_j is estimated as

$$(\hat{x}_j, \hat{y}_j, 1)^T = \hat{T}(\bar{u}_j, \bar{v}_j, 1)^T. \tag{4.10}$$

Let $t = (t_x, t_y, \theta, s) \sim N(\bar{t}, \Sigma_t)$ be the translation, rotation angle, and the scaling involved in the transformation T. Assume that $\hat{X}_j = (\hat{x}_j, \hat{y}_j, 1)^T$ can be approximated by a first-order Taylor series expansion about the mean of t, then the uncertainty with $\hat{X}_j$ can be approximated as

$$\Sigma_{\hat{X}_j} \approx J_t \Sigma_t J_t^T + \hat{T}\bar{U}_j\hat{T}^T \tag{4.11}$$

where $\bar{U}_j = (\bar{u}_j, \bar{v}_j, 1)^T$, J_t is the Jacobian of the transformation evaluated at $\bar{t}$.

The location of a subpart in the image coordinate frame $(\hat{x}_j, \hat{y}_j)$ can be inferred from that of a part f_i directly connected to it using Eq. (4.12) or from that of the extended body part f_k that covers it using Eq. (4.13). Here a scaled orthographic camera model is assumed.

$$(\hat{x}_j, \hat{y}_j) = (\hat{x}_i, \hat{y}_i) + \hat{l}_i(cos\hat{\theta}_i, sin\hat{\theta}_i) \tag{4.12}$$

$$(\hat{x}_j, \hat{y}_j) = Rl'[(\bar{u}_j, \bar{v}_j) - (\bar{u}_k, \bar{v}_k)] + (\hat{x}_k, \hat{y}_k), \tag{4.13}$$

where $R = \begin{pmatrix} cos\theta_k & sin\theta_k \\ -sin\theta_k & cos\theta_k \end{pmatrix}$ and $l' = \hat{l}_k/\bar{l}_k$.

Figs. 4.2(c) and 4.3(a) show the locations of the subparts derived from the main and extended body parts, respectively.

4.3.3 Contour Alignment

The orientations of the missed body parts cannot be predicted from the model and the identified parts, because the orientation relationships between the body parts are not encoded in the human model. This is solved in the second iteration of the RCR algorithm by aligning the predicted body part outlines with the detected edge features (as shown in Fig. 4.3). The procedure of the alignment is as follows. For each model body part f_i, run Steps 1 to 3:

Step 1: render the outline of body part f_i based on its parameters estimated in Sections 4.3.1 and 4.3.2. If body part f_i is not identified, then its orientation is initialized as the orientation of the torso.

Step 2: align the rendered outline with the edge features such that

$$\hat{\theta}_i = arg \max_{\theta} N(B_\theta \cap E),$$

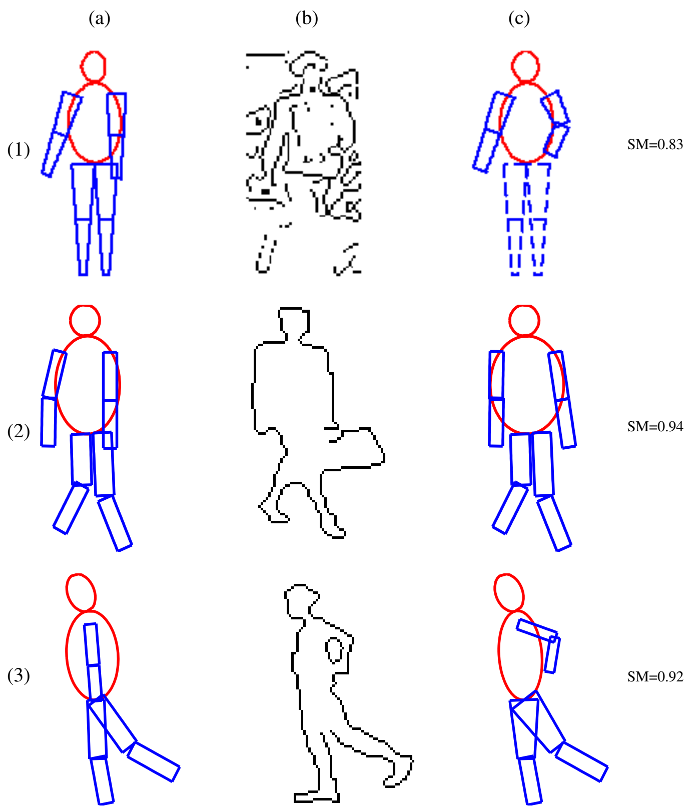

Figure 4.3: The second iteration: (a) the updated outlines of the body parts (b) the edge images (c) the aligned body parts

where B_θ is the rendered boundary of f_i at orientation θ, E is the set of edge pixels, and $N(s)$ is the number of points in the point set s.

Step 3: if $N(B_{\hat{\theta}_i} \cap E) > threshold$, then body part f_i is detected, and the points in $B_{\hat{\theta}_i} \cap E$ are removed from the edge image E. Otherwise, body part f_i is not detected.

Because of cluttered backgrounds, the alignment may be distracted by other objects. To avoid such situations, other cues such as stereo, motion, and the intensity pattern can be used to constrain the search of the body parts to be within the region of similar attributes. For example in Fig. 4.3(1), the search for the arms is constrained to be within a region having similar disparity as the torso region.

4.4 Example Runs of the RCR Algorithm

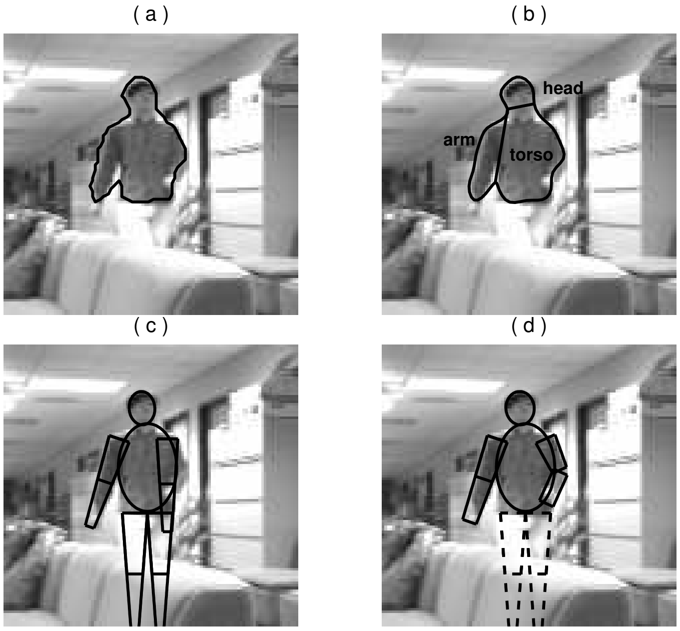

Figure 4.4: Human detection and body part localization: (a) initially detected contour (b) identified body parts (c) updated/predicted outlines of the body parts (d) aligned body parts

This section presents some example runs of the RCR algorithm. Fig. 4.4(a) shows an indoor scene with a cluttered background. The initial contour in Fig. 4.4(a) was extracted from stereo segmentation [123] and the ICP [84] algorithm. However, the outlines of the arms are not complete, and the legs are missed due to partial occlusion and the similar intensities between the clothes and the background. From this incomplete contour, three body parts are identified as indicated in Fig. 4.4(b), and the similarity between the extracted contour and the human model is 0.75 based on the Bayesian Similarity measure defined in Eq. (3.5). The updated/predicted outlines of the body parts are shown in Fig. 4.4(c). The orientations of the missed parts are set to be the same as that of the torso; their actual orientations are obtained by aligning the predicted contours with the edge features as shown in Fig. 4.4(d). In the second iteration, the estimated similarity between the set of extracted contours and the human model is 0.83. Because more body parts are detected from the image, the Bayesian similarity measure can be estimated more accurately as demonstrated by other examples given in Figs. 4.2 and 4.3.

