Statistical Modeling and Localization of Nonrigid and Articulated Shapes

Jiayong Zhang

CMU-RI-TR-06-18

March, 2006

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

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

Thesis Committee:
Robert T. Collins, Co-chair
Yanxi Liu, Co-chair
Takeo Kanade
James M. Rehg, GIT

Keywords: Shape representation, object localization, nonrigid object, articulated object, deformable template, mixture distribution, hierarchical model, sequential Monte Carlo, dynamic programming, hybrid search, gait, pose estimation, human motion.

Abstract

An articulated object can be loosely defined as a structure or mechanical system composed of *links* and *joints*. The human body is a good example of a nonrigid, articulated object. Localizing body shapes in still images remains a fundamental problem in computer vision, with potential applications in surveillance, video editing/annotation, human computer interfaces, and entertainment.

In this thesis, we present a 2D model-based approach to human body localization. We first consider a fixed viewpoint scenario (side-view) by introducing a triangulated model of the nonrigid and articulated body contours. Four types of image cues are combined to relate the model configuration to the observed image, including edge gradient, silhouette, skin color, and region similarity. The model is arranged into a sequential structure, enabling simple yet effective spatial inference through Sequential Monte Carlo (SMC) sampling.

We then extend the system to situations where the viewpoint of the human target is unknown. To accommodate large viewpoint changes, a mixture of view-dependent models is employed. Each model is decomposed based on the concept of parts, with anthropometric constraints and self-occlusion explicitly treated. Inference is done by direct sampling of the posterior mixture, using SMC enhanced with annealing. The fitting method is independent of the number of mixture components, and does not require the preselection of a "correct" viewpoint.

Finally, we return to the generic setting of a single image with arbitrary pose and arbitrary viewpoint. The constraints on the body pose and background subtraction that have been used in previous systems are no longer required. Our proposed solution is a hybrid search facilitated by a 3-level hierarchical decomposition of the model. We first fit a simple tree-structured model defined on a compact landmark set along the body contours by Dynamic Programming (DP). The output is a series of proposal maps that encode the probabilities of partial body configurations. Next, we fit a mixture of view-dependent models by SMC, which handles self-occlusion, anthropometric constraints, and large viewpoint changes. DP and SMC are designed to search in opposite directions such that the DP proposals are utilized effectively to initialize and guide the SMC inference. This hybrid strategy of combining deterministic and stochastic search ensures both the robustness and efficiency of DP, and the accuracy of SMC. Finally, we fit an expanded mixture model with increased landmark density through local optimization.

The models were trained on around 7500 gait images. Extensive tests on cluttered images with varying poses including walking, dancing and various types of sports activities demonstrate the feasibility of the proposed approach.

Acknowledgements

I would like to thank my advisors, Robert Collins and Yanxi Liu, for the opportunity to conduct this research and the constant support, motivation and encouragement throughout the course of this investigation. I would like to thank my committee members, Professors Takeo Kanade and James Rehg, for their invaluable comments.

Thanks to Jiebo Luo and Rodney Miller for the nice intern experience at Kodak.

Thanks to all my friends here at CMU: Jingcao Hu, Jie Yang, Ren Liu, Qifa Ke, Jing Xiao, Changbo Hu, Bing Wu, Yi Chang, Wen-Chieh Lin, Arlene Zhao, Leon Gu, Yan Li, Yanghai Tsin, Kang Li, Leonid Teverovskiy, and many more. They made my life enjoyable.

Thanks also to Suzanne Lyons Muth, Louise Ditmore and Janice Brochetti for their administrative help.

Finally, I would like to thank my family, and my wife Xiaofang, for their love and support through the years.

Contents

List of Figures

List of Tables

Chapter 1

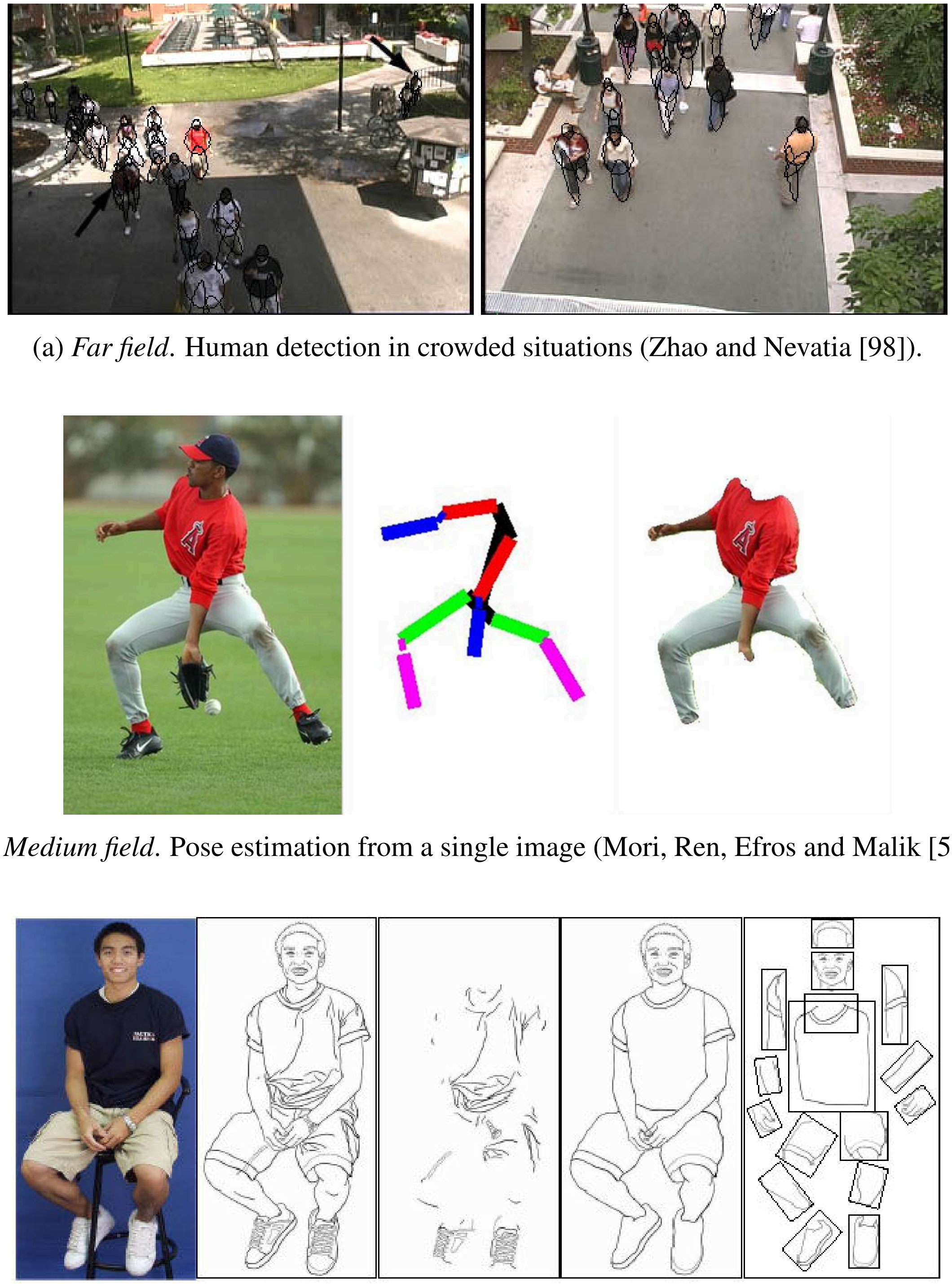
Introduction

An object is articulated if it consists of a set of moving parts (*links*) connecting to each other at certain articulation points (*joints*). The human body is often approximated as a rigid articulated object [16]. When body shapes are studied across subjects, or when clothing and muscle effects become prominent, the rigidity assumption of body parts no longer applies. As a result, each part must be treated as a nonrigid form. Local shape variations of the parts, together with the global articulation, constitute the complete nonrigid deformation of the human body.

1.1 Goal: Parse Pictures of People

This thesis deals with detecting and localizing human bodies and body parts in static images. Human body analysis has a 20-year history in computer vision, yet remains one of the fundamental unsolved problems. This problem has attracted increasing attention from researchers lately. This interest is motivated by a wide spectrum of potential applications, such as surveillance, video editing and annotation, human computer interfaces, entertainment, traffic monitoring, sports, medicine, and image compression.

Human body analysis can be classified into three regimes based on the relative distance between camera and subject (or input resolution). In the first regime ("far" field), targets are typically tens of pixels tall. Although the resolution is low, high-level tasks such as human detection and simple activity recognition are still feasible. As an example, we show in Figure 1.1a two result frames from [98] on human crowd segmentation. The second regime ("medium" field) contains human figures that are an order of magnitude taller, say 200 pixels. Enough image support can be attained to segment and label different body parts such as the head, torso, thighs, calves and arms. One typical task in this regime is to



(a) *Far field.* Human detection in crowded situations (Zhao and Nevatia [98]).

(b) *Medium field.* Pose estimation from a single image (Mori, Ren, Efros and Malik [59]).

(c) *Near field.* Cloth modeling and body sketch (Chen, Xu and Zhu [15]).

Figure 1.1: Example works of static human body analysis in three resolution regimes: far field (a), medium field (b), and near field (c).

estimate the body pose (i.e., a set of 2D or 3D joint angles) from a single image. Another typical task is human identification based on the body shape. In the third regime ("near" field), the input resolution and quality is so high that subtle edge and appearance features become observable. As a result, many otherwise impossible photo-editing tasks can be performed. For instance, Figure 1.1c is excerpted from a recent work on clothes recognition and automatic generation of human sketch [15]. Note that the three-regime classification introduced here is somewhat different from that in [26] for activity recognition, where the far field regime is defined as simple blob targets that cannot be articulated.

In this thesis, we consider the medium resolution regime. Particularly, we focus on localizing the 2D shapes and positions of the body parts. We seek a good summary of both body pose and shape in a given image, while avoiding the ill-posed problem of 3D recovery. We assume that:

1. The torso of the target is approximately parallel to the imaging plane;

2. There is no serious external occlusion.

Furthermore, we do not impose any constraint on the body pose or the viewpoint. No background subtraction (*e.g.,* from video) or depth information (*e.g.,* from stereo) is required. A typical example is shown in Figure 1.2. Note that the complete body boundary shape can

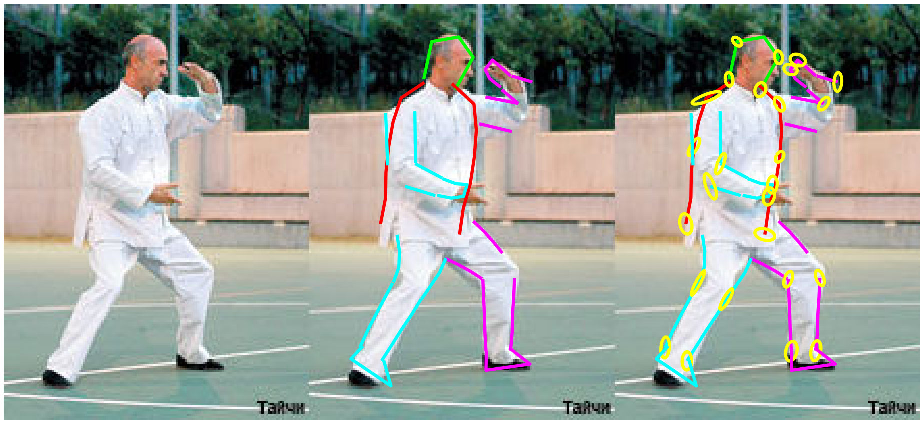


Figure 1.2: Illustration of the thesis goal. Given a single image (left) of a human target, we want to generate a boundary estimate (middle) of each body part, together with the estimated uncertainties (right; shown as error ellipses of selected landmarks on the boundary). The body boundary is partitioned into 14 parts: head, torso, left/right thighs, left/right calves, left/right foot, left/right upper arm, left/right lower arm, and left/right hand.

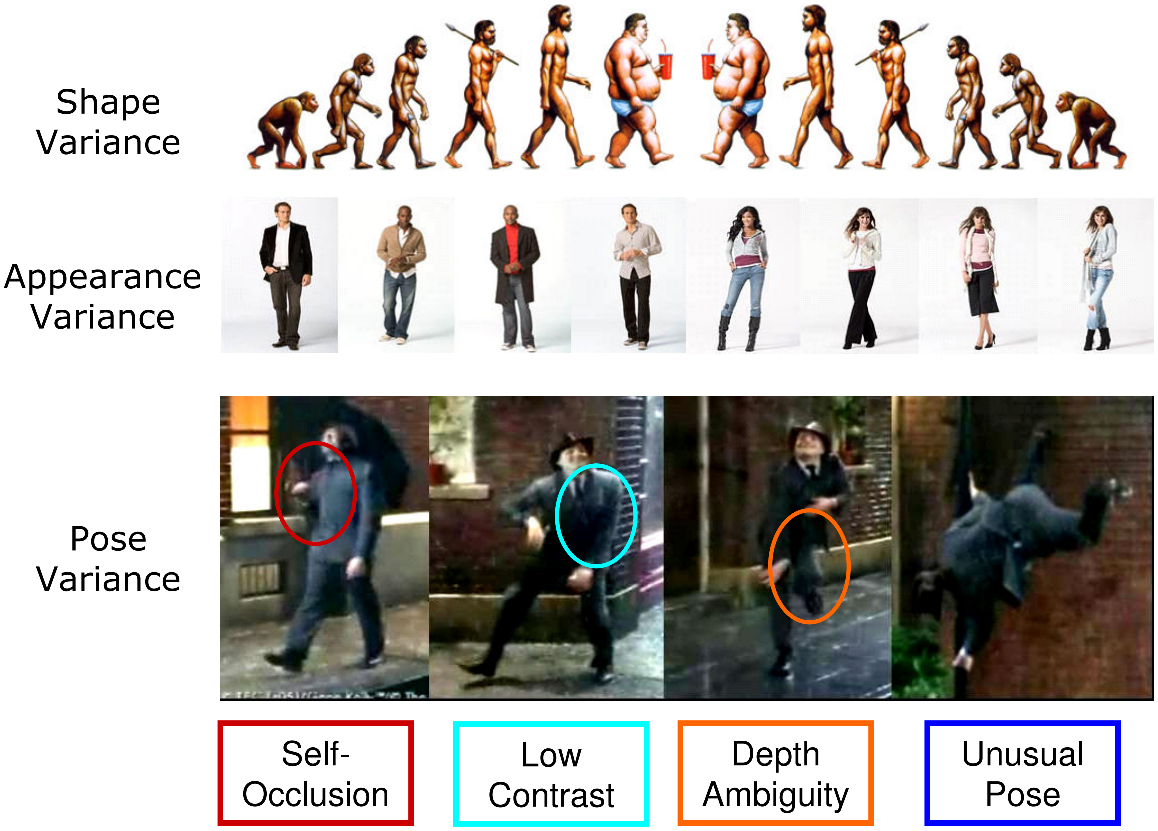


Figure 1.3: Challenges of human body localization in a generic setting: single image, arbitrary pose, and arbitrary viewpoint.

be constructed by stitching together these boundary pieces. Body joints can be localized from the open ends of adjacent parts.