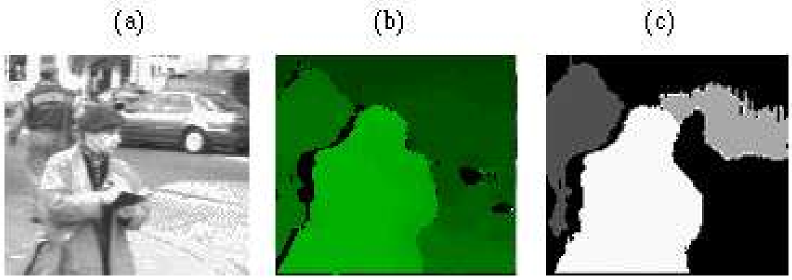

Figure 4.5: Stereo-based segmentation (a) the left image from the stereo cameras (b) the disparity map (c) the segmentation result

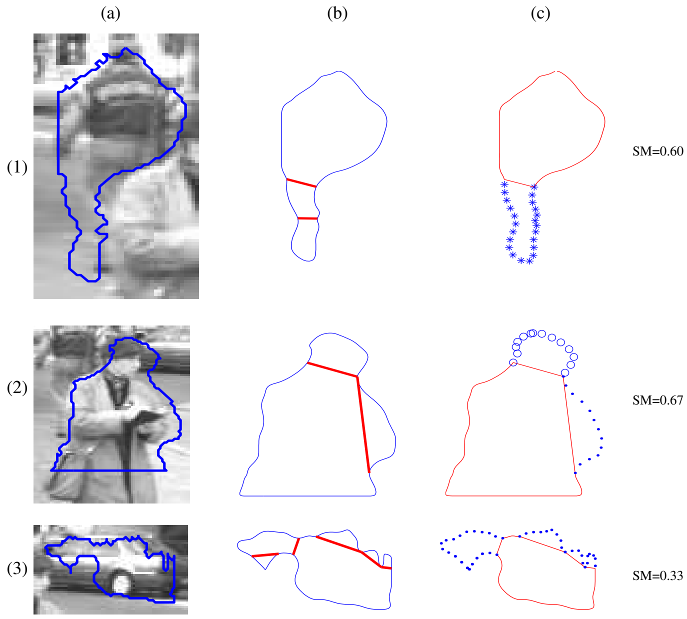

Figure 4.6: The first iteration of the RCR algorithm

Fig. 4.5(a) shows a street scene. In the first iteration, the initial outlines of the foreground objects were extracted from the cluttered background based on depth segmentation [123] (see Fig.4.5(c)). The contours of the objects are not complete and some body parts are missed due to partial occlusion or due to being out of the field of view of the camera.

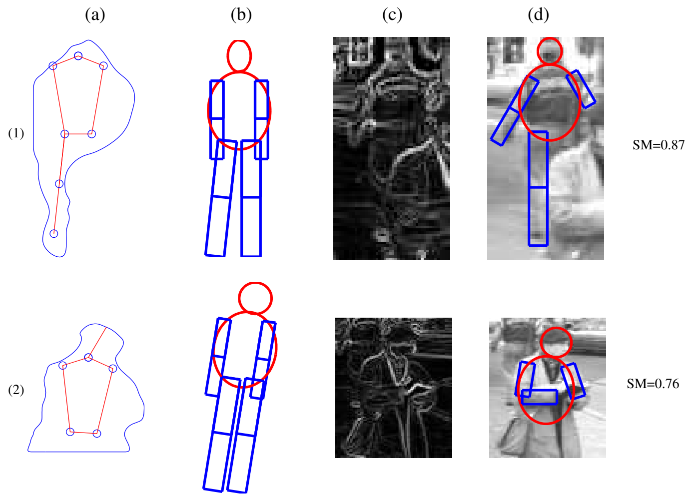

Figure 4.7: The second iteration of the RCR algorithm

From the incomplete contours, the human body parts are identified as shown in Fig. 4.6(c). According to the decision rule given in (3.19) (here $\lambda_1 = 0.4$, $\lambda_2 = 0.6$), the car is correctly identified as not a person, because the degree of similarity between it and the human model is 0.33 which is less than the lower threshold 0.4. No decision can be made on the two people in the scene in the first iteration because the degrees of similarity between their outlines and the human model (0.60 and 0.67, respectively) fall in the uncertainty interval $[0.4, 0.6]$. However, in the second iteration, the two people are correctly identified as humans as more body parts are found in the image and the degrees of similarity between them and the human model increase to 0.87 and 0.76, respectively. Usually, only two iterations are enough to achieve the correct decision.

The above examples demonstrate that despite the poor performance of the contour detector, people in an image can still be detected by exploiting the relationships between the body parts, and the RCR algorithm helps to locate the missed parts. More experimental results will be given in Chapter 5.

Chapter 5

Applications

5.1 Application I: Pedestrian Detection

Pedestrian detection is essential to avoid dangerous traffic situations; a driver assistance system can warn the driver of potential collision with nearby objects — especially pedestrians. However, detecting pedestrians from the cameras mounted on a moving vehicle is a challenging task because (1) the observer is moving or the background is changing, (2) when a pedestrian is close to the camera, some of his/her body parts will be out of the field of view, and (3) there is no control over what pedestrians wear. The following subsection presents related work. Subsection 5.5.2 describes the pedestrian detection system based on the RCR algorithm; the experimental results are given in subsection 5.5.3.

5.1.1 Related Work

Most human tracking and motion analysis systems [58, 59] employ a simple segmentation procedure such as background subtraction or temporal differencing to detect pedestrians. A serious problem with these approaches is the dynamic background caused by illumination changes or background (or camera) motion. Some techniques such as Pfinder [60], W^4 [80], and path clustering [62], have been developed to compensate for small, or gradual changes in the scene or the lighting. However, they cannot deal with large, sudden changes in the background. Although optical flow can be used to detect independently moving targets in the presence of camera motion, it is not feasible for non-rigid object extraction since the movements of the body parts are each different. Above all, a common drawback with the above approaches is the assumption that all detected objects are pedestrians; this limits the generalization and application of these schemes.

More sophisticated pedestrian detection systems include a recognition step to discriminate humans from other objects. These techniques can be classified into motion-based, shape-based, and multi-cue-based methods. Most motion-based approaches [64, 72, 73, 74] use cyclic features or motion patterns unique to human beings for human detection. However, there are several limitations with these schemes. First, the human's feet or legs should be visible in order to extract cyclic features. Second, the recognition procedure requires a sequence of images, which delays the identification until several frames later and increases the processing time. Third, the procedure cannot detect stationary humans and humans performing unconstrained and complex movement such as wandering around, turning, jumping, etc.

On the other hand, the shape-based approach relies on shape features to recognize humans. Thus, this approach can detect both moving and stationary people. The primary difficulty in this approach is accommodating the wide range of variations in human appearance due to pose, non-rigid motion, lighting, clothing, occlusion, etc. Hogg [58] and Rohr [59] used hand-crafted human models to detect humans. An advantage of these methods is that they can analyze the motion of each body part; the disadvantage is that the models only encode the shape of the human body and cannot be used to analyze people wearing loose clothes. Lipton [71] depends on a dispersedness defined as the ratio $perimeter^2/area$ to classify human and vehicle. This classification metric is easy to calculate, but fails to distinguish humans from other objects with similar dispersedness and tends to misclassify humans walking together as a vehicle. Papageorgious and Poggio [37] present a more robust human detection system based on wavelet analysis and the support vector machine (SVM) technique. However, the system has to search the whole image at multi-scales for humans.

This would be an extremely computationally expensive procedure, and it may cause multiple responses from a single human. More recently, Gavrila and Philomin [69] developed a real-time human detection algorithm based on Distance Transforms. The method includes an offline generation of the template hierarchy, and an online coarse-to-fine matching between the templates and the image. The algorithm is further sped up by hardware-specific means (i.e. SIMD instructions). Although the template hierarchy can capture the variety of object shapes, it can not handle large shape variations and partial occlusion appropriately when pedestrians are very close to the camera.

To increase reliability, some systems [63, 78] integrate multiple cues such as stereo, color, face, and shape to detect humans. However, skin color is very sensitive to illumination changes [63]; face detection can only identify people facing the camera. Felzenszwalb and Huttenlocher's method [87] can only detect people of fixed size and wearing special colored clothes, while Forsyth *et al.*'s approach [25, 26] combines skin color and a part-based method to detect naked people. These systems prove that stereo and shape are more reliable and helpful cues than color and face detection in general situations.

5.1.2 Pedestrian Detection System

The pedestrian detection system developed in this thesis includes two steps: first, separate foreground objects from the background; second, distinguish pedestrians from other objects in order to protect pedestrians in danger. The first task is a segmentation procedure. In much of this thesis, stereo-based depth segmentation is used to extract foreground objects. Using stereo to guide pedestrian detection carries with it some distinct advantages over conventional techniques. First, it allows explicit occlusion analysis and is robust to illumination changes. Second, the real size of an object derived from the disparity map provides a more accurate classification metric than the image size of the object. Third, using stereo cameras can detect both stationary and moving objects. Fourth, depth information helps to reduce the foreshortening problem facing 2D modeling of 3D articulated objects. Fifth, computation time is significantly reduced by performing recognition where objects are detected; it is less likely to detect the background area as a pedestrian since detection is biased toward areas where objects are detected. The algorithm proceeds in several stages of processing as explained below. In a collision warning system, we are only concerned about the objects within a close distance from our vehicle. Thus, the objects that are far away from the vehicle are first eliminated from the disparity image by range thresholding. Then a morphological closing operator is employed to remove the noise and to smooth the foreground regions. Then, a connected-component grouping operator is applied to find the foreground regions with smoothly varying range. Finally, small regions are eliminated through size thresholding. The size range of an average person is obtained from the statistical data given in [116]. Fig. 5.3 presents the segmentation results of some street scenes. The results demonstrate that the overlapping objects can be successfully separated if they are at different distances from the cameras.

The second task is a recognition procedure which is achieved through the recursive context reasoning algorithm proposed in this thesis. A sample run of the RCR algorithm is given in Figs. 4.5-4.7. The decision if an object is a person or not is based on the decision rule described in (3.19). The thresholds λ_1, λ_2 determine the trade off between the rate of pedestrian detection and the rate of false alarm; they are selected by evaluating the *receiver operating characteristics* (ROC) curve illustrated in Fig. 5.1 generated by testing the RCR algorithm on 20 sequences (100 frames/sequences) of urban street scenes. By setting the thresholds $\lambda_1 = 0.4, \lambda_2 = 0.6$ in the decision rule, the system achieves a pedestrian detection rate of 86.3% and a false alarm rate of 2.1% (the false alarm rate is estimated as the ratio between the false positives and the number of non-human object segmented from an image).

5.1.3 Experimental Results

The system has been implemented on a Pentium III 500 Mhz system under Microsoft Windows 98. It has been tested on the videos of urban areas obtained from a stereo system mounted on the top of a minivan. The stereo system which is developed by the Point Grey

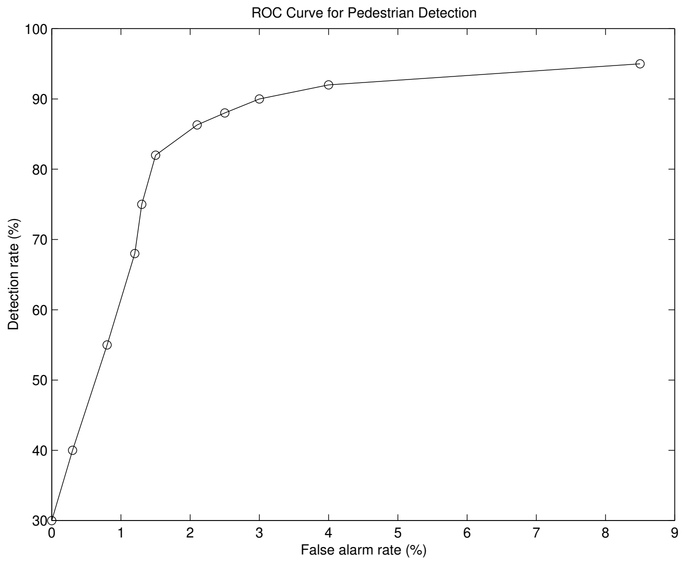

Figure 5.1: The ROC curve for threshold selection

Research Inc. consists of three cameras as shown in Fig. 5.2. The processing speed of the stereo system is 9 frames/second at 320x240 image resolution and 48 disparity range.

Figure 5.2: The Digiclops stereo system

The pedestrian detection system can detect and classify objects at a frame rate ranging from 1 frames/second to 6 frames/second, depending on the number of objects presented in the field of view of the cameras.

Fig. 5.3 illustrates some sample results of detecting pedestrians in urban areas. The

Figure 5.3: Sample results of pedestrian detection (the unidentified objects are circled by dotted lines)

results demonstrate that the developed pedestrian detection system can detect pedestrians in different sizes, clothing, and occlusion status. The reason of the uncertainty with an object's identification is that the object (such as a tree) resembles a human at a coarse level but not at a fine level. The main reasons for the failure of detecting a pedestrian are that (1) the pedestrian is not segmented from the background correctly because his/her clothes lack texture or have the similar color as the background, (2) the pedestrians walk in a group, or (3) the vehicle moves fast and the images are blurred. The false alarms are mainly due to the human-like shape faked from the cluttered scene.

Table 5.1: Comparing the performance of the CMU systems and MIT systems.

	Detection rate	**False alarms**	**Frame rate**
RCR	86.3 %	7	1-6f/s (PC 500MHz)
ANN	85.4%	18	3-8f/s (PC 500MHz)
MIT I	50%	100	10f/s (SUN)
MIT II	81.6%	14	1f/h (SGI)

It is worthwhile to compare this RCR-based pedestrian detection system with the previous work on Neural network-based pedestrian detection [123] and MIT's wavelet-based system [37]. To allow a fair comparison, let all methods work on the same set of test images. The test set includes 617 people candidates generated from 100 images through the range segmentation procedure. Among these candidates, there are 254 pedestrians and 363 non-pedestrians. Table 5.1 presents the results of comparing the performance of the above pedestrian detection systems. MIT I is a fast version which only uses intensity information, while MIT II employs color information for pedestrian detection. The results demonstrate that the RCR algorithm achieves the similar human detection rate and the lowest false alarm rate. This is because the RCR algorithm employs the context information to perform and improve both contour extraction and human detection iteratively. Thus, our pedestrian detection system is faster and more accurate than MIT's system. The advantage of the NN-based and wavelet-based methods is that no contours are required to extract from the image first. The disadvantage of the NN-based method is that it may generate many false alarms because the separation space is not closure [118]. For example, a parking meter can often be identified as a pedestrian using NN-based method, but it is labeled as non-pedestrian correctly as shown in Fig. 5.3 using the RCR algorithm. The disadvantage of the wavelet-based method is that it has to search for pedestrians in different scales in the whole image, so it is very slow. Another limitation with these two approaches is that they are global shape based and cannot deal with partial occlusion very well. However, for the application of pedestrian detection, it is more important to detect the pedestrians near the vehicle to avoid potential accident; when a pedestrian is close to the camera, some of his/her body parts tend to be out of the field of view of the camera.

5.2 Application II: Human Motion Capture

Human motion capture plays an important role in a wide spectrum of applications such as visual surveillance, performance measurement for atheletes and patients with disabilities, human-computer interfaces, figure animation, and video conference (see Aggarwal *et al* [57] for a general review). Many human motion capture devices need to employ special markers or magnetic sensor attachements around the joints of a subject. Thus, they impose physical restrictions on the subject. Vision-based body part localization is a way to enable mark-free human motion capture. However, locating the body parts especially the joints in an image is a difficult problem, because the joints are hidden by muscle, skin, and clothing. Many semi-automatic body part tracking systems [65, 88, 89] and 3D pose recovery systems [91, 83] require a user to locate the body parts/joints manually in the image, which is a time consuming and tedious task for the user. The RCR algorithm proposed in this thesis can be used as the automatic initialization of these systems. The following sections first discuss the related work, then present the experimental results of human motion capture.

5.2.1 Related Work

The earliest computer vision attempt to recognize human movements was reported by O'Rourke and Badler [19] working on synthetic images using a 3D structure of rigid segments, joints, and constraints between them. Previous approaches to automatic body part localization in real images can be classified into two categories. One is labeling or feature-to-feature matching [25, 60, 78, 79, 81, 90]. A set of body part candidates is first extracted from the image, then the candidates are matched against the modeled body parts. The advantage with this approach is that the search space is discrete and small, and the comparison is at the feature level. The main challenge faced by this approach is that not all body part candidates can be detected reliably and located accurately. The problem is how to perform matching in the absence of some body parts and/or presence of extra parts (such as the attached objects or clothes). Some systems [25, 60] employ skin color and shape configuration to extract body parts not covered by the clothes or other objects; some systems employ certain heuristic methods, such as the special order along the silhouette [78], a motionless support part [79], or hierarchical order [17, 24] to locate the body parts sequentially; some require an initial calibration [81, 60]. More recently Rosales and Sclaroff [90] employ a learned mapping function to do matching. However, these approaches are very sensitive to occlusion and shape deformation caused by clothes or segmentation errors, and they just provide coarse locations of the body parts.

Another approach is the alignment method [86, 87] or the feature-to-image matching. The matching is conducted by projecting the body part features into the image and comparing directly. The goal is to find the best pose that aligns the mapped body parts with the image. The advantage of this approach is that no figure/background segmentation and feature extraction are required. One of the disadvantages is that the search space is continuous and large, and the comparison is at the pixel intensity level. Both Cham and Rehg[86] and Felzenszwalb and Huttenlocher [87] have proposed efficient algorithms to perform alignment, but they need different color/template models for different people.

In contrast, the RCR algorithm proposed in this thesis is a combination of the labeling and alignment methods. The algorithm achieves both efficient and accurate body part localization by making use of the advantages of the above two approaches and overcoming their disadvantages. The combined approach has been used in [109] for 2D-2D feature matching and 2D-3D matching in [85]. The common feature is that the uncertainty information has been used for feature matching and for determining transformation uncertainty with which to predict the positions of adjacent features. The main differences of the RCR algorithm from other related work are: (1) the match is performed at object parts level instead of features (such as lines, arcs) level. The number of object parts is significantly smaller than that of low-level features; this enables a more efficient matching procedure. Parts are a more natural and stable representation of articulated objects than other features; this representation has found strong support from human vision [51]. (2) previous work uses a single model to perform matching while in the RCR algorithm, the model is dynamically assembled from the model parts according to the match hypothesis. (3) previous work usually assumes independent feature matching and conducts search in a depth-first manner. In contrast, in the RCR algorithm, the search of the matching of the same level body parts is conducted in a breadth-first manner to guarantee global optimal match and to avoid backtracking. (4) The propagation of position prediction is not only from connected parts but also from the extended body parts. This enables more robust joint localization against contour extraction and decomposition errors. The matching is conducted in a coarse-to-fine manner to further facilitate fast body part localization.