1.2 Challenges

We study the problem of human body localization in a generic setting: single image, arbitrary pose, and arbitrary viewpoint. This is a nontrivial task, even without torso foreshortening and external occlusion, because:

- Body shape may vary dramatically from person to person. Although skeleton structure is stable, it is hidden by muscle and clothing, and thus not directly observable. This difficulty is compounded by articulation;

- Due to the wide variety of color/texture of human clothing and skin, it seems computationally infeasible to obtain a general *a priori* appearance model for people;

- The projection from 3D to 2D results in ambiguity of depth and variation of shape with viewpoint;
- Self-occlusion leads to low contrast observation and feature invisibility;
- All limbs have the same shape of antiparallel lines, also called *apars* (a special case of ribbon). Left limbs and their right counterparts have the same appearance of clothing or skin. The resultant self-similarity causes a serious ambiguity in part labeling;
- Unusual poses are indeed possible, and their probability is much higher than zero (*e.g.,* posters, magazine ads, sports and entertainment fields).

Figure 1.3 is an illustration of the difficulties described above.

1.3 Three Stratified Goals

Given the difficulty of the thesis goal, we have studied three increasingly difficult versions of it. A simple comparison of these stratified versions is given in Table 1.1 and Figure 1.4.

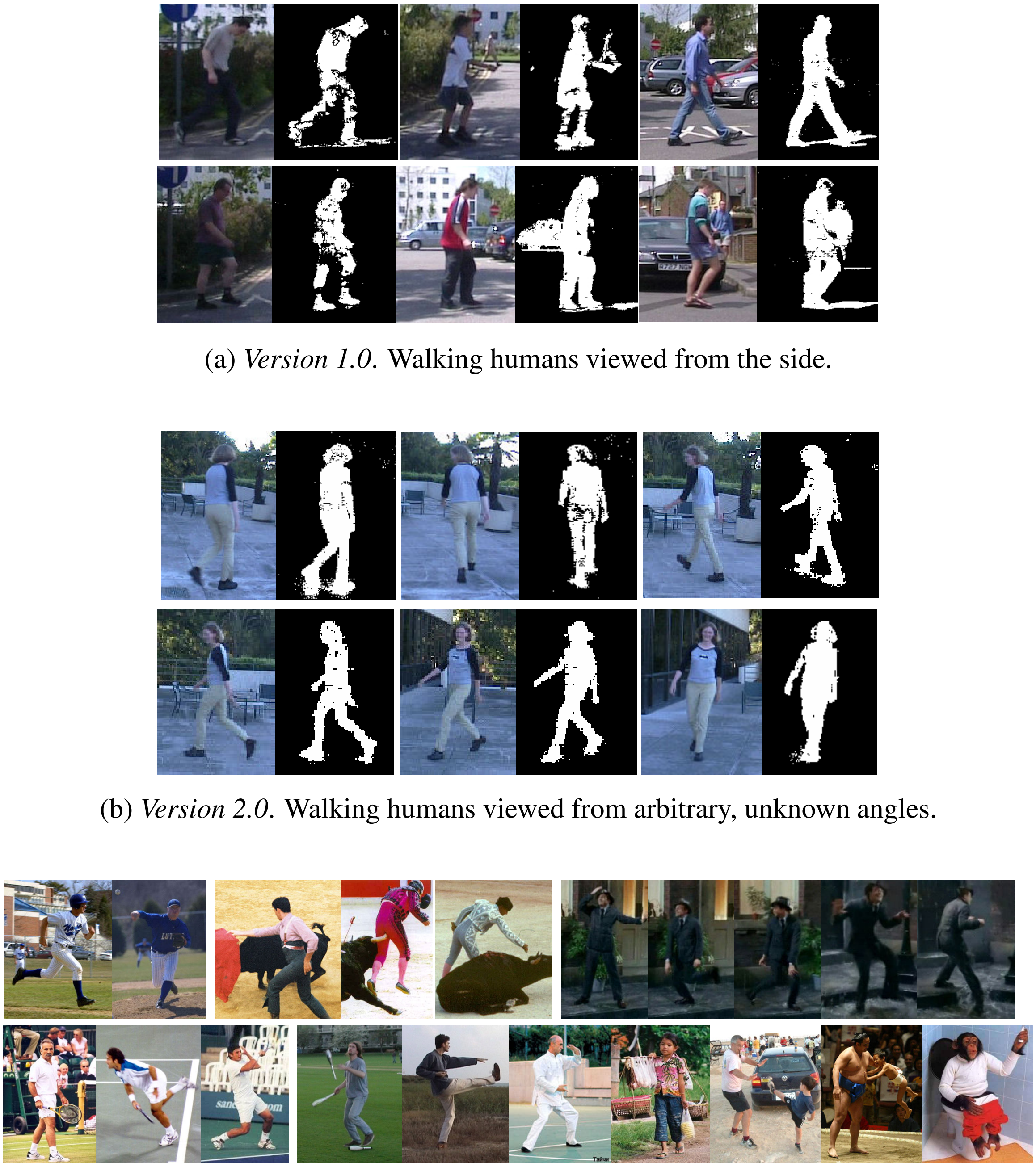
Table 1.1: Comparison of three stratified goals with increasing difficulty.

	Input	Viewpoint	Pose
Goal I (v1.0)	Single Image + Body Silhouette	Side View	Walking
Goal II (v2.0)	Single Image + Body Silhouette	Arbitrary	Walking
Goal III (v3.0)	Single Image	Arbitrary	Arbitrary

We first study a fixed viewpoint scenario by fitting walking humans viewed from the side (Figure 1.4a). The camera is stationary such that background subtraction can be applied. Given the availability of background subtraction and strong pose prior, the localization task can be greatly simplified. Such a scenario, though simple, occurs often in typical surveillance applications.

We then extend the fixed viewpoint system to situations where the viewpoint of the human target is unknown. The main problems to solve are the considerable shape variation and self-occlusion caused by viewpoint changes. An example of such a scenario is a random shot of a person walking in a circle (Figure 1.4b).

We finally remove the requirements of stationary camera and walking activity by handling arbitrary still images with clutter, *e.g.,* from the web or other sources (Figure 1.4c).



(a) *Version 1.0.* Walking humans viewed from the side.

(b) *Version 2.0.* Walking humans viewed from arbitrary, unknown angles.

(c) *Version 3.0.* Still images with varying poses and clutter.

Figure 1.4: Example inputs of the three versions of our proposed system.

The data is diverse and challenging, with poses varying from walking to various sports activities.

Table 1.2: Summary of the representations (*i.e.,* modeling) and inference algorithms (*i.e.,* localization) that have been used in this thesis.

	Chap. 3	Chap. 4	Chap. 5	Note
	v1.0	v2.0	v3.0	
Triangle-based Model	●			
Part-based Model		●	●	
Mixture Model		●	●	With part-based models as components.
Tree-structured Model			●	
Boundary Model			●	
Hierarchical Model			●	Combining tree-structured model, mixture model, and detailed boundary model.
Sequential Monte Carlo (SMC)	●	●	●	
Markov Chain Monte Carlo (MCMC)		●	●	
Dynamic Programming (DP)			●	
Reweighted SMC			●	
Local Optimization			●	
Hybrid Search			●	Combining DP, reweighted SMC, and local optimization.

1.4 Outline of the Thesis

We take a Bayesian approach to deformable template matching. The central component is a statistical landmark-based representation (*i.e.,* modeling) of the nonrigid and articulated body contours. Several image cues are combined to relate the body configuration to the observed image. The model is arranged into a sequential structure, enabling simple yet effective spatial inference (*i.e.,* localization) through stochastic and/or deterministic sequential search.

The remainder of the thesis is arranged as follows. In Chapter 2, we briefly review previous works on human body analysis in either static images or video sequences. Chapters 3 through 5 constitute the main technical part of the thesis. Three fitting algorithms are presented, targeting the three stratified goals respectively [95, 96, 97]. Each algorithm is developed by making substantial improvements to its predecessor. During the system "upgrades", a number of variants of the representations and inference algorithms are introduced (see Table 1.2 for a summary). Finally, we highlight the main contributions of this work in Chapter 6, and conclude with a discussion of our insights and possible future work.

Chapter 2

Background

The problem of human body analysis has a 20-year history in computer vision (*c.f.,* the work of Marr *et al.* [53] and Hogg [35]), yet remains one of the fundamental unsolved problems. In this chapter, we briefly review existing work on this topic, with a focus on still image analysis and more recent developments. Interested readers should refer to [3, 12, 29, 55] for surveys on earlier work, which is mostly on tracking and recognition of human motion. Our review is divided into four sections:

- Model-based vs Image-based
- Top-down vs Bottom-up
- 2D vs 3D
- Spatial vs Temporal

2.1 Model-based vs Image-based

The model-based approach assumes an explicit parametric model of the human body, and the best configuration is determined based on how well it predicts the observed image. When the motion is complex, multiple parametric models can be used [48]. Taking this idea further, every training example may be treated as a separate deformable model (or exemplar) [58, 85]. In general, model-based methods are computationally expensive, and may be easily trapped in local minima. As a pay-off, different effects such as articulated motion, anthropometric deformation, illumination and occlusion can be delineated and studied individually.

Image-based methods, sometimes called learning-based, aim at recovering body pose without extracting body parts. This is formulated as a high-dimensional regression problem, *i.e.,* to determine a mapping from the image (or image descriptor) space to the body configuration space. The solutions share a common architecture of a front end, which extracts features from the image and represents them as vectors in a high-dimensional space, and a regression engine. Features that have been used include Hu moments of silhouette images [70], concatenated coordinates of sampled boundary points [32], multi-scale edge direction histograms [73], distribution of shape contexts evaluated at sampled boundary points [1], and Harr-like features selected by AdaBoost [66, 90]. Example regression engines include robust Local Weighted Regression [73], BoostMap [4], perceptron mapping [70], and Relevance Vector Machine [1, 84].

Image-based methods are appealing because proven statistical learning techniques can be easily applied. With some care, they also can be made fast (in the test mode) and suitable for real-time applications. One weakness of this approach is that the ability to accurately represent the space of realizable shapes depends almost exclusively on the amount and representativeness of the training data. In fact, many works of this type use synthesized training examples from a motion capture database. In addition, most existing implementations use features extracted solely from silhouette images, and do not recover anthropometric information.

2.2 Top-down vs Bottom-up

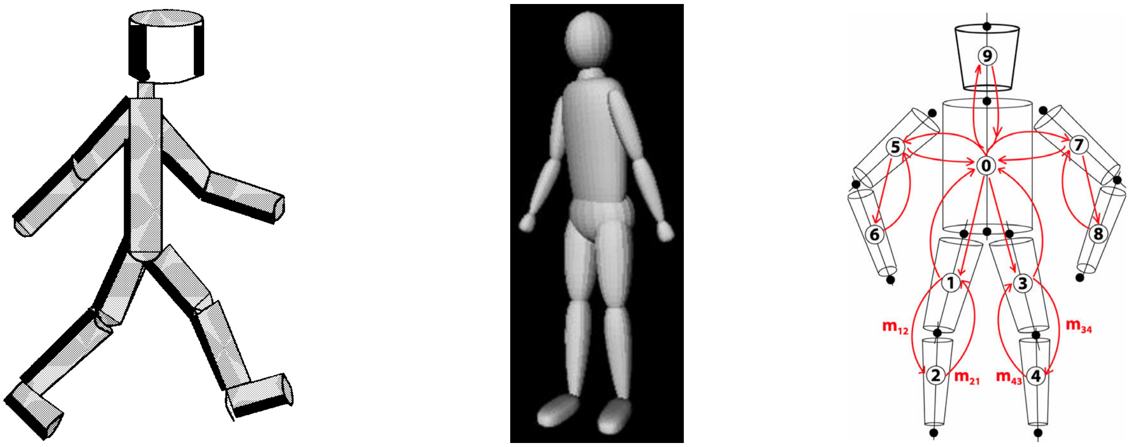
Traditional vision research emphasizes top-down recognition and tracking [53]. A top-down method directly explores a high-dimensional configuration space in order to optimize a complex objective function that measures the similarity between predicted and actual views. As an alternative, bottom-up methods offer the promise of significantly reduced search cost.

Most bottom-up methods assume “weak” models, where each body part is represented by a single rectangle or feature point, and the connections between parts are loose [38, 57, 69, 79]. To proceed, a candidate list of body parts is first detected. These candidates are then pruned and assembled into the best configuration with the guidance of global geometric constraints. Many good head detectors exist [94], and limb detectors have been built based on point feature tracking [79], template matching [37], hierarchical grouping of parallel edge elements [38, 67], probabilistic region similarity [68], image segmentation [59, 80], edgelet [92], and appearance based detectors using SVM [57, 69] and AdaBoost [54, 92].

The bottom-up approach highlights a simple and flexible structure. Therefore it often targets high-level tasks such as human detection. However, building a robust part detector is difficult in practice, and for complex problems it is unrealistic to assume that the entire recognition problem can be solved in a purely feed-forward fashion.

Recent attempts have been made to combine the top-down and bottom-up search strategies. Examples include: iteration between part detection and temporal pruning [64], the use of bottom-up proposals in Data-Driven MCMC [50], and the use of a stratified sampler in Nonparametric Belief Propagation [76]. In spite of these efforts, a proper balance of top-down and bottom-up processing remains to be defined.

2.3 2D vs 3D



(a) Elliptical Cylinders (b) Enhanced Ellipsoids (c) Loose Limbed

Figure 2.1: Selected volumetric representations. (a) Cylinder model used in the work of Hogg [35] to generate 3D description of a walking human. Originated by Marr and Nishihara [53], each part is defined by 3 shape parameters. Relative position of parts is determined by geometric transformations in embedded coordinate systems. (b) Sminchisescu and Triggs' model [78] for monocular body tracking. The limb is built from superquadric ellipsoids with additional tapering and bending parameters [5]. The model has around 30-35 joint parameters, plus 8 internal proportion parameters, plus 9 deformable shape parameters for each body part. (c) Loose-limbed model used by Sigal *et al.* [76], which can be considered as the 3D version of pictorial structure. Each part is modeled by a tapered cylinder with 5 shape and 6 pose parameters.

A persistent debate exists over the use of object centered models, such as representations of the objects' 3D structures in a coordinate frame independent of the viewing parameters [53]. Because 3D models have the attractive feature of leading to viewpoint independence, they have garnered much of the research effort in human motion analysis.