5.2.2 Experimental Results

The body parts of a person are located in a coarse-to-fine manner using the RCR algorithm. First, the person is segmented from the background (shown in Fig. 5.4(b)). This is done through background subtraction [71]. Second, the segmented region is decomposed into ribbons and these ribbons are matched with the modeled body parts including the extended parts (shown in Fig. 5.4(c)). The joints are initially located in the middle of a cut segment. Third, the locations of the joints are adjusted to achieve consistency with the modeled spatial and size relationships between the body parts. The locations and the sizes of the missed body parts are inferred from the extended body parts and the detected body parts (shown in Fig. 5.4(d)). Fourth, the predicted outlines of the body parts are aligned with the edge features in Fig. 5.4(e). The final results are shown in Fig. 5.4(f). Fig. 5.4 presents some examples of locating the body parts of a person walking in a parking lot. Because the human model proposed in this thesis is two dimensional, it cannot be used to distinguish front-view from the side-view very well as shown in Section 5.5. In this experiment, the view point is assumed to be the side-view. When the arms are overlapped with the torso, it is very hard to locate them, and they may be aligned with the outline of the torso by mistake. Another problem is that the left and right limbs tend to be confused in a side-view. Motion information obtained from previous frames can be used to predict the orientations of the limbs and to solve the ambiguity. Fig. 5.6 presents a full cycle of a walking person. In

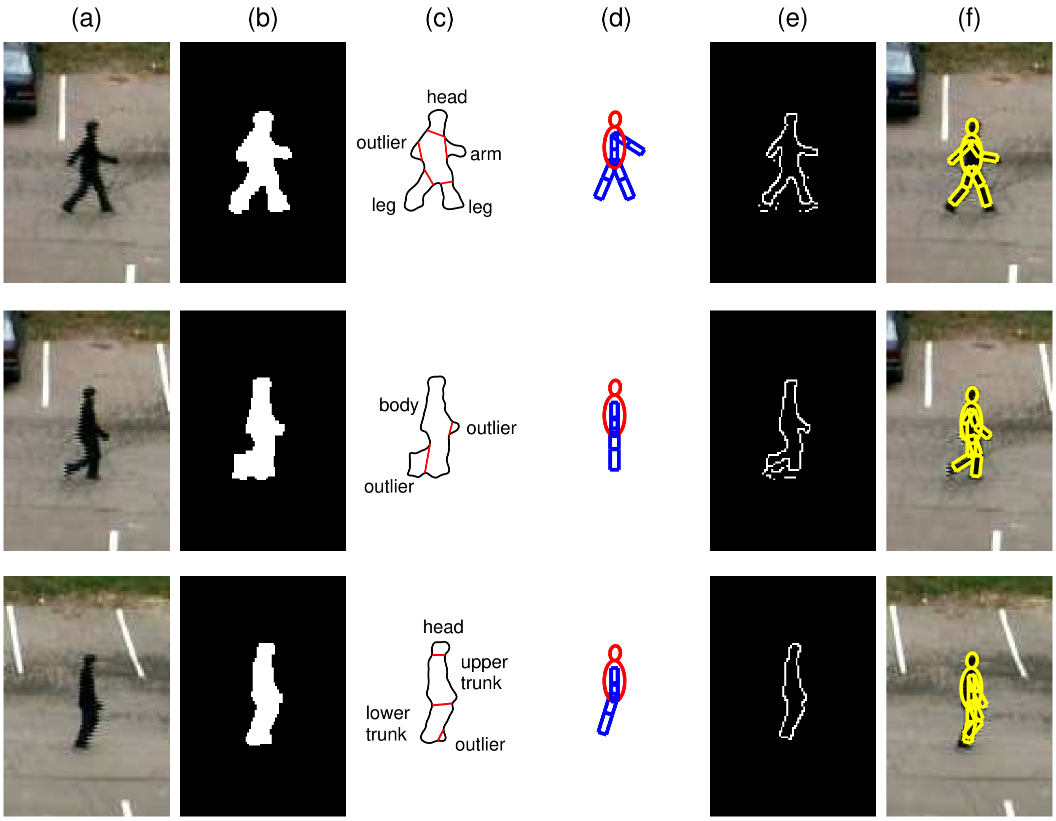

Figure 5.4: Body part localization: (a) images (b) foreground object detected from background subtraction (c) identified body parts (d) updated/predicted locations and outlines of the body parts (e) edge images (f) aligned body parts.

the first half cycle, no motion information is used to resolve the ambiguity with limbs' orientations, while in the second half of the cycle, the prediction from previous frames (constant angular velocity is assumed) is used to get better results of body part localization. Fig. 5.5 illustrates how the left knee angle changes with time. The motion information is not used from frames 1 to 30, but is from frames 31 to 190. Therefore, the left and right legs are switched sometimes during the first part. From frames 31 to 190, an obvious pattern of walking cycle can be observed.

To demonstrate that the RCR algorithm can also be used to analyze other motions besides walking, it is tested on the images of figure skating. This is a challenge task, be-

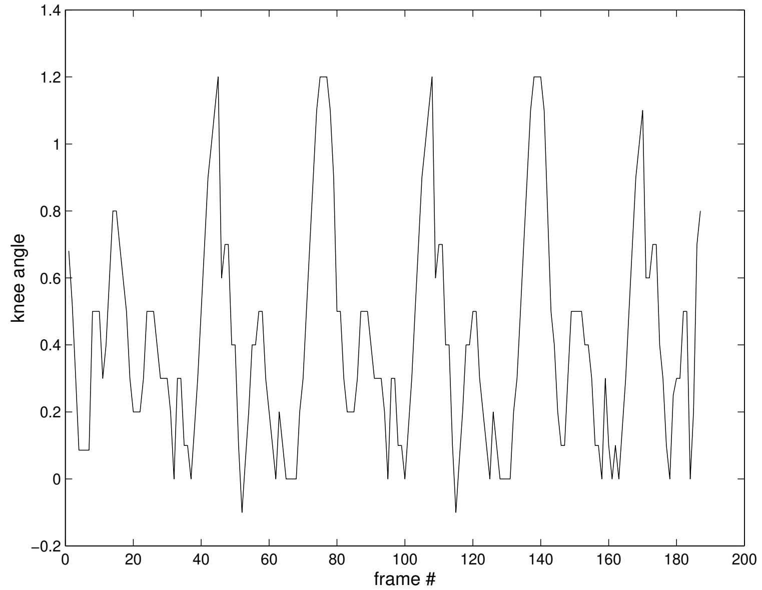

Figure 5.5: The change of angle at the knee with time

cause skaters may perform very complicate actions, wear loose clothes, and appear together (pairs). In the experiments, the silhouettes of skaters are first segmented from the background through intensity thresholding. Then the ribbons are extracted from the silhouette using a fine-to-coarse shape decomposition method described in Chapter 2. Then the body parts are identified based on the Bayesian similarity measure illustrated in Chapter 3; the locations of the joints are updated using the RCR algorithm described in Chapter 4. The results shown in Fig. 5.7 demonstrate that even if the skaters are in various postures and wear skirts flying in the air, the covered body parts can still be located accurately using the dynamically assembled human model, the Bayesian similarity measure, and the coarse-to-fine localization approach. The algorithm can also distinguish a pair of skaters skating together as shown in Fig. 5.7. To illustrate the accuracy of the joint locations, the 3D pose of the skaters are recovered with the approach proposed by Taylor [91]. Because only the 2D model and 2D image are used to perform joint localization, some joints cannot be located correctly. The next chapter will discuss the future work directions to overcome these limitations.

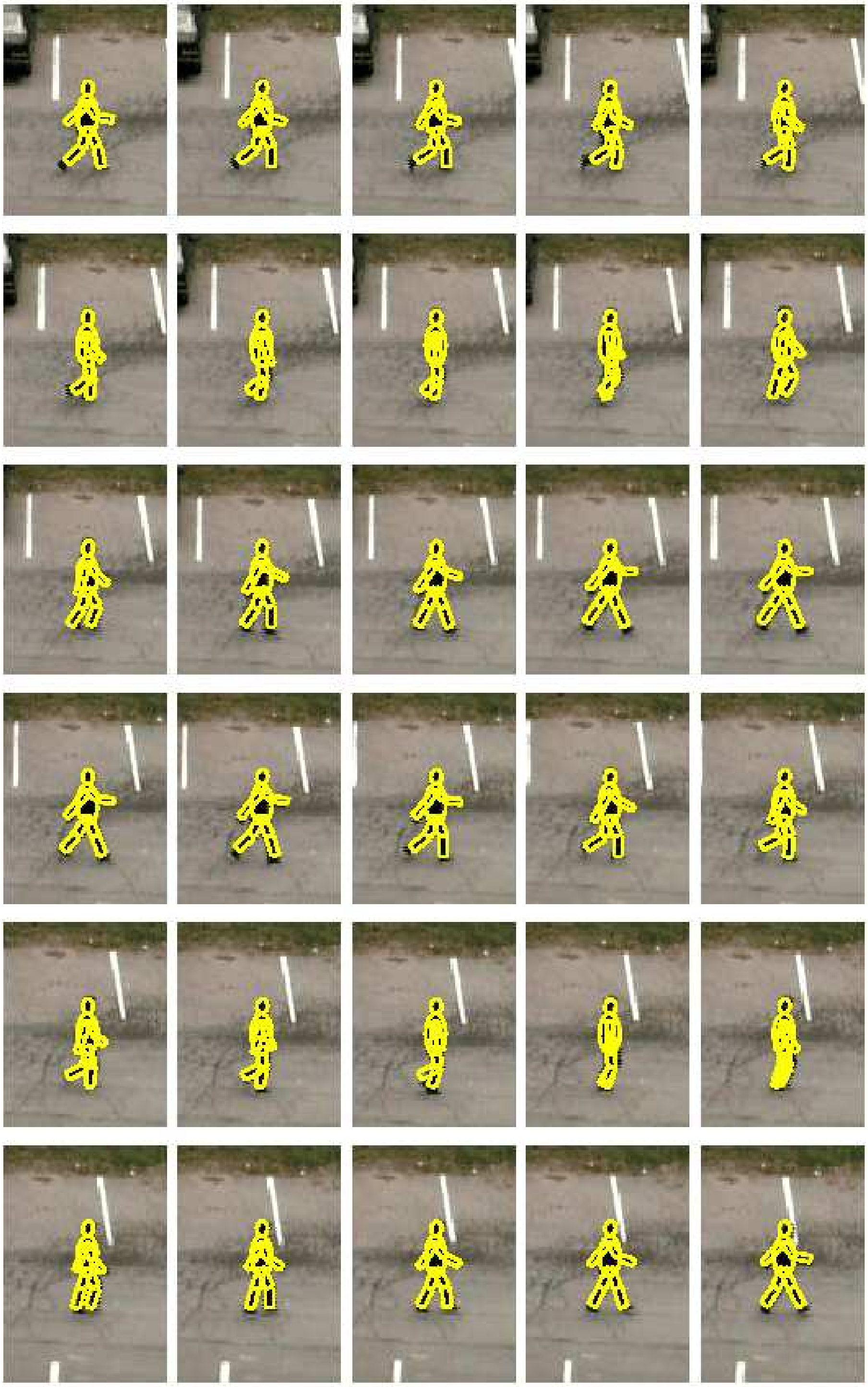

Figure 5.6: Markless human motion capture

Figure 5.7: Joint localization of figure skaters

Chapter 6

Conclusion

This thesis has addressed several computer vision problems: how to represent a class of deformable shapes, how to match a contour with a model and evaluate their similarity, and how to improve contour extraction and object detection by combining bottom-up and top-down procedures. Section 6.1 summarizes the contributions of this thesis in response to those questions. Section 6.2 discusses the limitations with this work and suggests future directions to improve and generalize this work.

6.1 Contributions

This thesis presents an integrated human shape modeling, detection, and body part localization vision system. It demonstrates that the system can (1) detect pedestrians in various shapes, sizes, postures, partial occlusion, and clothing from a moving vehicle using stereo cameras; (2) locate the visible joints automatically and accurately without employing any markers around the joints of a subject.

The following contributions distinguish this thesis from previous work:

1. Dressed human modeling and dynamic model assembling: Unlike previous work that employs a fixed human body model or global deformable template to perform human detection, in this thesis merged body parts are introduced to represent the deforma-

tions caused by clothing, segmentation errors, or low image resolution. A dressed human model is dynamically assembled from the model parts in the recognition step; the shapes of the body parts and the size and spatial relationships between them (the contextual information) are represented as invariant under translation, rotation, and scaling. Therefore, the system can detect people in different clothes, positions, sizes, and orientations.

2. Bayesian similarity measure: A probabilistic similarity measure is derived from the human model that combines the local shape and global relationship constraints into a single equation to guide body part identification and human detection. Thus, the identification of a part does not only depend on its own shape but also the contextual constraints from other parts. In contrast with previous work, the Bayesian similarity measure enables efficient shape matching and comparison robust to articulation, partial occlusion, and segmentation errors through coarse-to-fine human model assembling.

3. Recursive context reasoning algorithm: Contour-based human detection depends on reliable contour extraction, but contour extraction is an under-constrained problem without the knowledge about the objects to be detected. Unlike previous work that assumes perfect and complete contours are available, this thesis proposes a recursive context reasoning (RCR) algorithm to solve the above dilemma. A contour updating procedure is introduced to integrate the human model and the identified body parts to predict the shapes and locations of the parts missed by the contour detector; the refined contours are used to reevaluate the Bayesian similarity measure and to determine if a person is present or not. Therefore, contour extraction, body part localization, and human detection are improved iteratively.

While there are many areas to which human detection and body part localization can be applied, the chosen domain for this thesis is vehicle safety. Despite the specific nature of the chosen application area, the underlying ideas and technologies presented in this thesis are readily applicable to other tasks and to detecting other objects. First, the TRS-invariant probabilistic representation can be used to model other object categories as well. Second, the Bayesian similarity measure can be used to classify other objects and to perform shape retrieval. Third, the RCR algorithm is useful for vision tasks such as object tracking as well by incorporating motion information into the model.

6.2 Limitations and Future Work

The experimental results demonstrate the effectiveness of the proposed techniques, but also reveal their limitations.

First, the shapes of body parts are represented with their aspect ratios. This representation reduces the number of parameters used in the model, but it is too general to distinguish body parts of different objects. Although size and spatial relationships between body parts help to identify body parts, it is difficult to distinguish two animals with similar spatial arrangement and size proportions. Therefore, a more accurate representation of the shape of a body part such as a deformable template is required when there is ambiguity with a coarse-level representation.

Second, the current human model does not incorporate the orientation relationships among body parts. As a result, the identification of a limb being at the left or the right side of the body may be wrong due to the absence of posture constraints. In the future, the posture constraints will be incorporated into the human model and into the Bayesian similarity measure to correct this kind of errors.

Third, the human model has only two dimensions, so it has difficulty to analyze viewpoint changes and foreshortening. The future work may extend the human model to three dimensional and incorporate the three dimensional data to handle the above difficulties. Because shape matching is at the parts-level instead of the point-level, 3D shape reconstruction and object recognition would be more robust and efficient.

Fourth, the body part identification procedure does not handle the situation when multiple ribbons correspond to the same body part. As a result, those ribbons will be identified as outliers. Although the RCR algorithm can correct some of these errors, a more reliable solution would be to allow the union of multiple ribbons to be matched with a single body part, or to incorporate a grouping operator to group those ribbons into a single body part.

Fifth, only the contour cue is used for body part localization and human detection. However, there is inherent ambiguity with some contours such as the one illustrated in Fig. 6.1. The contour looks like a left-facing duck as well as a right-facing rabbit. Therefore, it is necessary to keep multiple hypotheses instead of just the best one to overcome the ambiguous and noisy data. The future work may combine other cues such as motion and color to resolve the inherent uncertainty with contour and contour detection errors. The combination of the color and spatial information has been successfully used to segment people in a group [76].

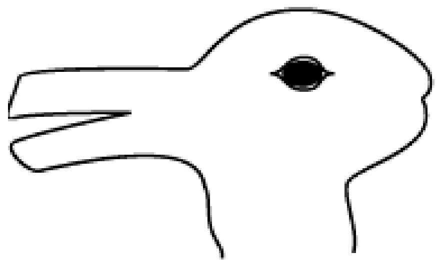

Figure 6.1: Duck-Rabbit example of the uncertainty with contour

Sixth, the RCR algorithm requires an initial contour extraction which is not available in some cases such as photo analysis. One way to overcome this limitation is to use a face detection system to locate the faces of people in the image, then use the approach proposed in this thesis to check the presence of the remaining body parts to verify if the detected face is a human's face.

Seventh, current pedestrian detection system is slow (1-6 frame/second) because it relies on a stereo vision system (9 frames/second) to perform initial contour extraction. One way to improve the speed of the system is to use face detection first as mentioned in the previous paragraph. Another way is to use the thermal Infrared image instead to perform initial contour extraction.

It is an open question whether there is a uniform approach to general object detection or not. Previous efforts toward general object recognition and detection do not work very well at the category/functional level. For example at the category/functional level, the representations/models of trees, clouds, rivers are significantly different from that of chairs [47]. Therefore, it is more effective to develop different methods to analyze different classes of objects.

The techniques developed in this thesis are useful for objects for which the decomposition into natural parts is well-defined such as animals. Figs. 3.8 and 3.9 demonstrate that the designed Bayesian similarity measure can be used to distinguish different animals. Thus, this work has potential to be applied to detect other articulated objects by building a corresponding object model and using the appropriate feature detectors. For example, in order to detect horses in an image, a horse model similar as the human model can be built to encode the shapes of horse body parts and their geometrical relationships; a Bayesian similarity measure based on the horse model can be designed to identify the body parts of the horse and to determine if the assembly of the parts corresponds to a horse. Furthermore, the RCR algorithm can be used to improve horse detection iteratively.

The work can be extended to include a relational database of object models. Then context reasoning can be performed both within an object and among objects. For example, if a school bus is detected, then it is highly possible to see children around. The work can also be extended to object tracking by incorporating dynamic object models into the system [124]. The above extensions and improvements will lead us to a more reliable object detection and tracking system.