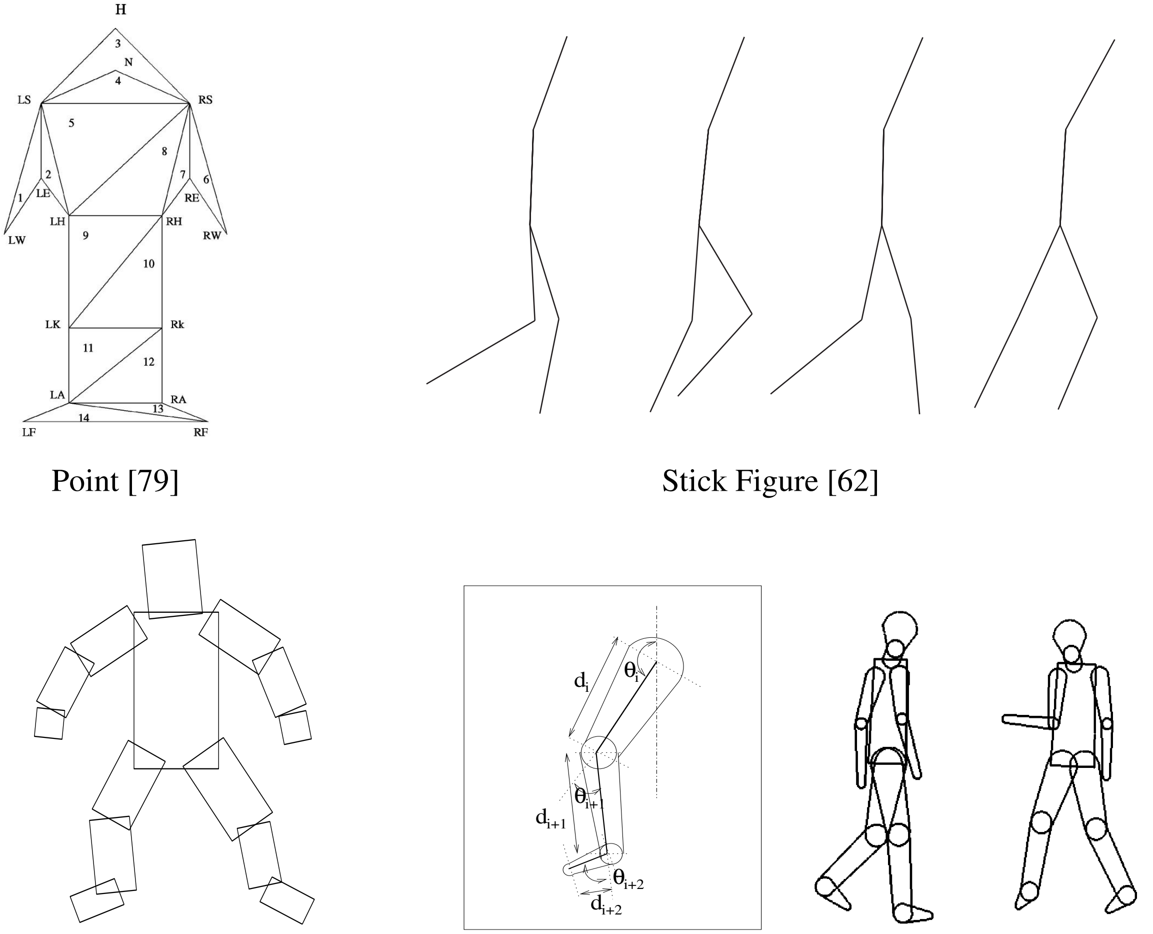
The simplest 3D representation of a human body is the stick figure, which consists of line segments connected at their endpoints (joints). Stick figure models can be described using only a few parameters, *i.e.,* the 3D position of each joint, assuming the connectivity is known in advance. However, the extraction of a stick figure from real images is rather difficult due to the lack of any shape model. This problem is avoided in the case where the trajectory of each joint is given, such as in Moving Light Displays (MLD) studies [11, 41].

Volumetric models are expected to better represent the complexity of the human body. They are built around the stick figure by fleshing out its line segments. The "flesh" is often modeled using the class of tapered super-quadrics [30], including cylinders, spheres, ellipsoids, and hyper-rectangles (see Figure 2.1 for some examples). The cost of better representation is an increase in the number of parameters in order to describe the part and the associated deformation, as well as resultant issues like body part collision.

In contrast to object centered representation, 2D approaches directly model the projection of the human figure in images. This avoids 3D ambiguity while still capturing natural degrees of freedom. Specifically, a body projection is modeled as a collection of 2D links with or without a depth ordering. Commonly used link representations include points, line segments, rectangles, ribbons or blobs, or rounded trapezoids (see Figure 2.2 for examples). Since each link in the 2D model typically describes the projected image appearance of a corresponding rigid link in a 3D kinematic model, these approaches are, by necessity, viewpoint-specific.

We may note that all the articulated models discussed above look to some extent "unnatural", robotic, or at best, humanoid. Some unique anthropometric properties of the human subject under study are ignored or improperly modeled. This is not a problem if the anthropometric deformation is small and therefore can be well accommodated in the matching process. However, when body shapes are studied across subjects, the problem becomes evident due to the dramatic shape variance from person to person. In this case, it is desirable to have a new representation that better delineates and captures natural human body variance while preserving compactness, high-level interpretability and computational simplicity of the model.

Conventionally, articulated motion is studied as an independent topic. On the other hand, there has been a rich body of research on modeling arbitrary deformable shapes. Important to mention here are the theory and practice of the statistical analysis of shapes



Point [79]

Stick Figure [62]

Pictorial Structure [28, 69]

Scaled Prismatic Model [2, 60]

Figure 2.2: Typical 2D representations of body structure.

developed by Kendell [43], Bookstein [8], Dryden and Mardia [25], Cootes, Taylor and colleagues [18]. These works focus on the situation where the objects are summarized by a set of key points called landmarks. They are well known in the computer vision community for their use in Active Shape Model (ASM) and Active Appearance Model (AAM), where shape variability is learned through labeled training examples by applying PCA to Procrustes residuals. Another method that is closely related to our work is the polygon representation proposed by Felzenszwalb [27]. Using the constrained Delaunay triangulation, this method has an attractive property that the globally optimal match of a model to the image can be found via Dynamic Programming (DP), since the dual graph representation of a triangulated polygon is a tree. However, the author only discussed simple polygons without self-intersections, and the use of DP constrained the definition of energy terms to two or three neighboring vertices. Also, due to the computational cost of

DP, the discretization resolution of the configuration space is largely limited.

2.4 Spatial vs Temporal

Human body analysis is fundamentally a problem of reasoning under uncertainty. Recent research in this area is dominated by Bayesian statistical inference. However, optimizing the posterior in a high dimensional configuration space is intrinsically difficult. There are three main types of search strategies: *gradient descent* incrementally improves an existing estimate, *regular sampling* evaluates the cost function at predefined points in the configuration space, and *stochastic sampling* generates random sampling points according to some proposal distribution that indicates good places to look. Whichever strategy is used, effective focusing is the key to high-dimensional search.

Most work on articulated human body fitting focuses on temporal tracking through video sequences (e.g., [13, 22, 76, 78, 88]). In this case, search is constrained by the strong prior propagated from the past and/or the future through temporal dynamics. Dynamic models of body motion vary in complexity; ranging from a simple random walk, to constant velocity, to nonlinear models learned from training examples such as Switching Linear Dynamic System [62], mixtures of autoregressive processes [2], motion graph [75], and Scaled Gaussian Process Latent Variable Models [87].

Systems that track 3D kinematic body models are often brittle because the likelihood surface relating a high degree of freedom 3D articulated body model to 2D body shape in an image is fraught with local minima [77]. Given the complexity of the likelihood, Monte Carlo sampling techniques for representing the posterior distribution demonstrate the most promising results [22, 49, 74, 78]. Even then, robust fitting is typically achieved only by imposing additional information, such as the use of multiple simultaneous views [22, 49], or strong constraints on the temporal dynamics [74].

One alternative is to track a 2D articulated body model instead, in the hope that the likelihood surface will be better behaved [10, 13, 42]. Nonetheless, the degrees of freedom left in the projected model are still high enough that gradient descent tracking [10, 42] needs a good initial pose estimate and small inter-frame motion. Methods that recognize that the solution space is multi-modal [13], particularly in the presence of background clutter, seem to be the most promising.

Over the last decade, there has been a large number of papers in computer vision on SMC (or particle filters) and their applications. Basic particle filters may not work well for complex problems like body tracking, and many variations and improvements can be found in the literature [24]. Recently, Nonparametric Belief Propagation was proposed to generalize particle filters to arbitrary graphs with pair-wise formulations rather than a simple chain [39, 81]. The algorithm was used by Sigal *et al.* in loose-limbed body tracking [76].

Spatial body fitting typically handles static images without a dynamic model. This is desired in situations where only a single image is available, or when we need to automatically initialize/re-initialize an online body tracker. Spatial fitting relies purely on kinematic constraints, and represents an important component in a successful tracking system.

Most bottom-up and learning-based methods work in static mode. There has also been a significant amount of research into the registration of nonrigid objects. However, only a few have addressed the problem of fitting articulated body models to static images [28, 50, 58]. In the work of Mori and Malik [58], a large number of models were stored. Each model (exemplar) was represented by edge pixels sampled from the body contour. Model fitting was then posed as a point-set matching problem, which was solved using shape context descriptors. Felzenszwalb and Huttenlocher [28] consider the problem of fitting a pictorial structure to a background subtraction mask. They showed that, for some restricted form of deformation cost, the global optimal match of the structure could be found efficiently via Viterbi recurrence over a standard discretization of the configuration space. The recent work by Lee and Cohen [50] attempts to fit a volumetric 3D model to static 2D images. They employed Data-Driven MCMC to find the MAP solution. Various information sources such as face detection, color segmentation, curve fitting, blob and ridge detection are used to form better proposals to facilitate the MCMC search.

SMC is used quite often to perform inference over a temporal chain of poses [24]. Different from this common trend, in this thesis, we apply SMC over a *spatial* chain for shape fitting. In fact, the earliest application of SMC was in the spatial domain, *i.e.,* the computer simulation of a long-chain polymer on a d-dimensional lattice space [34, 45]. Similar ideas have been used by Perez and colleagues [63] to apply particle filters in the problem of interactive contour extraction. Ioffe and Forsyth [38] used importance sampling to incrementally update a set of candidate assemblies. MacCormick and Isard [52] proposed partitioned sampling to track articulated objects, which is in essence a Monte Carlo smoother in the spatial domain.

Chapter 3

Body Fitting Using Sequential Monte Carlo

3.1 Introduction

In this chapter, we address the first and simplest subgoal of the thesis, *i.e.,* to simultaneously locate body parts and associated shape boundaries of walking humans viewed from the side (Figure 1.4a). The camera is stationary such that background subtraction can be applied. However, even in such a simplified scenario, accurate human body extraction is a nontrivial task, due to large variation in observed body shapes caused by articulated motion, anthropometric body variation, and clothing.

Two questions immediately arise. Firstly, why do we want to find the detailed shape boundary? Conventionally, body parts are approximated by straight lines, 2D rectangles or blobs, or generalized 3D cylinders. The focus of these models is to estimate the body pose (i.e., a set of joint angles), even though they are crude to the eye, or robot-like. For some applications, body pose is the only information required, and the unique body shape of the subject under study can be ignored (e.g., human motion control [66]). In some cases, however, shape information plays an important role (e.g., gait identification [89]). For this reason, we propose a joint encoding of both shape and pose, which can provide discriminative cues for human identification from gaits, or can be used to initialize a kinematic body tracker for activity analysis. A second argument is that an appropriate model of the boundary deformation can help localize body parts. Decoupling geometric deformation from appearance variation is one of the key issues in the class of methods called *deformable templates* [40]. A good example of deformable template is the Active Appearance Model [17], which has been proven to be successful for face image interpretation.

Secondly, how detailed should the shape boundary be? An ideal case is to recover the body part labels in pixel or subpixel resolution, which is desirable in many graphics applications. Unfortunately this remains an elusive goal without high input resolution and user interaction. In this work, we choose to model the body shape by a set of piecewise linear boundary curves, which reside locally in a low-dimensional space. The benefit is to represent the body shape and characterize its deformation with limited representational overhead. We believe this piecewise linear representation attains a reasonable equilibrium between its modeling ability and simplicity, and at the same time provides an important intermediate step for more advanced needs.

3.2 Overview of the Approach

We take a 2D model-based approach to fitting walking humans viewed from the side (Figure 3.1). The body shape is represented by a set of landmarks along the boundary curves.

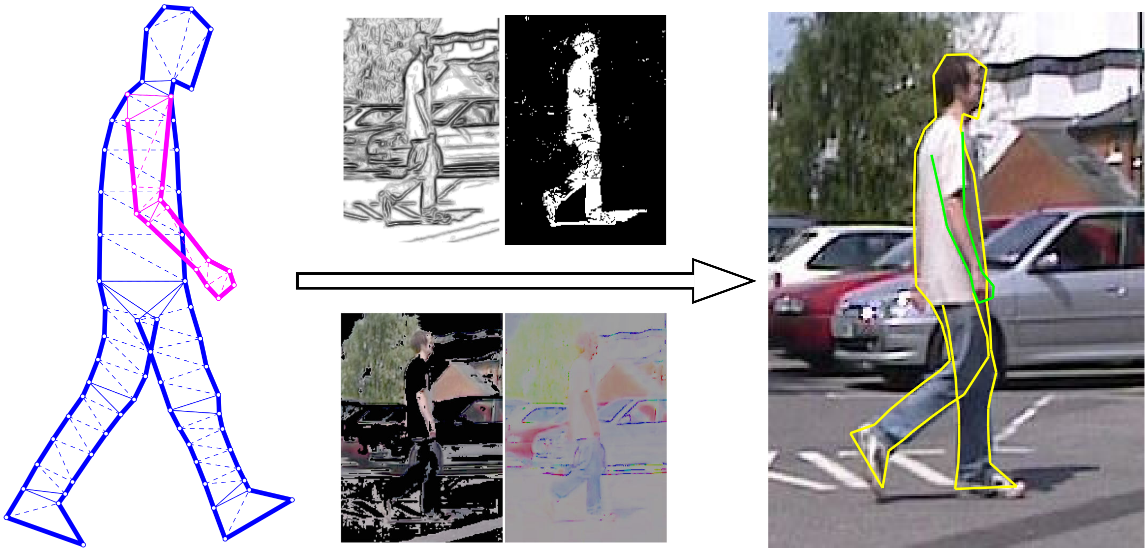


Figure 3.1: Overview of our approach. A nonrigid, articulated contour model (left) and local image cues (middle) are combined via Bayes formulation. The model is fit using Sequential Monte Carlo to a sample image (right) taken in a cluttered, outdoor scene.

The deformation of the model is constrained by the joint probabilistic distribution of landmark positions. To simultaneously accommodate anthropometric deformation and articulated motion, this distribution is inevitably complex and highly nonlinear. We apply graphical models to the shape representation to factor the joint distribution of all landmarks into a series of marginal and conditional distributions. Each of these distributions is specified according to one of two deformation mechanisms: local nonrigid deformation, and rotation motion of each joint.

We formulate the shape model matching to the observed image in a Bayesian framework. The likelihood is computed from several cues, including edge gradient, silhouette, skin color and region similarity. Due to the high degree of freedom of the model, optimizing the posterior is intrinsically difficult. Therefore, we impose a spatially sequential structure on the model. This sequential arrangement enables us to expand the configuration space and collect image information incrementally using Sequential Monte Carlo (SMC) sampling.

The proposed approach differs from conventional body localization methods in several aspects. First, we study the body shape across subjects. To this end, the model prior is learned from a large number of real gait images that have been automatically labeled by bootstrapping. Second, our model is designed to capture detailed body boundaries. Third, the posterior of model-to-image matching is decomposed in such a way that spatial inference can be performed effectively via SMC sampling. It is important to note that we are using SMC to perform inference over a spatial chain for shape fitting, rather than over a temporal chain of varying poses.

3.3 Bayesian Formulation

We adopt a Bayesian approach to deformable template matching, as conveyed by the formula,

$$p(\Omega|\mathcal{I}) \sim p(\Omega)\, p(\mathcal{I}|\Omega), \tag{3.1}$$

where Ω denotes a configuration of the model, and $\mathcal{I}$ denotes an input image. The shape prior $p(\Omega)$ encodes our knowledge of possible shape deformations, while the imaging likelihood $p(\mathcal{I}|\Omega)$ measures how compatible a given model configuration is with respect to observed image features. The desired model-to-image matching is then found by searching for a configuration $\hat{\Omega}$ that maximizes the posterior probability, or by sampling the posterior at random. In this section, we discuss the parameterization, shape prior and imaging likelihood cues. Section 3.4 presents our sequential Monte Carlo approach to find the desired model-to-image fitting.