Appendix: Parameters of the Human Model

The parameters of the human body model are estimated based on a large quantity of data accumulated over more than 40 years by Henry Dreyfurs, Associates and published by Tilley [116]. Tilley provides both the body measurements of people at different ages and the clothing corrections. According to NASA [117], human variability falls into three categories:

1. Intra-individual: size change during adult life;
2. Inter-individual: there are big difference due to sex and ethnic and racial membership;
3. Secular variability: changes occur from generation to generation for various reasons. Since the pace of these changes is relatively slow, they have a limited effect on the human body model.

The parameters of the dressed human model are learned from the 500 training data collected from many web-sites and catalogs (Fig. 6.2 shows some examples). The following tables list the parameters used in this thesis (Symbols used in the tables: ts: torso, hd: head, am: arm, lg: leg, hn: hand, ft: foot, ut: upper trunk, lt: lower trunk, bd: body, la: left arm, ll: left leg, rl: right leg, ra: right arm).

Figure 6.2: Examples of the training data for human model learning

Table 6.1: The means and the standard deviations of the aspect ratios

		ts	hd	am	lg	hn	ft	ut	lt	bd
front	$\bar{a}$	.61	.78	.25	.25	.70	.42	.92	.43	.30
view	σ_a	.10	.09	.12	.08	.20	.13	.08	.10	.08
side	$\bar{a}$	.45	.78	.25	.25	.70	.42	.73	.22	.26
view	σ_a	.11	.09	.12	.08	.20	.13	.09	.11	.09

Table 6.2: The means of the length ratios

	ts	hd	la	ll	rl	ra	ut	lt	bd
ts	1.0	.52	.95	1.47	1.47	.95	1.0	1.47	3.0
hd	1.92	1.0	1.83	2.84	2.84	1.83	1.92	2.84	5.76
la	1.05	.55	1.0	1.55	1.55	1.0	1.05	1.55	3.16
ll	.68	.36	.66	1.0	1.0	.66	.68	1.0	2.04
rl	.68	.36	.66	1.0	1.0	.66	.68	1.0	2.04
ra	1.05	.55	1.0	1.55	1.55	1.0	1.05	1.55	3.16
ut	1.0	.52	.95	1.47	1.47	.95	1.0	1.47	3.0
lt	.68	.36	.66	1.0	1.0	.66	.68	1.0	2.04
bd	.33	.18	.32	.49	.49	.32	.33	.49	1.0

Table 6.3: The standard deviations of the length ratios

	ts	hd	la	ll	rl	ra	ut	lt	bd
ts	0	.05	.05	.09	.09	.05	.01	.09	.08
hd	.18	0	.21	.36	.36	.21	.18	.36	.53
la	.05	.07	0	.04	.04	.01	.05	.04	.10
ll	.04	.06	.02	0	.01	.02	.04	.01	.10
rl	.04	.06	.02	.01	0	.02	.04	.01	.10
ra	.05	.07	.01	.04	.04	0	.05	.04	.10
ut	.01	.05	.05	.09	.09	.05	0	.09	.08
lt	.04	.06	.02	.01	.01	.02	.04	0	.10
bd	.01	.02	.01	.02	.02	.01	.01	.02	0

Table 6.4: The means of the coordinates of the body parts in the normalized torso coordinate system

		ts	hd	la	ll	rl	ra	ut	lt
front	$\bar{x}$	0	0	-.31	-.163	.163	.31	0	0
view	$\bar{y}$	0	.5	.353	-.472	-.472	.353	0	-.5
side	$\bar{x}$	0	0	0	0	0	0	0	0
view	$\bar{y}$	0	.5	.353	-.472	-.472	.353	0	-.5

Table 6.5: The covariance of the coordinates of the body parts in the normalized torso coordinate system (front view)

	$\times 10^{-2}$									
x_{hd}	0.34	0.11	0.08	0.03	0.04	0.13	-0.06	0.15	-0.09	0.04
y_{hd}	0.11	0.92	0.09	0.56	0.13	0.91	-0.15	0.92	-0.10	0.55
x_{la}	0.08	0.09	0.30	0.07	0.16	0.19	-0.17	0.14	-0.33	0.06
y_{la}	0.03	0.56	0.07	0.95	0.05	1.13	-0.03	1.01	-0.07	0.91
x_{ll}	0.04	0.13	0.16	0.05	0.39	0.23	-0.35	0.27	-0.16	0.05
y_{ll}	0.13	0.91	0.19	1.13	0.23	1.82	-0.27	1.89	-0.19	1.13
x_{rl}	-0.06	-0.15	-0.17	-0.03	-0.35	-0.27	0.41	-0.32	0.17	-0.06
y_{rl}	0.15	0.92	0.14	1.01	0.27	1.89	-0.32	1.27	-0.20	1.23
x_{ra}	-0.09	-0.10	-0.33	-0.07	-0.16	-0.19	0.17	-0.20	0.35	-0.08
y_{ra}	0.04	0.55	0.06	0.91	0.05	1.13	-0.06	1.23	-0.08	0.90

Table 6.6: The covariance of the coordinates of the body parts in the normalized torso coordinate system (side view)

	$\times 10^{-2}$									
x_{hd}	0.18	0.10	0.06	0.02	0.02	0.09	-0.08	0.13	-0.07	0.02
y_{hd}	0.10	0.85	0.07	0.45	0.11	0.87	-0.13	0.88	-0.19	0.50
x_{la}	0.06	0.07	0.13	0.08	0.14	0.15	-0.11	0.17	-0.25	0.07
y_{la}	0.02	0.45	0.08	0.91	0.06	1.11	-0.04	1.06	-0.04	0.83
x_{ll}	0.02	0.11	0.14	0.06	0.20	0.13	-0.30	0.25	-0.13	0.06
y_{ll}	0.09	0.87	0.15	1.11	0.13	1.67	-0.25	1.71	-0.12	1.02
x_{rl}	-0.08	-0.13	-0.11	-0.04	-0.30	-0.25	0.22	-0.21	0.15	-0.05
y_{rl}	0.13	0.88	0.17	1.06	0.25	1.71	-0.21	1.18	-0.21	1.14
x_{ra}	-0.07	-0.19	-0.25	-0.04	-0.13	-0.12	0.15	-0.21	0.16	-0.07
y_{ra}	0.02	0.50	0.07	0.83	0.06	1.02	-0.05	1.14	-0.07	0.92

Bibliography

[1] R.A. Brooks, "Symbolic Reasoning Around 3-D Models and 2-D Images," *Artificial Intelligence J.*, Vol. 17, pp. 285-348, 1981.

[2] D.G. Lowe, "Three-Dimensional Object Detection from Single Two-Dimensional Images," *Artificial Intelligence J.*, Vol. 31, pp. 355-395, 1987.

[3] Y. Lamdan and H.J. Wolfson, "Geometric Hashing: A General and Efficient Model-Based Recognition Scheme," *Proc. of the 2'nd Int. Conf. on Computer Vision*, pp. 238-249, 1988.

[4] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson, "On Recognition of 3-D Objects from 2-D Images," *Proc. of Int. Conf. on Robotics and Automation*, Vol. 3, pp. 1407-1413, Philadelphia, April, 1988.

[5] D.D. Ballard, "Generalizing the Hough Transform to Detect Arbitrary Shapes," *Pattern Recognition*, Vol. 13, No. 2, pp. 111-122, 1981.

[6] D.P. Huttenlocher and S. Ullman, "Object Recognition Using Alignment," *Proc. of the 1'st Int. Conf. on Computer Vision*, pp. 102-111, London, 1987.

[7] D.P. Huttenlocher and S. Ullman, "Recognizing Solid Objects by Alignment," *Proc. of the DARPA Image Understanding Workshop*, Vol. II, pp. 1114-1122, Cambridge, MA, April, 1988.

[8] T. Silberberg, D. Harwood, and L. Davis, "Object Recognition Using Oriented Model Points," *Computer Vision, Graphics and Image Processing*, Vol. 35, pp. 47-71, 1986.

[9] S. Linainmaa, D. Harwood, and L.S. Davis, "Pose Determination of a Three-Dimensional Object Using Triangle Pairs," *CAR-TR-143*, Center for Automation Research, University of Maryland, 1985.

[10] D. Thompson and J.L. Mundy, "Three-Dimensional Model Matching from an Unconstrained Viewpoint", *Proc. IEEE Conf. Robotics and Automation*, pp. 208-220, 1987.

[11] W.E.L. Grimson, D.P. Huttenlocher, and D.W. Jacobs, "Affine Matching with Bounded Sensor Error: A Study of Geometric Hashing and Alignment," *Technical Report 1250*, M.I.T. Artificial Intelligence Laboratory, August, 1991.

[12] http://www.dai.ed.ac.uk/CVonline/recog.htm (a list of object recognition methods).

[13] http://www.dai.ed.ac.uk/CVonline/repres.htm (a list of shape representations).

[14] L. Schomaker, E. Leau, L. Vuurpijl, "Using Pen-based Outlines for Object-based Annotation and Image-based Queries," *Proc. VISUAL'99*, pp. 585-592, 1999.

[15] C. Mertz, S. McNeil, and C. Thorpe, "Side Collision Warning Systems for Transit Buses," *IV 2000, IEEE Intelligent Vehicle Symposium*, October, 2000.