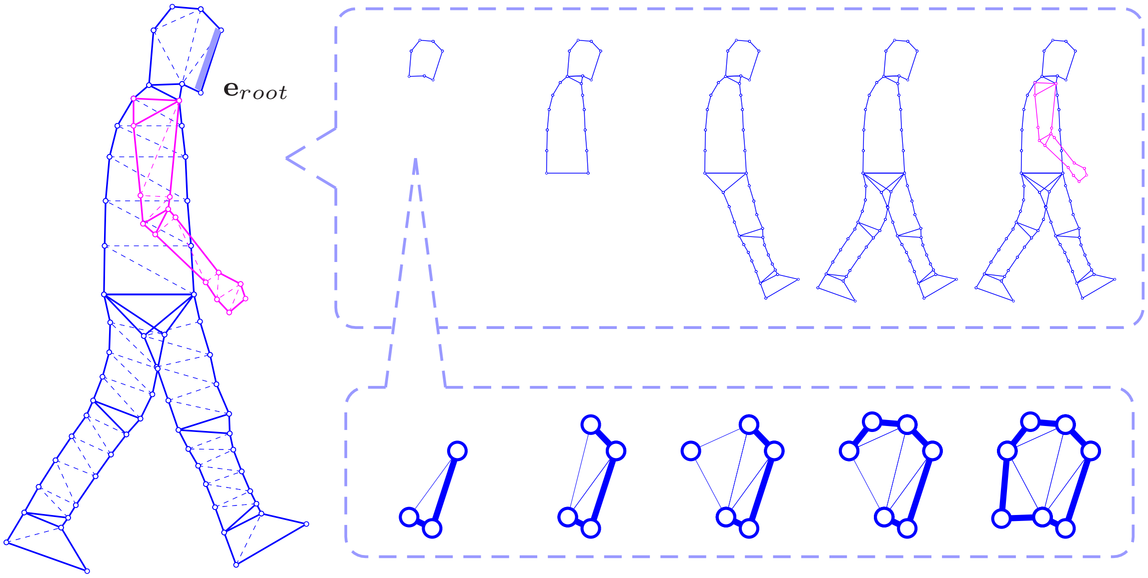


Figure 3.2: Shape triangulation specifying the growing order of vertices. Given root edge $\mathbf{e}_{root}$, the shape is constructed sequentially by growing one triangle (vertex) at a time. Note that this is not the connectivity graph of the shape prior. Only one, non-occluded arm is modeled.

3.3.1 Landmark-based Representation

We represent a body shape by a set of nonrigid boundary curves, as depicted in Figure 3.1. These curves are assumed to be piecewise linear, and thus are completely described by a set of K landmarks $\mathbf{v}_{1:K} = \{\mathbf{v}_k\}_{k=1}^K$. The 2D coordinates of these landmarks, $\{(x_k, y_k)\}_{k=1}^K$, specify the configuration $\Omega \in \Re^{2K}$ of our body model.

Such a landmark-based shape representation is not new in computer vision. For example, it has been used in Active Shape Models (ASM) and Active Appearance Models (AAM) for face image interpretation (e.g., [18]). A common practice in these methods is to first remove the Euclidean similarity transformations (translation, rotation and scaling) and then model the shape residuals using some low dimensional linear model. However, direct application of this global analysis to the body shape is difficult, because the articulated motions of body parts are so large and independent that the shape residuals no longer reside in a low dimensional linear subspace.

Instead, we apply graphical modeling to the shape representation to factor the joint distribution of all landmarks $p(\Omega)$ into a series of marginal and conditional distributions. This is intuitive since the human body represents a typical disaggregated structure. To this end, we specify a growing order of vertices by triangulation, as depicted in Figure 3.2. This is inspired by the work of Felzenszwalb [27], where the author applied the constrained Delaunay triangulation [7] to single polygons. In our case, the landmark positions are selected by hand, and distributed almost uniformly along the external boundary sides of each body part. The triangulation is designed such that,

- The shape can be constructed sequentially by growing one landmark or triangle at a time (Figure 3.2);
- Each landmark, say $\mathbf{v}_k$, is connected to an unique parent edge, say $(\mathbf{v}_i, \mathbf{v}_j)$, $1 \leq i < j < k$.

In this way, the model is arranged into a sequential chain-like structure. This sequential arrangement enables us to expand the configuration space and collect image information incrementally, which is essential to our sampling algorithm described below.

3.3.2 Shape Prior

The shape prior knowledge is encoded by the joint density distribution of the locations of K landmarks, *i.e.,* $p(x_1, y_1, \dots, x_K, y_K)$, or equivalently $p(\mathbf{v}_{1:K})$, where $\mathbf{v}_k = (x_k, y_k)$. Given the fixed landmark ordering, this joint distribution can be expanded as,

$$p(\mathbf{v}_{1:K}) = p(\mathbf{v}_1, \mathbf{v}_2) \prod_{k=3}^{K} p(\mathbf{v}_k | \mathbf{v}_{1:k-1}) \,. \tag{3.2}$$

We do not model any preference over the absolute location, scale or orientation of the human target in the image. This means that $p(\mathbf{v}_1, \mathbf{v}_2)$ is a constant. For simplicity, we let $p(\mathbf{v}_1, \mathbf{v}_2) = 1$. The prior distribution is then given by,

$$p(\mathbf{v}_{1:K}) = \prod_{k=3}^{K} p(\mathbf{v}_k | \mathbf{v}_{1:k-1}) \,. \tag{3.3}$$

Note that $p(\mathbf{v}_{1:K})$ is an improper prior since its integral is infinite [6].

To further specify the complete conditional $p(\mathbf{v}_k | \mathbf{v}_{1:k-1})$, we introduce two types of deformation mechanisms. The first type is designed to model rotation motion of the joints. We select nine joint triangles (Figure 3.3), with the index set denoted as $\mathcal{J}$, corresponding to $\{$neck, shoulder, elbow, left/right hip, left/right knee, left/right ankle$\}$. These joint triangles divide the body shape into ten parts. For each $k \in \mathcal{J}$, $\mathbf{v}_k$ is connected to a unique *parent* edge, $\mathbf{e}_k^P = (\mathbf{v}_i, \mathbf{v}_j)$, and is predicted by perturbing $\mathbf{v}_j$ with (ρ_k, θ_k) in the local polar coordinates determined by $\vec{\mathbf{e}}_k^P$.

$$\mathbf{v}_k = \rho_k \cdot Rot(\theta_k) \cdot (\mathbf{v}_j - \mathbf{v}_i) + \mathbf{v}_i \,. \tag{3.4}$$

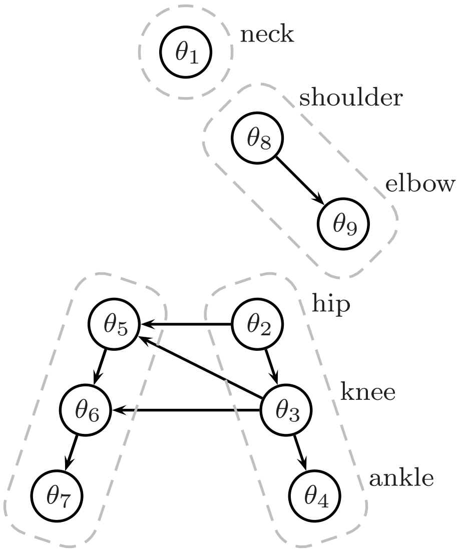


Figure 3.3: Graph structure specifying ordering and dependency relations among 9 joint angles Θ.

Although it seems safe to assume that the local lengths ρ_k are independent, we cannot ignore the long range dependencies among joint angles, $\Theta = \{\theta_k : k \in \mathcal{J}\}$. Therefore another Bayes network capturing body topology is designed to model $p(\Theta)$ (Figure 3.3).

The second type of mechanism models local non-rigid deformation. For each landmark within the body parts, say $\mathbf{v}_k$, we specify a parent triangle $\mathbf{t}_k^P$. We then assume the Markov property,

$$p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{t}_k^P), \tag{3.5}$$

which implies that the position of $\mathbf{v}_k$ can be completely predicted from it's parent triangle $\mathbf{t}_k^P$. Our prediction method uses an affine transformation in the local landmark coordinate system:

$$\mathbf{v}_k = (A_k \cdot \bar{\mathbf{v}}_k + b_k) + \mathbf{n}_k\,, \tag{3.6}$$

where $\bar{\mathbf{v}}_k$ is the reference position of the k-th landmark. To predict the position of $\mathbf{v}_k$, the reference landmark $\bar{\mathbf{v}}_k$ goes through a linear transformation A_k followed by a shift b_k, and then perturbed by noise $\mathbf{n}_k$. Note that the conditioning variables $\mathbf{t}_k^P$ are implicitly encoded in A_k and b_k. (A_k, b_k) is determined by either 1) the affine transformation from the triangle $\bar{\mathbf{t}}_k^P$ in the reference model to the triangle $\mathbf{t}_k^P$ fit previously to the data, or 2) the similarity transform from the reference edge $\bar{\mathbf{e}}_k^P$ to the fitted edge $\mathbf{e}_k^P$. The latter is used for the first triangle of each body part, whose parent is a joint triangle. The noise term $\mathbf{n}_k = (n_k^x, n_k^y)$ is applied in the local Cartesian coordinates determined by $\vec{\mathbf{e}}_k^P$.

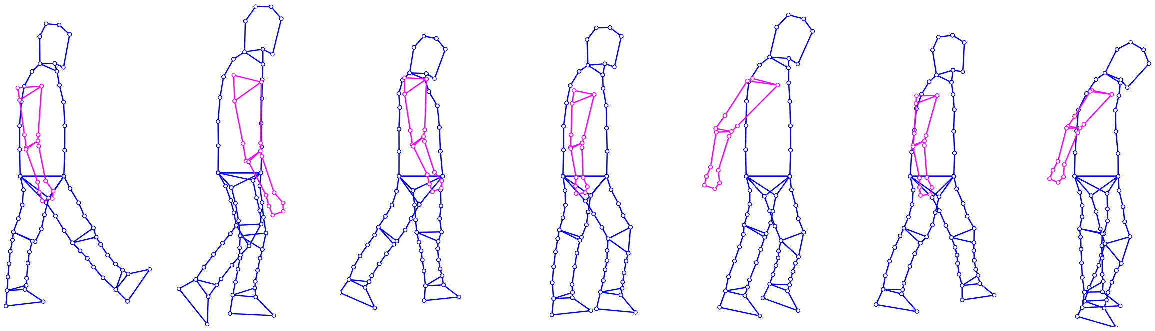


Figure 3.4: Selected random samples from the learned shape prior $p(\Omega)$. Each shape is normalized by aligning the bottom edge of the torso with line segment $(0,0)(1,0)$.

Using the deformation mechanisms described above, a complete sample shape can be sequentially constructed starting from a given position, scale and orientation of the root triangle $\mathbf{t}_{root} = \mathbf{v}_{0:2}$, which is defined on the face in our shape model (Figure 3.2). At each step, a vertex sample $\tilde{\mathbf{v}}_k$ is generated according to either (3.4) or (3.6), depending on whether the current triangle is a joint or body triangle.

To summarize, the shape prior can be formulated as

$$p(\mathbf{v}_{1:K}) = p(\Theta) \prod_{k \in \mathcal{J}} p(\rho_k) \prod_{k \notin \mathcal{J}} p(\mathbf{n}_k), \qquad (3.7)$$

Note that the proposed model is translation invariant because $p(\mathbf{v}_{1:K})$ involves no absolute landmark positions. By expressing $\mathbf{n}_k$ in the local coordinate system of $\vec{\mathbf{e}}_k^P$, the model is also made rotation and scale invariant.

We estimate the densities $p(\mathbf{n}_k)$, $p(\rho_k)$ and $p(\Theta)$ in equation (3.7) from a set of training images. The details are described in Section 3.5.2. Figure 3.4 shows several samples randomly drawn from the learned shape prior. Note that only one arm is modeled.

3.3.3 Imaging Likelihood

Let $\Lambda = \{(i,j)\}$ be the image lattice associated with the image $\mathcal{I}$, and let $\mathcal{I}_R$ denote the image patch defined on a region $R \subset \Lambda$. As depicted in Figure 3.2, the sequential structure of the model insures that each vertex $\mathbf{v}_k$ is connected to a unique parent edge $(\mathbf{v}_i, \mathbf{v}_j)$. $\mathbf{v}_k$ and $(\mathbf{v}_i, \mathbf{v}_j)$ specify a triangle $\mathbf{v}_{T_k} = (\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k)$, where $T_k = (i,j,k)$, and the associated region R_k. These triangles partition the image into two areas: the body foreground, $R_{FG} = \cup_k R_k$, and the background, $R_{BG} = \cap_k \overline{R_k}$.

Similar to the prior, we seek a marginal and conditional decomposition of the likelihood. We start from the simplest case. Suppose,

1. There is no overlap between foreground regions;

2. $\mathcal{I}_{R_k}$ is an independent realization from a probabilistic foreground model $p(\mathcal{I}_{R_k}|FG)$;

3. $\mathcal{I}_{R_{BG}}$ is an independent realization from a background model $p(\mathcal{I}_{R_{BG}}|BG)$;

4. The probability to observe $\mathcal{I}$ given the background model is irrelevant to image partitioning.

The likelihood simplifies to,

$$p(\mathcal{I}|\Omega) \propto \prod_k \frac{p(\mathcal{I}_{R_k}|FG)}{p(\mathcal{I}_{R_k}|BG)} = \prod_k \phi(\mathbf{v}_{T_k}). \tag{3.8}$$

This means the likelihood function can be factored into the products of many local terms, each of which is a likelihood ratio defined on a local triangular image region. Since $\mathcal{I}$ is constant, the k-th likelihood term, $p(\mathcal{I}_{R_k}|FG)/p(\mathcal{I}_{R_k}|BG)$, only depends on the position of the k-th triangle, $\mathbf{v}_{T_k}$. Therefore it will be simply denoted as $\phi(\mathbf{v}_{T_k})$.

In the following, we will gradually increase the complexity of the decomposition given by equation (3.8). First, visual patterns from different parts may not be coherent, and thus should be explained by different models. Accordingly we replace the homogeneous likelihood term $\phi(\mathbf{v}_{T_k})$ with $\phi(\mathbf{v}_{T_k}; \ell_{T_k})$, where ℓ_{T_k} is the observation model index for region R_k.

Second, foreground regions come from the same object so they are likely to be correlated. This can be modeled by merging multiple regions, or by using conditional terms like $p(I_{R_k}|I_{R_{k-1}})$. In this case, it is more convenient to assume that the shape is covered by a set of clusters $\mathcal{C}$ (Figure 3.5). Each cluster $C \in \mathcal{C}$ contains a small number of related vertices, on which a likelihood ratio $\phi(\mathbf{v}_C)$ can be defined. There is no one-to-one relationship between clusters and vertices. However, we can still impose a sequential structure on $\mathcal{C}$. Let $\mathcal{C}_k$ be those clusters that are completely covered only at step k, i.e., $\mathcal{C}_k = \{C|k \in C, C \subseteq [1:k], C \in \mathcal{C}\}$. It is easy to show that $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for $\forall i \neq j$, and $\mathcal{C} = \cup_k \mathcal{C}_k$.

Third, due to self-occlusion, foreground regions do overlap. The effect can be modeled by introducing correction terms in the sequential process of shape construction. Suppose at step t we visit a new cluster C which covers the region R_C. By inspecting $\mathbf{v}_C$ and $\mathbf{v}_{\mathcal{C}_{1:k-1}}$,

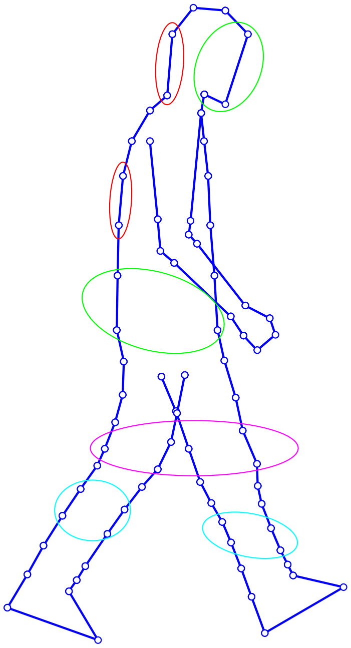


Figure 3.5: The body shape is covered by a set of clusters. Each cluster (large colored ellipse) contains a small number of related vertices (small circles), on which a likelihood ratio can be defined.

we may detect that R_C overlaps with a cluster region, say $R_{C'}$, that has been visited. In this case, we compute a correction term as follows and multiply it with the likelihood function,

$$\psi(\mathbf{v}_C, \mathbf{v}_{\mathcal{C}_{1:k-1}}) = \frac{\phi(\mathbf{v}_C, \mathbf{v}_{C'})}{\phi(\mathbf{v}_C)\phi(\mathbf{v}_{C'})}. \tag{3.9}$$

In fact, $\phi(\mathbf{v}_C)$ does not have to be a precise Bayesian generative model. An approximate measure may be good enough in practice, such as the foreground likelihood $p(\mathcal{I}_R|FG)$ alone, or a subjective energy term that may not be justified by statistics of the training data. A better alternative is to extract features $\mathcal{F}_R$ from the image patch $\mathcal{I}_R$, and replace the likelihood (ratio) to observe $\mathcal{I}_R$ by the likelihood (ratio) to observe $\mathcal{F}_R$. Thus our definition of ϕ can be modified as,

$$\phi(\mathbf{v}_C) = \frac{p(\mathcal{F}_{R_C}|FG)}{p(\mathcal{F}_{R_C}|BG)}. \tag{3.10}$$

Taking this idea further, we may extract features from different types of image cues. For each cue z, we may define a cluster structure $\mathcal{C}^z$, and a set of likelihood terms $\phi^z(\mathbf{v}_C)$. Assuming these cues are independent, the joint likelihood can be computed as their product.

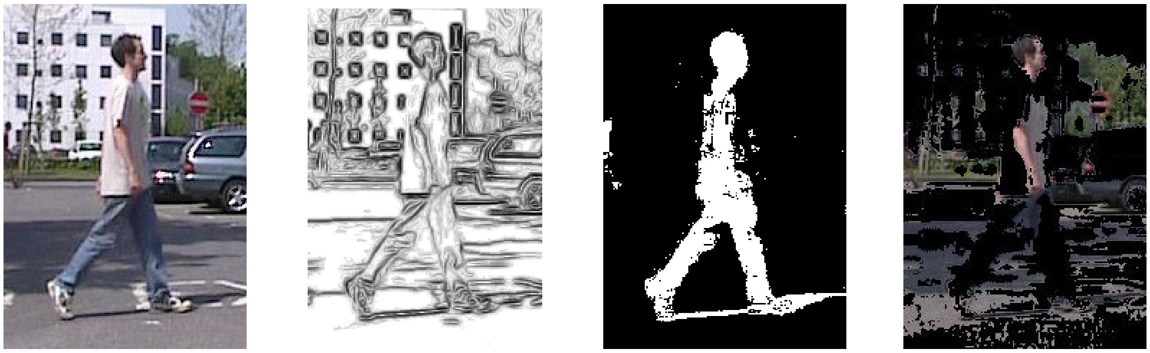
Taking all of the above into consideration, the likelihood model is expressed as,

$$p(\mathcal{I}|\Omega) \propto \prod_k \prod_z \prod_{C \in \mathcal{C}_k^z} \phi^z(\mathbf{v}_C; \ell_C)\, \psi^z(\mathbf{v}_C, \mathbf{v}_{\mathcal{C}^z_{1:k-1}}). \tag{3.11}$$

In this paper, we implement a simplified version of the observation model by: 1) assuming homogeneous likelihood, 2) dropping self-occlusion correction terms, and 3) computing $\phi(\mathbf{v}_C)$ by foreground likelihood alone. Accordingly equation (3.11) simplifies to,

$$p(\mathcal{I}|\Omega) \propto \prod_{k} \prod_{z} \prod_{C \in \mathcal{C}_k^z} \phi^z(\mathbf{v}_C). \tag{3.12}$$

Four types of image cues are used in computing $\phi(\mathbf{v}_C)$: edge gradient, silhouette, skin color, and region similarity. These cues are discussed in the following subsections.



(a) raw image (b) edge gradient (c) silhouette (d) skin/hair mask

Figure 3.6: A sample image (a) and three low-level visual cues (b)–(d) that are combined in the imaging model.

Edge Gradient

The edge potential ϕ^e is defined on the external boundary side of each triangle. We use a color edge detector called the compass operator [72]. At each pixel, this operator outputs a vector $\mathbf{u}$ ($\|\mathbf{u}\| \in [0,1]$) which encodes the strength and orientation of the edge feature at that point. Fig. 3.6b shows an example strength image. Given a line segment $\mathbf{e}$, we compute the boundary energy,

$$E(\mathbf{e}) = \int_s \mathbf{u}(s) \cdot \mathbf{e}/\|\mathbf{e}\|^2 \, ds, \tag{3.13}$$

and then model $\phi^e(\mathbf{e})$ with a truncated Gaussian,

$$\phi_e(\mathbf{e}) \propto \exp\{-[1 - E(\mathbf{e})]^2/\sigma_e^2\}, \quad E(\mathbf{e}) \in [0,1]. \tag{3.14}$$

Silhouette

The body silhouette potential ϕ^f is computed from a binary foreground mask $\mathcal{B}$ that labels pixels as 1 if they are likely to be on the person, and 0 if they are more likely to come from the background. This mask could be computed from a prior model of the color distribution of the person's clothing, via histogram backprojection [83]. However, in our experiments, we use a static camera and compute the mask using background subtraction. A standard background model of the mean and covariance of each pixel is used, and a binary mask $\mathcal{B}$ is generated by thresholding the Mahalanobis distance. Further, we assume that each pixel in $\mathcal{B}$ is drawn independently from the Bernoulli distribution $\{p_{10}, p_{11}\}$ if the pixel is in the foreground, or $\{p_{00}, p_{01}\}$ if it is in the background $(p_{\cdot 0}$+$p_{\cdot 1}$=1, p_{00}>p_{01}, p_{10}<$p_{11})$. Given a model configuration Ω, the probability of observing foreground mask $\mathcal{B}$ is derived as,

$$p(\mathcal{B}|\Omega) = \gamma \left(p_{10}/p_{00}\right)^{N_{10}} \left(p_{11}/p_{01}\right)^{N_{11}} , \tag{3.15}$$

where N_{10} is the number of pixels inside the model that are labeled background, N_{11} is the number of pixels inside the model that are labeled foreground, and γ is a constant independent of Ω. Noting that $N_{1\cdot}$ can be decomposed as $N_{1\cdot} = \sum_k N_{1\cdot}(\mathbf{t}_k)$, we have,

$$\phi^f(\mathbf{t}_k) \propto \exp\{\alpha_f N_{10}(\mathbf{t}_k) + \beta_f N_{11}(\mathbf{t}_k)\}, \tag{3.16}$$

where α_f and β_f are coefficients depending on p_{10} and p_{00}.

Skin Color

The skin potential ϕ^s helps to locate the head and arm. We use a simple skin detector based on a color histogram. The detector is learned from a training set of hand-labeled skin pixels. Because the face area is often very small in gait images shot from a side view, we extend the training set with hair pixels such that the resulting detector detects both skin and hair. Note that the skin/hair color mask can be very noisy and contain large false positive areas (Figure 3.6d). However, this is not a problem when complemented by other image cues. As the detector outputs a binary mask, a potential function similar to ϕ^f is used.

$$\phi^s(\mathbf{t}_k) \propto \exp\{\alpha_s N_{10}(\mathbf{t}_k) + \beta_s N_{11}(\mathbf{t}_k)\}, \tag{3.17}$$

Note that we only count skin and non-skin pixels in head and arm regions.

Region Similarity

The region similarity potential ϕ^r is defined by comparing appearances of image patches. It reflects the observations that: 1) appearances of adjacent triangles are likely to be similar; 2) appearances of symmetrically corresponding leg triangles are likely to be similar; and 3) appearance of foot and leg triangles are likely to be different. Given two triangles $\mathbf{t}_i$ and $\mathbf{t}_j$, we first compute the normalized color histograms h_i and h_j. Their distance is then defined using Bhattacharya coefficient $d_{ij} = \sqrt{1-\rho_{ij}}$, where $\rho_{ij} = \sum_k \sqrt{h_i(k)h_j(k)}$. Finally we model d_{ij} with a truncated Gaussian

$$\phi^r(\mathbf{t}_i, \mathbf{t}_j) \propto \exp\{-d_{ij}^2/\sigma_c^2\}, \quad d_{ij} \in [0,1]. \tag{3.18}$$

3.4 Inference by Sequential Monte Carlo

We now present our approach for finding modes in the posterior using stochastic search. Combining the equations for shape prior (3.7) and imaging likelihood (3.12) with the Bayes equation (3.1), the posterior distribution can be written as,

$$p(\Omega|\mathcal{I}) \propto \prod_k \Gamma_k \cdot \Phi_k, \tag{3.19}$$

where,

$$\Gamma_k = p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = \begin{cases} p(\mathbf{v}_k|\mathbf{e}_k^P, \Theta_{k-1}) & \text{if } T_k \text{ is joint} \\ p(\mathbf{v}_k|\mathbf{t}_k^P) & \text{otherwise} \end{cases}$$

$$\Phi_k = \prod_z \prod_{C \in \mathcal{C}_k^z} \phi^z(\mathbf{v}_C)$$

and Θ_k denotes the subset of joint angles that are visited as of step k, i.e., $\Theta_k = \{\theta_i | i \leq k, i \in \mathcal{J}\}$.

Equation (3.19) shows that the prior and likelihood terms are factored into a series of simple terms with the same sequential structure. This makes the methods of Sequential Monte Carlo (SMC) [24] especially attractive. SMC methods are flexible, easy to implement, parallelizable, and have the special property of drawing simultaneously a population of independent samples from the posterior distribution. Over the last decade, there has been a large number of papers in computer vision on SMC methods and their applications, under the names of *condensation* and *particle filters*. Once again, it is important to distinguish our use of SMC for body model fitting from the usual use in tracking body pose across time. Here, our chain is spatial, representing the sequential decomposition of contour landmark points, instead of a temporal chain of poses across time. Another difference is that we are using SMC for smoothing rather than filtering [44].

We employ the most basic version of SMC smoother. We traverse the shape model in K steps. At step k, we grow one landmark, expanding the configuration space by one more dimension. A natural choice for the proposal function π_k is the partial shape prior on $\mathbf{v}_{1:k}$, which has an iterative form,

$$\pi_k = \pi_{k-1} \cdot \Gamma_k, \tag{3.20}$$

with the (unnormalized) importance weights,

$$w_k \propto w_{k-1} \cdot \Phi_k. \tag{3.21}$$

Another key element in SMC is resampling in order to deal with a high number of dimensions. We use stratified resampling proposed in [44], which is optimal in terms of variance in the class of unbiased resampling schemes.

The inference procedure is summarized as follows.

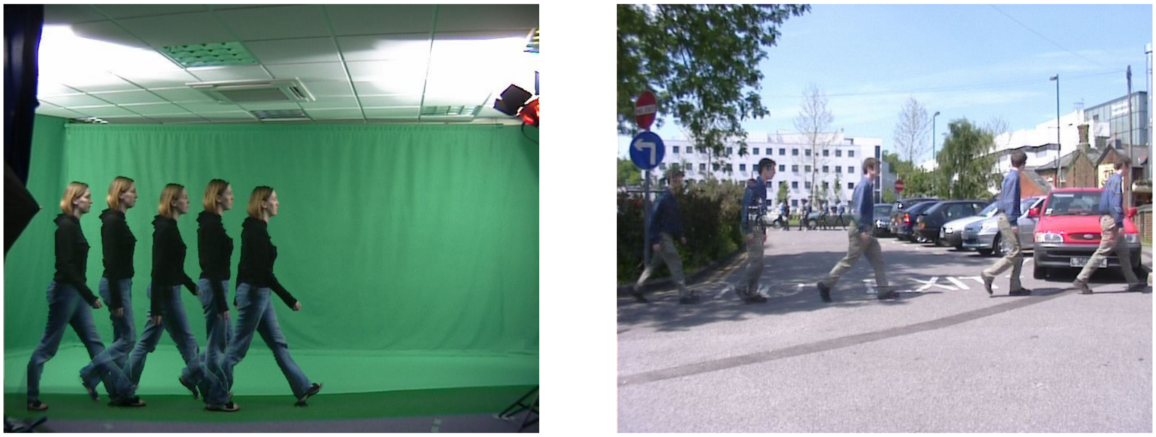
SMC INFERENCE PROCEDURE

1. INITIALIZATION.
   - For n = 1 to N, sample $\mathbf{v}_{1:2}^{(n)} \sim p_0(\mathbf{v}_{1:2}|\mathcal{I})$ and set $k = 3$.
2. IMPORTANCE SAMPLING.
   - For n = 1 to N, if T_k is joint, sample $\tilde{\mathbf{v}}_k^{(n)} \sim p(\mathbf{v}_k|\mathbf{e}_k^{P(n)}, {\Theta_{k-1}}^{(n)})$, otherwise sample $\tilde{\mathbf{v}}_k^{(n)} \sim p(\mathbf{v}_k|\mathbf{t}_k^{P(n)})$. Set $\tilde{\mathbf{v}}_{1:k}^{(n)} = \left(\mathbf{v}_{1:k-1}^{(n)}, \tilde{\mathbf{v}}_k^{(n)}\right)$.
   - For n = 1 to N, evaluate the importance weights $\tilde{w}_k^{(n)} = \prod_z \prod_{C \in \mathfrak{C}_k^z} \phi^z(\tilde{\mathbf{v}}_C^{(n)})$. Normalize the importance weights.
3. STRATIFIED RESAMPLING.
   - Resample N particles $\left\{\mathbf{v}_{1:k}^{(n)}\right\}_{n=1}^{N}$ from the set $\left\{\tilde{\mathbf{v}}_{1:k}^{(n)}\right\}_{n=1}^{N}$ according to the importance weights.
   - Set $k \leftarrow k + 1$ and go to step 2.