[16] S. McNeil, C. Thorpe, and C. Mertz, "A New Focus for Side Collision Warning Systems for Transit Buses" *ITS2000, Intelligent Transportation Society of America's Tenth Annual Meeting and Exposition*, May, 2000.

[17] D. Marr and H.K. Nishihara, "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes," Proc. Roy. Soc. B, B-200, pp. 269-294, 1977.

[18] D. Marr, *Vision*, W. H. Freeman and Co., 1982.

[19] J. O'Rourke and N.I. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, Vol 2, No. 6, pp. 522-536, Nov. 1980.

[20] A. P. Pentland, "Perceptual Organization and the Presentation of Natural Form," *Artificial Intelligence*, 28, pp. 293-331, 1986.

[21] A. Pentland, "Recognition by Parts," *Proc. Int'l Conf. on Computer Vision*, pp. 612-620, 1987.

[22] J. Segen, "Model Learning and Recognition of Nonrigid Objects," CVPR'89, pp. 597-602, 1997.

[23] S. C. Zhu and A. L. Yuille, "Forms: A Flexible Object Recognition and Modeling System," Int'l Journal of Computer Vision, Vol.20, No.3, 1996.

[24] A. Hauck, S. Lanser, C. Zierl, "Hierarchical Recognition of Articulated Objects from Single Perspective Views," CVPR'97, San Juan, Puerto Rico, June 17-19, pp. 870 - 876, 1997.

[25] D.A. Forsyth, M.M. Fleck, "Body Plans," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[26] S. Ioffe, D.A. Forsyth, "Finding People by Sampling," *Proc. Int'l Conf. on Computer Vision*, 1999.

[27] T.K. Leung, M.C. Burl, and P. Perona, "Finding Faces in Clutteres Scenes Using Random Labeled Graph Matching," *Proc. Int'l Conf. on Computer Vision*, June 1995.

[28] T.K. Leung, M.C. Burl, and P. Perona, "Probbilistic Affine Invariants for Recognition," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 1998.

[29] M.C. Burl, M. Weber and P. Perona, "A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry," *Proc. European Conf. on Computer Vision*, Vol. 2, pp. 628-642, 1998.

[30] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships for Objec t Recognition," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 45-51, July, 1998.

[31] T-L. Liu and D. Geiger, "Approximate Tree Matching and Shape Similarity," *Proc. Intl. Conf. on Computer Vision*, Kerkyra, Greece, 1999.

[32] A. Baumberg, D. Hogg, "Learning Flexible Models from Image Sequences," *Proc. European Conf. on Computer Vision*, pp. 299-308, 1994.

[33] M. Sullivan, C. Richards, C. Smith, O. Masoud, N. Papanikolopoulos, "Pedestrian Tracking from a Stationary Camera Using Active Deformabel Models," *Proc. Intelligent Vehicles*, pp. 90-95, 1995.

[34] S. Ju, M. Black, and Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated Image Motion," *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pp. 38-44, 1996.

[35] I.A. Kakadiaris, D. Metaxas, "Model-Based Estimation of 3D Human Motion with Occlusion Based on Active Multi-Viewpoint Selection," CVPR'96, pp. 81-87, 1996.

[36] D.M.Gavrila and L.S. Davis, "3-D Model Based Tracking of Humans in Action: A Multi-View Approach," CVPR'96, pp. 73-80, 1996.

[37] C. Papageorgiou, T. Evgeniou, T. Poggio, "A Trainable Pedestrian Detection System," *1998 IEEE Int'l Conference on Intelligent Vehicles*, pp. 241-246, 1998.

[38] R. Bowden, T. A. Mitchell, M. Sarhadi, "Reconstructing 3D Pose and Motion from a Single Camera View," *BMVC'98*, pp. 904-913, 1998.

[39] E.J. Ong and S. Gong, "A Dynamic 3D Human Model using Hybrid 2D-3D Representations in Hierarchical PCA Space," *Proc. British Machine Vision Conference (BMVC)*, 1999.

[40] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models - their training and applications," *Computer Vision and Image Understanding*, Vol. 61, No. 2, January 1995.

[41] D.M. Gavrila and V. Philomin, “Real-Time Object Detection for “Smart” Vehicles,” *Int’l Conf. on Computer Vision*, Corfu, Greece, 1999.

[42] F. Mokhtarian and A. K. Mackworth, “A Theory of Multi-Scale, Curvature-Based Shape Representation for Planar Curves,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 14, No. 8, pp. 789-805, 1992.

[43] T.J. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, “Training Models of Shapes from Sets of Examples,” *BMVC*, pp. 9-18, 1992.

[44] G. Chuang and C-C Kuo, “Wavelet Descriptor of Planar Curves: Theory and Applications,” *IEEE Trans. Image Processing*, Vol. 5, pp. 56-70, 1996.

[45] H. Blum, “Biological Shape and Visual Science,” *J. Theor. Biol.*, Vol. 38, pp. 205-287, 1993.

[46] P. Dimit, C. Phillips, K. Siddiqi, “Robust and Efficient Skeletal Graphs,” *CVPR’00*, Hilton Head, South Carolina, June, 2000.

[47] L. Stark, K. Bowyer, “Achieving Generalized Object Recognition through Reasoning about Association of Function to Structure,” PAMI, 13(10), pp. 1097-1104, 1991.

[48] D.D. Hoffman, W.A. Richards, “Parts of Recognition,” *Cognition*, Vol. 18, pp. 65-96, 1984.

[49] D.D. Hoffman, W.A. Richards, “Salience of visual parts,” *Cognition*, Vol. 63, pp. 29-78, 1997.

[50] K. Siddiqi, B. B. Kimia, “Parts of Visual Form: Computational Aspects,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 3, pp. 239-251, 1995.

[51] K. Siddiqi, B. B. Kimia, and K. J. Tresness, “Parts of Visual Form: Psychophysical Aspects,” *Perception*, Vol. 25, No. 4, pp. 399-424, 1996.

[52] M. Singh, G. D. Seyranian, D. D. Hoffman, "Parsing Silhouettes: the Short-Cut Rule," *Perception and Psychophysics*, Vol. 61, No. 4, pp. 636-660, May 1999.

[53] T. Pavlidis, "A Review of Algorithms for Shape Analysis," *Computer Graphics and Image Processing*, Vol. 7, Nov. 2, pp. 243-258, 1978.

[54] L. J. Latecki and R. Lakmper, "Convexity Rule for Shape Decomposition Based on Discrete Contour Evolution," *Computer Vision and Image Understanding (CVIU)*, Vol. 73, No. 3, pp. 441-454, 1999.

[55] R. Malladi and J.A. Sethian, "A Unified Approach for Shape Segmentation, Representation, and Recognition," Report LBL-36069, Lawrence Berkeley Laboratory, University of California, Berkeley, CA, August 1994.

[56] S. Pei and C. Lin, "The Detection of Dominant Points on Digital Curves by Scale Space Filtering," *Pattern Recognition*, No. 25, Vol. 11, pp. 1307-1414, 1992.

[57] J.K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *IEEE Nonrigid and Articulated Motion Workshop*, June 1997.

[58] D. Hogg, "Model-based Vision: a Program to See a Walking Person," *Image and Vision computing*, Vol. 1, No. 1, pp. 5-20, 1983.

[59] K. Rohr, "Towards Model-based Recognition of Human Movements in Image Sequences," *CVGIP: Image Understanding*, Vol. 59, No. 1, pp. 94-115, Jan. 1994.

[60] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, "Pfinder: Real time Tracking of the Human Body," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, Vol. 19, No. 7, pp. 780-785, 1997.

[61] O. Masoud, N. P. Papanikolopoulos, "Robust Pedestrian Tracking Using a Model-Based Approach," *IEEE Conf. on Intelligent Transportation Systems*, pp. 338-343, 1997.

[62] J. Segen, S. Pingali, "A Camera-Based System for Tracking People in Real Time," *Proc. of the 13th Int. Conf. on Pattern Recognition*, pp. 63-67, 1996.

[63] T. Darrell, G. Gordon, M. Harville, J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 601-608, 1998.

[64] C. Wohler, J. K. Aulanf, T. Portner, U. Franke, "A Time Delay Neural Network Algorithm for Real-time Pedestrian Recognition," *International Conference on Intelligent Vehicle*, Germany, 1998.

[65] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps," *CVPR'98*, 1998.

[66] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, Y. Osaki, "Incremental Tracking of Human Actions from Multiple Views," *CVPR'98*, pp. 2-7, 1998.

[67] S. Wachter and H. H. Nagel, "Tracking Persons in Monocular Image Sequence," *Computer Vision and Image Understanding*, Vol. 74, No. 3, pp. 174-192, 1999.

[68] J. Deutscher, B. North, B. Bascle, A. Blake, "Tracking through Singularities and Discontinuities Random Sampling," *ICCV'99*, pp. 1144-1149, 1999.

[69] D.M. Gavrila and V. Philomin, "Real-Time Object Detection for "Smart" Vehicles," *Int'l Conf. on Computer Vision*, Corfu, Greece, 1999.