The procedure is initialized by uniformly sampling the root edge $\mathbf{v}_{1:2}$ over a range of position, rotation and scale. In our experiments, we sample $\mathbf{v}_{1:2}$ from within the top 1/3 portion of the image, oriented between $-\pi/3$ and $\pi/3$. The scale was chosen to satisfy $P(0.75 < l/h < 1) > 0$, where h is the image height, and l is the body height.

3.5 Experiments

3.5.1 Image Dataset



(a) Indoor (training)

(b) Outdoor (testing)

Figure 3.7: Example sequences from the Southampton gait database. Each displayed image is merged from five selected frames of the same sequence respectively, including the starting and ending frames.

We evaluate our deformable model using the Southampton HumanID gait database (available online at `http://www.gait.ecs.soton.ac.uk`), which was originally collected for research in automatic gait recognition. The database contains video sequences of walking individuals. Only sequences filmed from the side view are used in our experiments. The training data consists of 112 sequences of 28 subjects (3126 frames) filmed inside the lab, under controlled lighting with a green chroma-key backdrop (Figure 3.7a). The test data consists of 10 sequences of 10 subjects (963 frames) shot outdoors with cluttered background and natural lighting (Figure 3.7b).

Although the raw data are video sequences, we do not impose dynamic constraints on the body pose over time. For the purpose of body contour fitting, each frame is treated independently.

3.5.2 Learning Model Parameters

The body shape model was created by the following bootstrapping procedure. First, we built the triangulated body contour and identified its rotation joints by hand-labeling one frame of the indoor data. We then fit this model to all 3,126 indoor training frames using a uniform shape prior. Good fitting was obtained since the indoor green-screen images are very clean (Figure 3.8). The fits obtained were then used to learn a more informative empirical prior distribution on body shape parameters. We represent densities $\{p(\mathbf{n}_k), p(\rho_k), p(\Theta)\}$ in the shape prior by discrete probability tables. For each fit in the training set, a set of deformation parameters $\{\mathbf{n}_k, \rho_k, \Theta\}$ was calculated based on the posterior mean estimate, then discretized and pooled to compute the probability tables. Note that each table's dimension is at most three. The final model, including the learned shape prior, was then used for testing in cluttered scenes.

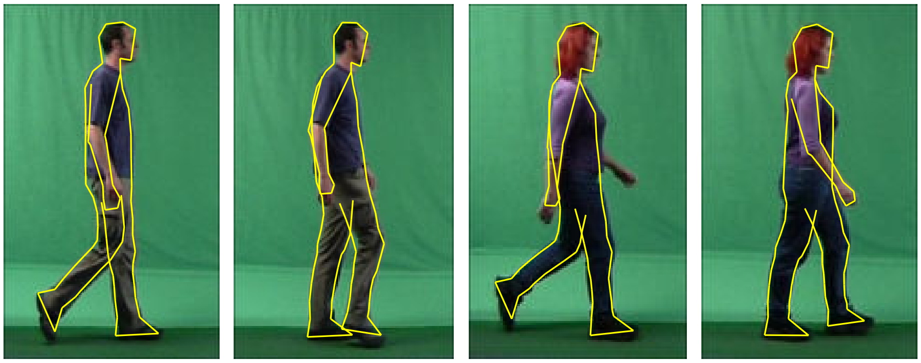


Figure 3.8: Sample results on fitting the indoor training set, using a uniform shape prior. Plotted are the posterior means.

We also trained the skin/hair color model using the indoor images. This leads to a weak classifier, since the lighting conditions of the indoor training images are very different from the outdoor natural illumination of the test scenes. Other parameters of the imaging model were also determined experimentally.

3.5.3 Test Result and Quantitative Evaluation

We applied the proposed algorithm to 963 images taken from the cluttered, outdoor gait sequences. Figure 3.9 shows two examples illustrating the incremental SMC inference procedure. For each step k, we plot the mean shape up to $\mathbf{v}_k$, with the marginal distribution of $\mathbf{v}_j$ $(j \leq k)$ summarized by its covariance ellipse (i.e., error ellipse). In the first image, the two legs are close to each other and a large uncertainty is observed when fitting the front leg. This uncertainty diminishes after both legs are fit. The second image has a background color similar to that of human skin, thus the head is not reliably detected until the body information has been incorporated.

A simple post processing procedure was used to deal with cases where both arms are

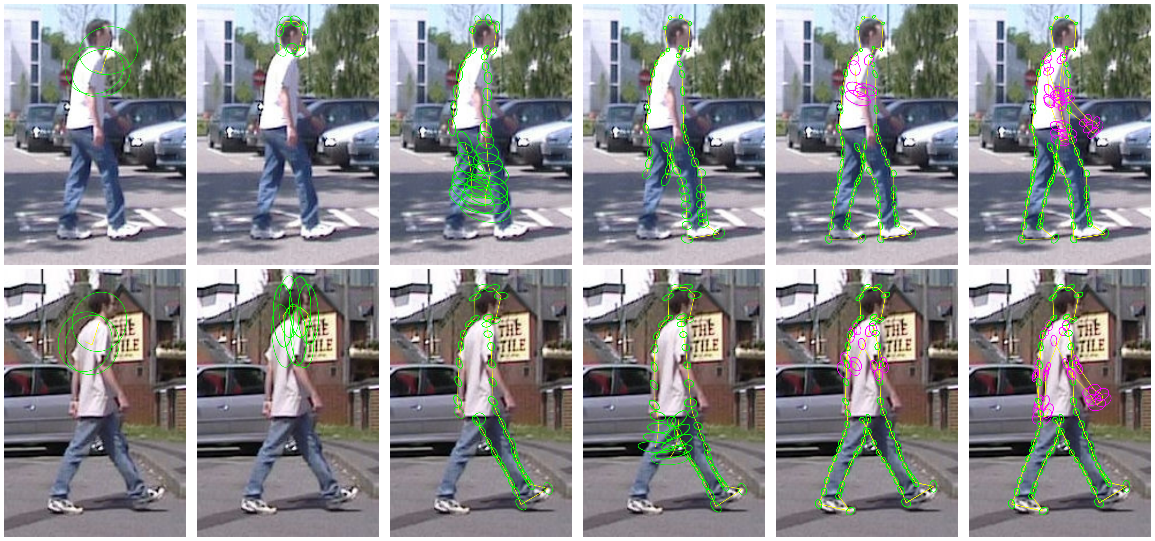


Figure 3.9: Two examples demonstrating the inference process of Sequential Monte Carlo. Plotted are the posterior means up to the k-th step, with $k \in \{1, 5, 32, 46, 64, 72\}$ for the first example, and $k \in \{1, 5, 40, 47, 64, 72\}$ for the second one. The distribution of each vertex is summarized by the shape of its covariance ellipse (error ellipse).

visible. First the sampled arm shapes are divided into two clusters based on hand positions, and the mean shape of each cluster is computed. Then we compare the hand distance between these two mean shapes to the width of the torso. If the ratio is above a threshold (empirically set to 0.6 throughout the experiment), then both arms are assumed to be detected.

To quantitatively evaluate the proposed model, we randomly selected 50 images and hand-labeled the ground truth boundaries of body parts. The posterior distribution computed by the SMC algorithm for each image is then summarized by a mean contour, which is compared to the ground truth using two types of metrics. One is symmetric Chamfer distance reflecting the global average error, and the other is symmetric Hausdorff distance reflecting the local worst-case error. Given two point sets $\mathcal{U}$ and $\mathcal{V}$, the Chamfer distance $d_{cham}(\mathcal{U}, \mathcal{V})$ is defined as the mean of the distances between each point in $\mathcal{U}$ and its closest point in $\mathcal{V}$. The symmetric distance is obtained by averaging $d_{cham}(\mathcal{U}, \mathcal{V})$ and $d_{cham}(\mathcal{V}, \mathcal{U})$. If two point sets are the same, $d_{cham}(\mathcal{U}, \mathcal{V}) = d_{cham}(\mathcal{V}, \mathcal{U}) = 0$. The Hausdorff distance is defined similarly except that we replace the mean with the maximum. We evaluate the fitting errors of body and arm separately, since it was expected that the core body shape (head, torso and legs) would be fit more accurately than the arms. Evaluation results are summarized in the last row of Table 3.1. To interpret these scores, note that

Table 3.1: Evaluation of model fitting by symmetric Chamfer and Hausdorff distances between mean contours and hand-labeled ground truth. The mean and standard deviation (in pixels) over 50 images are given in the form of MEAN ±STD. Each row corresponds to one combination of image cues. If selected, the source is marked with '•'.

$\phi^e\phi^f\phi^s\phi^r$	Chamfer		Hausdorff	
	Body	Arm	Body	Arm
● ○ ○ ○	4.00±3.52	4.41±4.08	16.7±13.7	10.7±7.20
○ ● ○ ○	2.53±0.91	6.25±5.95	10.2±3.52	13.0±9.20
○ ● ● ●	2.19±0.61	2.36±0.94	8.80±2.59	7.13±2.77
● ○ ● ●	2.77±1.62	4.13±6.83	11.8±6.14	9.49±8.41
● ● ○ ●	2.00±0.59	2.96±1.51	9.17±2.49	8.30±3.88
● ● ● ○	2.02±0.53	2.25±1.30	8.81±2.19	6.77±3.52
● ● ● ●	1.87±0.42	2.18±0.99	8.35±2.07	6.62±2.88

average body height is roughly 200 pixels in the dataset. Fitting results of all the 50 selected images are given in Figure 3.10 and Figure 3.11. More results are available online at http://www.cs.cmu.edu/~zhangjy/cvpr04/.

We also evaluated the utility of each image cue by comparing fitting accuracy both with and without that source of data. The results are shown in the first six rows of Table 3.1. It is observed that removing the foreground mask information decreases the performance accuracy the most, while appearance consistency affects performance the least. The scatter plot in Figure 3.12c suggests that, even without foreground/background information, we can still get reasonable fittings on a considerable portion of the testing images.

It is important to realize that the SMC inference procedure does not produce only a single estimate of model configuration, but an entire population of samples from the posterior distribution for the configuration. These samples can be summarized either by the mean or by the maximum a posteriori (mode). We observed that considerable differences between the mean and mode occasionally occur, indicating that the underlying posterior is indeed multimodal (Figure 3.13). Hence, representing the result of shape matching by a distribution may be preferable if, e.g., the shape model is biased or the available data is

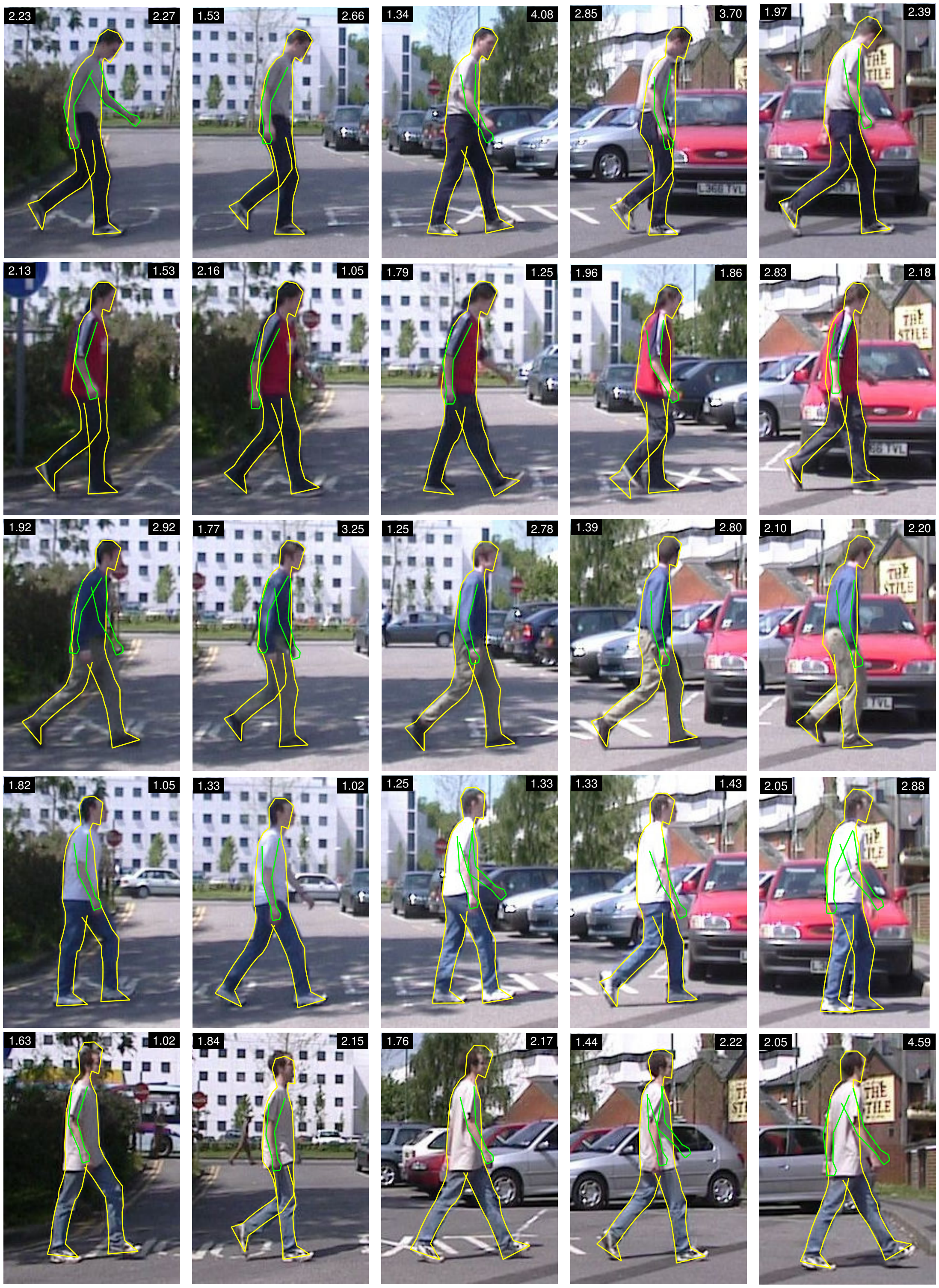


Figure 3.10: Results on the outdoor test set (Subjects 01–05). Plotted are the posterior means, with symmetric chamfer distance scores shown in the top corners (body on the left, and arm on the right). A lower score usually indicates a better fit.

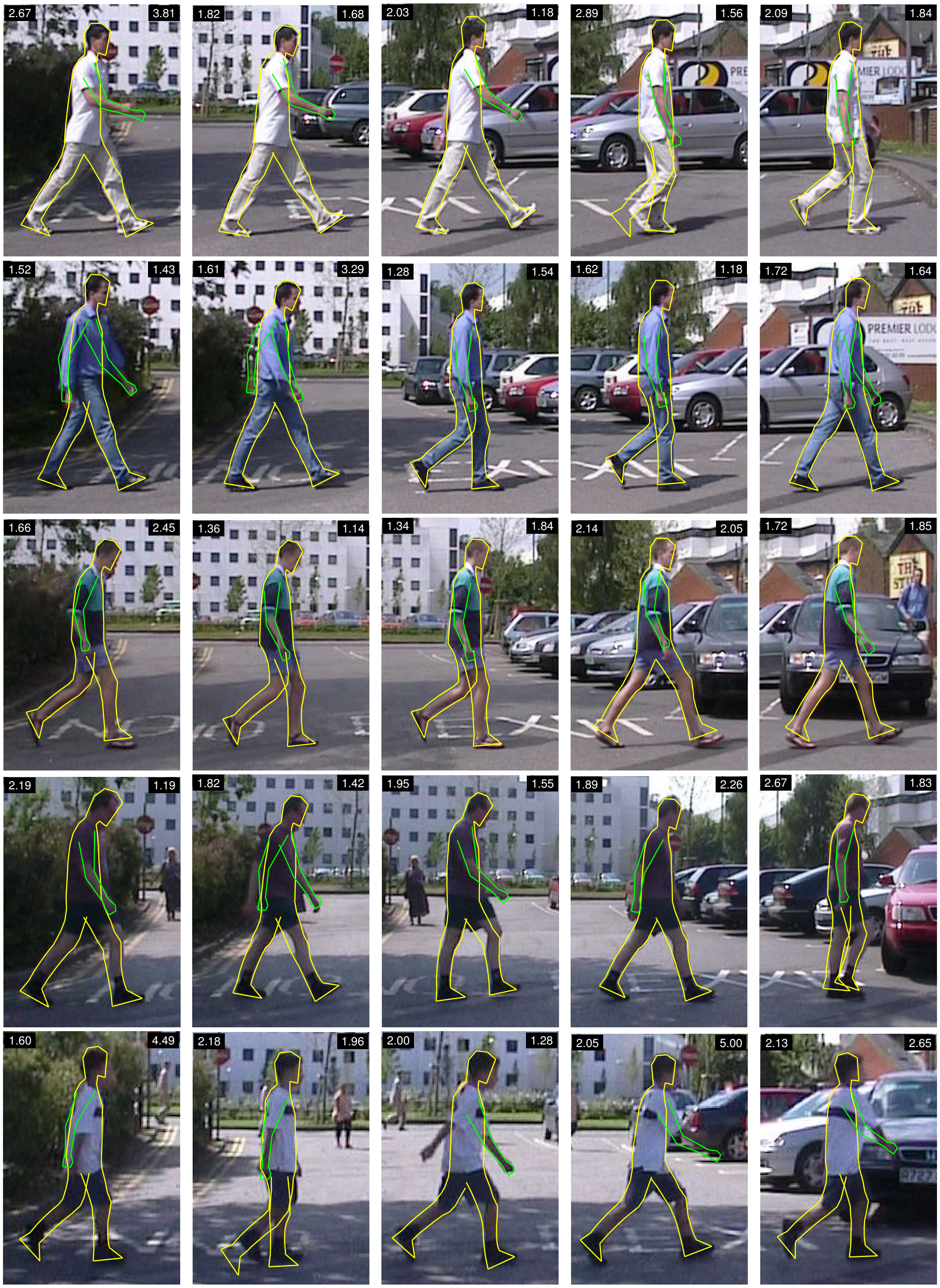
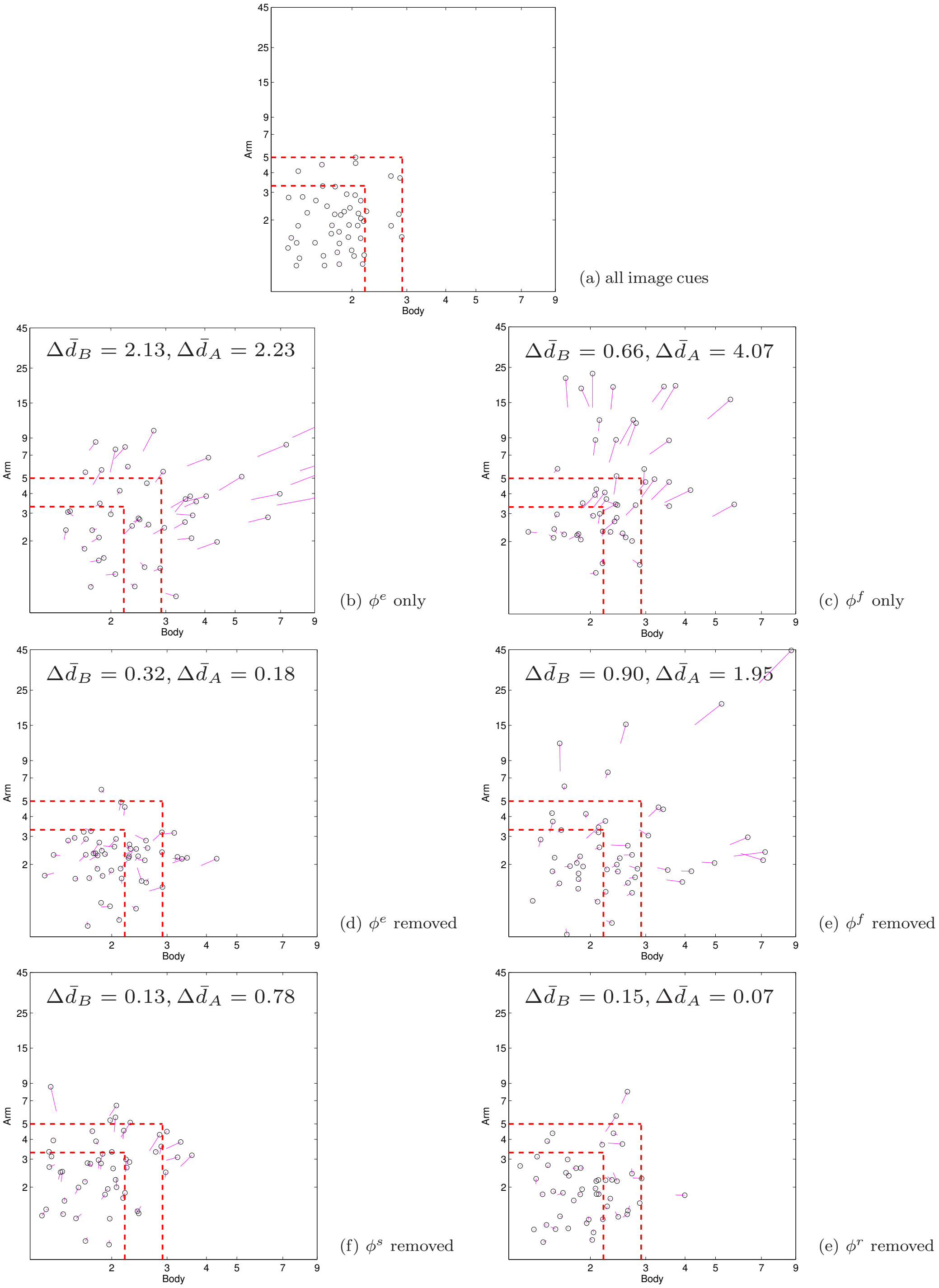


Figure 3.11: Results on the outdoor test set (Subjects 06–10). Plotted are the posterior means, with symmetric chamfer distance scores shown in the top corners (body on the left, and arm on the right). A lower score usually indicates a better fit.



(a) all image cues

(b) ϕ^e only

(c) ϕ^f only

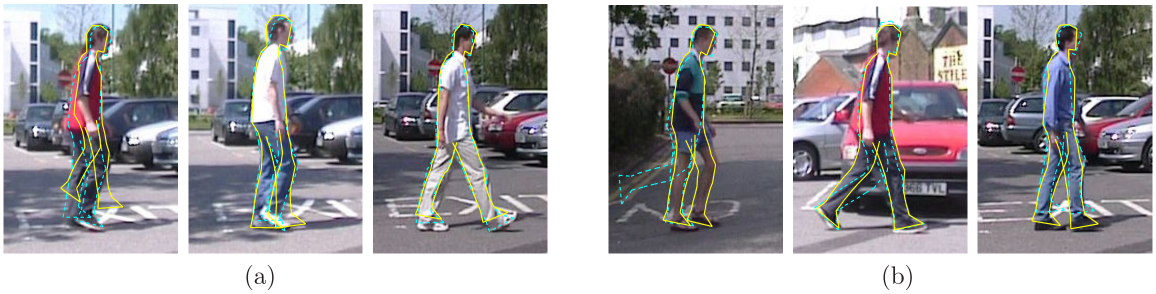
(d) ϕ^e removed

(e) ϕ^f removed

(f) ϕ^s removed

(e) ϕ^r removed

Figure 3.12: Scatter plots of Chamfer scores on 50 hand-labeled images. Both axes are in logarithmic scale. For each point in (b-g), a short "tail" is attached, directed to its corresponding point in (a), and scaled proportional to the distance (score change) between these two points. Average score increases are shown on top of each plot ($\Delta\bar{d}_B$ for the body, and $\Delta\bar{d}_A$ for the arm). Score bounds in (a) are shown as dotted red lines.



(a)

(b)

Figure 3.13: Some cases where discrepancies exist between the mean (solid yellow) and the maximum (dotted cyan) a posteriori of the SMC output. (a) Maximum (mode) is significantly better. (b) Mean is significantly better.

insufficient. Alternatively, methods for more intelligent mode selection could be used [56].

For the results shown here, we used on the order of 10^4 particles during sampling, and the inference algorithm took around one minute for each image on a 2GHz PC.

3.6 Summary

We have presented a 2D model-based approach for localizing the articulated and deformable shape of a walking human body in side-view images. The body shape is directly represented by the positions of landmarks densely sampled along the body contours. This representation provides a joint encoding of both pose and shape, while avoiding the ill-posed problem of 3D recovery. A learned shape prior and four types of local image cues are combined in a Bayesian framework. The model is decomposed into a chain-like structure, enabling simple spatial inference through SMC sampling. This stochastic search strategy can handle complex prior/likelihood definitions. It is also more efficient than regular sampling such as DP, which evaluates predefined points in the configuration space.

The method can be tailored to situations where only a single image or stereo image pair is available, noting the fact that foreground masks (or, equivalently, foreground/background appearance models) can be generated from many sources other than background subtraction, e.g., stereo depth maps, color segmentation, or robust model fitting [65].

Chapter 4

Mixture Shape Model

4.1 Introduction

Chapter 3 demonstrated, in a fixed viewpoint scenario, the effectiveness of the proposed 2D model-based approach to body localization. The model trained for one view works well for a small range of view angles. To tolerate a wider range of angles, we need to train new models with more flexible prior constraints.

In this chapter, we extend this approach to situations where the viewpoint of the human target is unknown. An example of such a scenario is to fit a random shot of a person walking in a circle (Figure 1.4b). Two main problems arise:

- How to handle the considerable shape variation caused by viewpoint changes with a 2D model?
- How to accommodate self-occlusion of body parts, which becomes unavoidable in this arbitrary-view scenario?

4.2 Overview of the Approach

The basic idea of our proposed solution is straightforward (Figure 4.1). First, we build a finite mixture of 2D view-dependent models. Each component of this mixture works for a small range of viewpoints. Then, we apply these component models simultaneously to the given image, and take the combination of their outputs.

Inference is done by direct sampling of the posterior mixture via SMC, searching in parallel through all view-dependent models. Resources are dynamically allocated according to the scores of their partial fits. In addition, we employ the well-known technique of

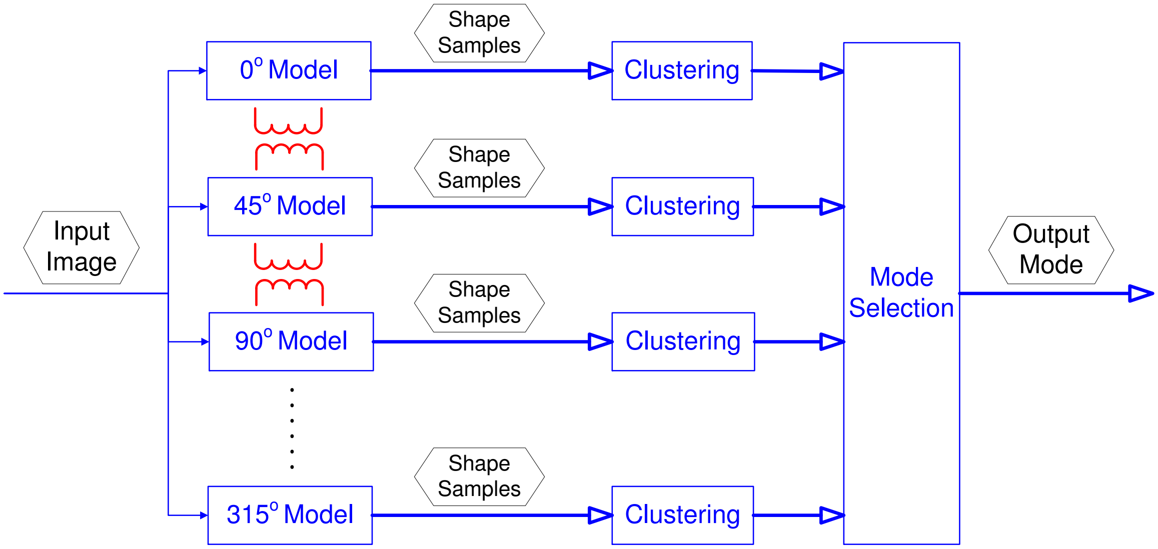


Figure 4.1: A multi-model approach to localizing the human body in images viewed from arbitrary and unknown angles. A number of 2D view-dependent models are constructed. Each model works for a small range of view angles. The models are fit simultaneously to the input image via SMC with dynamic resource allocation. Their outputs are combined via mode analysis and selection.

annealing and MCMC move to enhance the SMC inference performance. Note that our use of multiple deformable models does not computationally depend on the number of models, nor does it require preselection of a “correct” viewpoint. Therefore it is potentially easy to increase the number of mixture components in order to increase the modeling accuracy.

We also introduce an improved decomposition of the body shape representation. Specifically, landmarks are grouped into *parts* and *joints*, thus the nonlinear deformation of the model can be factored into shape variations of the parts and articulated motions of the joints. The deformation of each part/joint is modeled by either one or a mixture of simple distributions conditioned on the deformation of other parts. This conditioning is designed to impose anthropometric constraints on the relative lengths of the limbs. A depth ordering of parts is specified to accommodate self-occlusion when computing image likelihood terms. Our part-based model is conceptually similar to pictorial structures [28]. However,

1. Our part parameterization is more flexible to capture natural body deformation;

2. Our joint constraints are tight to preserve boundary continuity;

3. To handle self-occlusion and anthropometric constraints, our model is no longer a simple tree structure for inference purpose.

4.3 Bayesian Formulation

Let χ be a viewpoint index and $p(\chi)$ be the prior probability that the image is obtained from a particular viewpoint. The Bayesian formulation in Equation (3.1) is modified as,

$$p(\Omega|\mathcal{I}) \sim \sum_{\chi} p(\mathcal{I}, \Omega|\chi)\, p(\chi) = \sum_{\chi} p(\mathcal{I}|\Omega, \chi)\, p(\Omega|\chi)\, p(\chi). \tag{4.1}$$

This indicates that the posterior is a mixture of distributions with χ as the component index. Each component $p(\mathcal{I}, \Omega|\chi)$ corresponds to a different view-dependent model.

Currently we use eight component models from angles uniformly distributed in [0,2π). These are further simplified to five basic models (as depicted in Figure 4.2), by noting the fact that left facing models can be constructed by flipping their right counterparts. The remainder of this section specifies the component prior $p(\Omega|\chi)$ and likelihood $p(\mathcal{I}|\Omega, \chi)$. Note that all these models are parameterized in the same way, and the viewpoint index χ will be dropped for simplicity.