[70] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, Vol. 73, No. 1, pp. 82-98, 1999.

[71] A. J. Lipton, H. Fujiyoshi, R. S. Patil, "Moving Target Classification and Tracking from Real-Time Video," *Workshop on Applications of Computer Vision*, Princeton, NJ, Oct. 1998.

[72] H. Mori, N. M. Charkari, T. Matsushita, "On-Line Vehicle and Pedestrian Detection Based on Sign Pattern," *IEEE Trans. on Industrial Electronics*, Vol. 41, No. 4, pp. 384-391, Aug. 1994.

[73] S. A. Niyogi, E. H. Adelson, "Analyzing and Recognizing Walking Figures in xyt," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 469-474, 1994.

[74] S. A. Niyogi, E. H. Adelson, "Analyzing Gait with Spatiotemporal Surfaces," *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 64-69, Austin, 1994.

[75] S. Huwer, H. Niemann, "3D Model based Detection and Tracking of People in Monocular Video Sequences," *IASTED Conference on Signal and Image Processing*, pp.172-177, November 2000.

[76] A.M. Elgammal, L.S. Davis, "Probabilistic Framework for Segmenting People Under Occlusion," *Proc. Int'l Conf. on Computer Vision*, 2001.

[77] I.A. Kakadiaris, D. Metsxas and R. Bajcsy, "Active Motion-Based Segmentation of Human Body Outlines," *Workshop on Articulated Motion*, 1994.

[78] I. Haritaoglu, D. Harwood, and L. Davis, "Ghost: A Human Body Part Labeling System Using Silhouettes," *Proc. Int'l Conf. on Pattern Recognition*, pp. 77-82, 1998.

[79] M.K. Leung, Y.H. Yang, "A Model Based Approach to Labeling Human Body Outlines," *Workshop on Articulated Motion*, pp. 57-62, 1994.

[80] I. Haritaoglu, D. Harwood, L. Davis, "W^4—Real Time Detection and Tracking of People and their Parts," *Technical Report*, University of Maryland, Aug. 1997.

[81] S. Iwasawa, "Real-time Estimation of Human Body Posture from Monocular Thermal Images" CVPR'97, pp. 15-20, 1997.

[82] Y. Song, X. Feng, P. Perona, "Towards Detection of Human Motions," CVPR'00, pp. 810-817, 2000.

[83] C. Barron, I. A. Kakadiaris, “Estimating Anthropometry and Pose from a Single Image,” *CVPR’00*, pp. 669 - 676, 2000.

[84] P.J. Besl, N.D. McKay, “A Method for Registration of 3-D Shapes,” *IEEE Trans. on Pattern Recognition and Machine Intelligence*, Vol. 14, No. 2, pp. 239-256, 1992.

[85] Y. Hel-Or and M. Werman, “Constraint Fusion for Recognition and Localization of Articulated Objects”, IJCV 19(1), pp. 5-28, 1996.

[86] T. Cham and J. Rehg, “Dynamic Feature Ordering for Efficient Registration”, ICCV’99, pp 1084-1091, Corfu, Greece, 1999.

[87] P.F. Felzenszwalb and D.P. Huttenlocher, “Efficient Matching of Picturial Structures,”, CVPR’00, pp. 66-75, 2000.

[88] T.J. Cham and J. Rehg, “A Multiple Hypothesis Approach to Figure Tracking,” CVPR’99, 1999.

[89] A. Pentland and B. Horowitz, “Recovering of Non Rigid Motion and Structure,” PAMI, Vol. 2, No. 6, 1980.

[90] R. Rosales and S. Sclaroff, “Inferring Body Pose without Tracking Body Parts,” CVPR’00, pp. 721-727, 2000.

[91] C.J.Taylor, “Reconstruction of Articulated Objects from Point Correspondences in a single Uncalibrated Image,” CVPR’00, pp. 677-684, 2000.

[92] B. Moghaddam, C. Nastar, and A. Pentland, “ A Bayesian Similarity Measure for Direct Image Matching,” *International Conference on Pattern Recognition*, Vienna, Austria. IEEE Computer Society Press, pages 350– 358, 1996.

[93] L. J. Latecki and R. Lakmper, “Shape Similarity Measure Based on Correspondence of Visual Parts,” *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 22, No. 10, October 2000.

[94] S. Abbasi, and F. Mokhtarian, "Shape Similarity Retrieval under Affine Transform: Application to Multi-View Object Representation and Recognition", *Proc. International Conference on Computer Vision*, pp. 450-455, Corfu, Greece, 1999.

[95] M. Pelillo, K. Siddiqi, S. W. Zucker, "Matching Hierarchical Structures Using Association Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No.11,1105-1120, 1999.

[96] R. O. Duda, and P. E. Hart. *Pattern Classification and Scene Analysis*, Wiley-Interscience Publication, John Wiley and Sons, Inc., 1973.

[97] R. Bolles and R. Cain. Recognizing and Locating Partially Visible Objects: the Local-Feature-Focus Method. *International Journal of Robotics Research*, 1(3):57–82, 1982.

[98] D. Mumford, Mathematical Theories of Shape: Do They Model Perception? *SPIE vol. 1570: Geometric Methods in Computer Vision*, 2–10, 1991.

[99] S. Loncaric, A Survey of Shape Analysis Techniques. *Pattern Recognition*, 25:17–23, 1992.

[100] D. Huttenlocher, G. Klanderman and W. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 15(9):850–863, 1993.

[101] R. Basri, L. Costa, D. Geiger and D. Jacobs. Determining the Similarity of Deformable Shapes. *IEEE Workshop on Physics Based Modeling in Computer Vision*, 135–143, 1995.

[102] Y. Lin, J. Dou, and H. Wang. Contour Shape Description Based on an Arch Height Function. *Pattern Recognition*, 25:17–23, 1992.

[103] C. Uras, and A. Verri. Computing Size Functions from Edge Maps. *International Journal of Computer Vision*, 23(2):169–183, 1997.

[104] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem, and J. S. Mitchell. An Efficiently Computable Metric for Comparing Polygonal Shapes. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(3):209–216, March 1991.

[105] K. Yoshida, and H. Sakoe. Online Handwritten Character Recognition for a Personal Computer Systems. *IEEE Transactions on Consumer Electronics*, CE-28(3):202–209, 1982.

[106] C. Tappert, Cursive Script Recognition by Elastic Matching. *IBM Journal of Research Development*, 26(6):765–771, 1982.

[107] W. Tsai and S. Yu. Attributed String Matching with Merging for Shape Recognition. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 7:453–462, 1985.

[108] D. Geiger, A. Gupta, L. Costa and J. Vlontzos. Dynamic Programming for Detecting, Tracking and Matching Deformable Contours. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 17:294–302, 1995.

[109] A. Pope and D. Lowe. Learning Object Recognition Models from Images. *International Conference on Computer Vision*, 296–301, 1993.

[110] M. Koch and R. Kashyap. Using Polygons to Recognize and Locate Partially Occluded Objects. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 9:483–494, 1987.

[111] A.P. Dempster, N.M. Laird, D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM algorithm,” *Journal of the Royal Statistical Society*, Series B, 39(1), 1-38, 1977.

[112] W. Freeman, “Exploiting the Generic Viewpoint Assumption,” Technical Report TR 93-15a, Mitsubishi Electric Research Labs, 1993.

[113] P. Maybeck, *Stochastic Models, Estimation and Control*, Academic Press, New York, 1982.

[114] A. Rosenfeld, "Axial Representation of Shape," *Computer Vision, Graphics, and Image Processing*, Vol. 33, pp. 156-173, 1986.

[115] J. Ponce, "On Characterizing Ribbons and Finding Skewed Symmetries," *Computer Vision, Graphics, and Image Processing*, 52(3), pp. 328-340, 1990.

[116] A.R. Tilley, "The Measure of Man and Woman: Human Factors in Design," H.D. Associates, NY, 1993.

[117] NASA, *Anthropometric Source Book*, NASA Reference Publication 1024, Washington, D.C.

[118] A. M. Gori, F. Scarselli, "Are Multilayer Perceptrons Adequate for Pattern Recognition and Verification?" *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 20(11), pp. 1121-1132, Nov. 98.

[119] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active Contour Models," *Proc. Int'l Conf. on Computer Vision*, pp. 259-268, 1987.

[120] A. Yuille, P.W. Hallinan and D. S. Cohen, "Detecting Facial Features Using Deformable Templates," *Intern. J. on Computer Vision*, Vol. 8, No. 2, pp. 99-112, 1992.

[121] A. Blake and A. Yuille, eds, *Active Vision,* MIT Press, Cambridge, MA, 1992.

[122] K. F. Lai, "Deformable Contours: Modeling, Extraction, Detection and Classification," P h.D. Thesis, Electrical and Computer Engineering Department, University of Wisconsin at Madison, August 1994.

[123] L. Zhao and C. Thorpe, "Stereo and Neural Network-Based Pedestrian Detection," *IEEE Trans. on Intelligent Transportation Systems*, Sept. 2000.

[124] L. Zhao and C. Thorpe, "Qualitative and Quantitative Car Tracking from a Range Image Sequence," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1998.