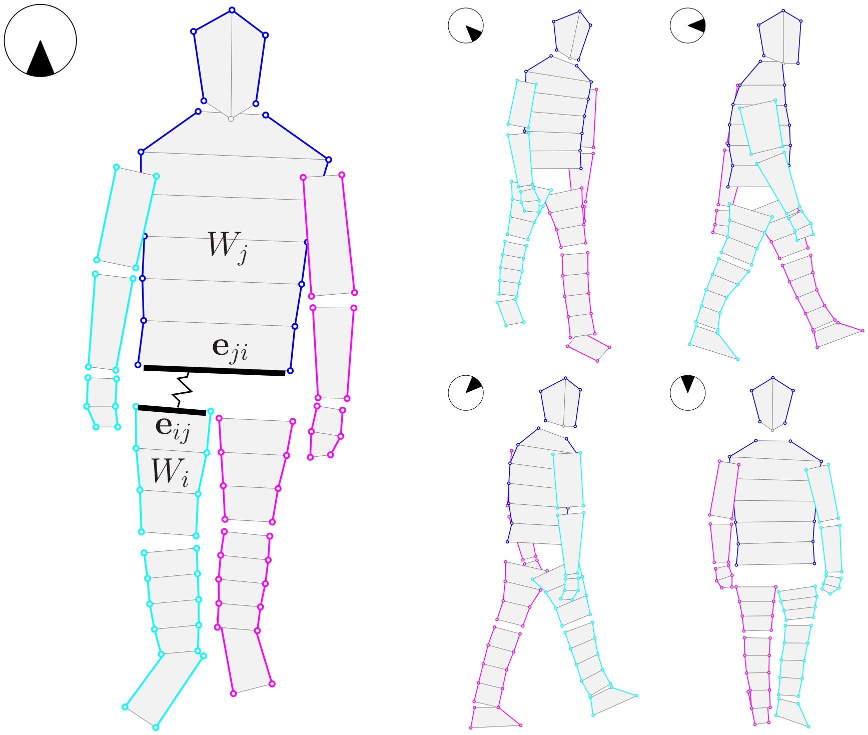


Figure 4.2: Topology of five basic component models. Landmarks are grouped into a collection of parts with depth order. A fixed landmark ordering is specified such that the shape can be traversed by growing one quadrilateral at a time.

4.3.1 Decomposing Prior by Parts

We note several limitations of the triangle-based model described in Section 3.3.2:

- A parent triangle needs to be manually specified for each within-part vertex;
- The local deformation energy based on affine prediction errors is somewhat *ad hoc*;
- The joint constraints from side views are inflexible and do not generalize well to other views.

Here we propose an improved prior decomposition to handle these issues.

We divide $\mathbf{v}_{1:K} = \{\mathbf{v}_k\}_{k=1}^{K}$ into M sequentially ordered *parts*, $\mathcal{W} = \{W_i\}_{i=1}^{M}$, where $W_i = \{\mathbf{v}_{i,k}\}_{k=1}^{K_i}$ consists of K_i sequentially ordered vertices (Figure 4.2). W_i is virtually attached to a particular parent part, say $W_j(j < i)$, through two edges, say $\mathbf{e}_i^j$ and $\mathbf{e}_j^i$. $\mathbf{e}_i^j$ is specified by the first two vertices of W_i, and $\mathbf{e}_j^i$ is specified by some pair of vertices from W_j. $(\mathbf{e}_i^j, \mathbf{e}_j^i)$ constitute a flexible *joint* that connects W_i and W_j. The M parts are connected into a "tree" structure by a total of $(M-1)$ joints $\mathcal{J} = \{(i,j)\}$. This tree structure can be traversed sequentially by visiting $\{\mathbf{v}_{1,1} \cdots \mathbf{v}_{1,K_1}\}\{\mathbf{v}_{2,1} \cdots \mathbf{v}_{2,K_2}\} \cdots \{\mathbf{v}_{M,1} \cdots \mathbf{v}_{M,K_M}\}$.

Given the fixed landmark ordering, the prior can be decomposed into a series of marginal and conditional distributions. Assuming the following Markov properties,

$$p(W_i|\mathbf{e}_i^j, W_{1:i-1}) = p(W_i|\mathbf{e}_i^j), \tag{4.2}$$

$$p(\mathbf{e}_i^j|W_{1:i-1}) = p(\mathbf{e}_i^j|\mathbf{e}_j^i), \tag{4.3}$$

the shape prior can be decomposed as,

$$p(\mathbf{v}_{1:K}) = p(W_1) \prod_{(i,j)\in\mathcal{J}} p(\mathbf{e}_i^j|\mathbf{e}_j^i) p(W_i|\mathbf{e}_i^j). \tag{4.4}$$

This suggests two types of deformation mechanisms. The first mechanism, encoded by $p(\mathbf{e}_i^j|\mathbf{e}_j^i)$, specifies the joint motion. We parameterize this motion by a similarity transform that maps $\mathbf{e}_j^i$ to $\mathbf{e}_i^j$ with the probability given by,

$$p(\mathbf{e}_i^j|\mathbf{e}_j^i) = p(x_i, y_i, \rho_i, \theta_i) = p(x_i, y_i)p(\rho_i)p(\theta_i), \tag{4.5}$$

where (x_i, y_i) is translational offset, ρ_i is scale and θ_i is rotation angle.

The second mechanism, encoded by $p(W_i|\mathbf{e}_i^j)$, models the local part deformation. We parameterize W_i by its Procrustes residuals $\mathbf{r}_{i,:} = \{\mathbf{r}_{i,k}\}_{k=1}^{K_i}$ and $\mathbf{e}_i^j$, where $\mathbf{r}_{i,:}$ is modeled as multivariate normal. To predict W_i, the mean shape of the i-th part is shifted by $\mathbf{r}_{i,:}$, followed by a similarity transform that maps the first two vertices of the shifted mean shape to $\mathbf{e}_i^j$. Assuming that the shape of W_i is independent of its location, rotation and scale. This implies

$$p(\mathbf{r}_{i,:}|\mathbf{e}_i^j) = p(\mathbf{r}_{i,:}), \tag{4.6}$$

The local deformation probability simplifies to,

$$p(W_i|\mathbf{e}_i^j) = p(\mathbf{r}_{i,:}) = \prod_k p(\mathbf{r}_{i,k}|\mathbf{r}_{i,1:k-1}). \tag{4.7}$$

Plugging Equations (4.5) and (4.7) into Equation (4.4) we get,

$$p(\mathbf{v}_{1:K}) = \prod_i p(x_i, y_i)p(\rho_i)p(\theta_i) \prod_k p(\mathbf{r}_{i,k}|\mathbf{r}_{i,1:k-1}).$$

Now we examine the assumptions we made in deriving the above decomposition. Although the human body possesses a disaggregated structure, there exist strong dependencies among the body parts. For example, contours of two adjacent parts are mostly continuous at their connection, and anthropometric constraints exist on the relative lengths of the limbs. The continuity constraint can be imposed in our model by proper choice of the origins of joint transforms and labeling of training data. However, the limb length constraint obviously invalidates our independence assumptions in Equations (4.2), (4.3) and (4.7). By parameterizing W_i with Procrustes residuals, its length l_i becomes a nonlinear function of both the shape $\mathbf{r}_{i,:}$ and the "scale" $\|\mathbf{e}_i\|$. As a result, imposing constraints on the limb length will induce a correlation between $\mathbf{r}_{i,:}$ and $\|\mathbf{e}_i\|$.

Based on this consideration, we modify Equations (4.2), (4.3) and (4.7) as,

$$\begin{aligned} p(W_i|\mathbf{e}_i^j, W_{1:i-1}) &= p(W_i|\mathbf{e}_i^j, l_1), \\ p(\mathbf{e}_i^j|W_{1:i-1}) &= p(\mathbf{e}_i^j|\mathbf{e}_j^i, l_1), \\ p(\mathbf{r}_{i,:}|\mathbf{e}_i^j) &= p(\mathbf{r}_{i,:}|\gamma_i^j), \end{aligned}$$

where l_1 is the length of W_1, and $\gamma_i^j = \|\mathbf{e}_i^j\|/l_1$. The final form of prior is,

$$p(\mathbf{v}_{1:K}) \propto \prod_{(i,j)\in\mathcal{J}} p(x_i, y_i|\gamma_j^i)p(\rho_i|\gamma_j^i)p(\theta_i) \prod_k p(\mathbf{r}_{i,k}|\mathbf{r}_{i,1:k-1}, \gamma_i^j). \tag{4.8}$$

We estimate densities in Equation (4.8) from labeled gait images. Figure 4.3 shows some random samples drawn from this learned shape prior. Note that we assume independent joint motion, thus the model is able to generate poses of activities other than walking.

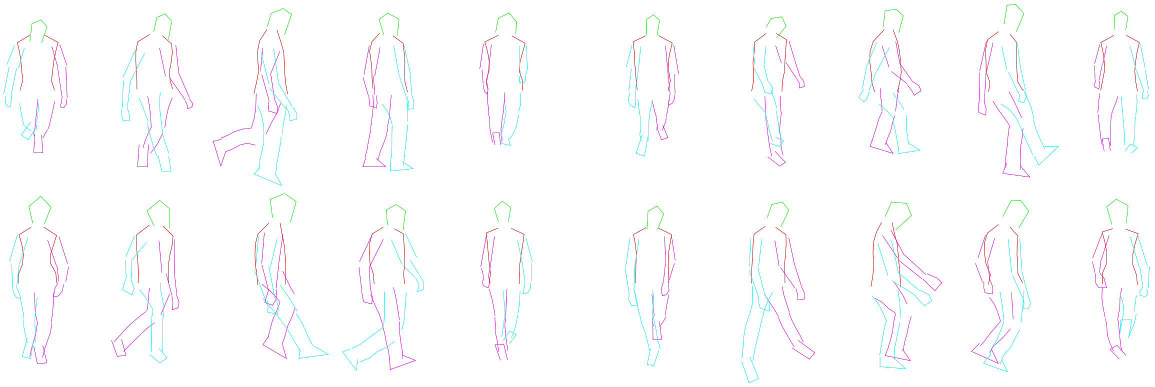


Figure 4.3: Selected random samples from the learned shape prior. Each row contains five samples corresponding to five component models. Each shape is normalized by aligning the torso with the associated mean shape.

4.3.2 Improved Likelihood

We implement the full observation model given by Equation (3.11). The body shape is divided into $T = K/2$ quadrilateral regions $Q_{1:T}$ (Figure 4.2). Four types of image features are computed similarly to the triangulated model as described in Section 3.3.3. Instead of using Gaussian-like foreground likelihood functions, image features are quantized and indexed into precomputed likelihood ratio tables. In addition, self-occlusion correction terms are taken into account for different types of features. As an example, the silhouette potential (3.16) is modified as follows. Let $\widetilde{Q}_t$ be the area within Q_t which is not covered by visited quadrangles, i.e., $\widetilde{Q}_t = Q_t \cap (\cap_{i<t} \overline{Q_i})$. Noting that $N_{1\cdot}$ can be decomposed as $N_{1\cdot} = \sum_t N_{1\cdot}(\widetilde{Q}_t)$, we have,

$$\phi^f(\mathbf{v}_{Q_t}) \propto \exp\{\alpha_f N_{10}(\widetilde{Q}_t) + \beta_f N_{11}(\widetilde{Q}_t)\}. \tag{4.9}$$

4.4 Simultaneous Model Fitting and Model Selection

There are two common strategies in using multiple deformable models. One is to fit each model completely then select the one that fits the best. This approach requires high computational cost when the model is complex. The other is to identify the “correct” model by a preprocessing step. However, sometimes it might not be possible to completely remove the uncertainty without fitting the model.

Our formulation of the fitting problem leads to the exploration of a posterior mixture given by Equation (4.1). To seek a sequential structure similar to Equation (3.19), we expand the configuration space Ω with the viewpoint index χ into an augmented configuration $\underline{\Omega} = \{\chi, \mathbf{v}_{1:K}\}$. Combining the equations for shape prior (4.8) and imaging likelihood (3.11), the posterior of $\underline{\Omega}$ can be written as,

$$p(\underline{\Omega}|\mathcal{I}) = p(\chi, \Omega|\mathcal{I}) \propto p(\chi) \prod_t \Gamma_t \cdot \Phi_t, \tag{4.10}$$

where,

$$\Gamma_t = \begin{cases} p(x_i, y_i|\gamma_j^i, \chi) p(\rho_i|\gamma_j^i, \chi) p(\theta_i|\chi) & Q_t \text{ is a joint polygon} \\ p(\mathbf{r}_{i,k-1:k}|\mathbf{r}_{i,1:k-2}, \gamma_i^j, \chi) & \text{otherwise} \end{cases}$$

$$\Phi_t = \prod_z \prod_{C \in \mathcal{C}_t^z} \phi^z(\mathbf{v}_C|\ell_C, \chi)\, \psi^z(\mathbf{v}_C, \mathbf{v}_{\mathcal{C}_{1:t-1}^z}|\chi).$$

Similar to the side-view case, this posterior can be sampled using SMC, which is equivalent to searching in parallel through all component shape models.

We traverse the shape model in $T = K/2$ steps. At step t, we grow two landmarks or equivalently a quadrilateral Q_t, expanding the configuration space by four dimensions. The proposal function π_t is the partial shape prior on $\mathbf{v}_{1:2t}$, which has an iterative form $\pi_t = \pi_{t-1}\Gamma_t$. The (unnormalized) importance weights are $w_t \propto w_{t-1}\Phi_t$. The output of SMC inference procedure $\left\{\chi^{(i)}, \mathbf{v}_{0:K}^{(i)}\right\}_{i=1}^N$ is the sample representation of the posterior mixture. Note that the viewpoint parameter χ can be marginalized out from the posterior distribution if we are only interested in localizing the positions of body contours.

For the complex mixture model, basic particle filters may not work well. Here we employ two well-known techniques to enhance the SMC performance: annealing and MCMC move.

Annealing

The basic idea of annealing is to gradually increase the peakiness of the likelihood term in order to avoid being trapped in local maxima during the early stage of the search. At each step, we compute a correction term from all visited clusters based on the change in their observation model, and multiply this correction term to the importance weight. For silhouette potential ϕ^f, we adjust the parameters $\{p_{00}, p_{01}\}$ in Equation (3.15). The reason is that, when we fit a partial shape, the foreground area that the partial shape did not cover should be considered as background. As a result, a pixel in this background is more probable to be labeled as 1. This implies that using the same background model during the search procedure is inherently inappropriate. For other image cues, we use the formula $\ln \phi(t) = \xi_t \cdot \ln \phi$, where ξ_t increases linearly from $1/T$ to 1.

MCMC Move

In standard SMC procedure, all new samples of a vertex, say $\mathbf{v}_j$, are generated at step j. The number of distinct values of these samples, say S_j, is finite. As every resampling after step j results in a decrease in S_j, sample density will gradually diminish and eventually we lose the accuracy of the distribution of $\mathbf{v}_j$. This phenomenon is sometimes referred to as sample attrition, or particle degeneracy. To alleviate this problem, we move each particle once after every resampling procedure, using Metropolis update [31]. Specifically, given a particle $\{\chi^{(i)}, \mathbf{v}_{0:2t}^{(i)}\}$ at step t, a new particle $\{\chi^{(i)}, \tilde{\mathbf{v}}_{0:2t}^{(i)}\}$ is generated from a Gaussian proposal density $N(\mathbf{v}_{0:2t}^{(i)}, \eta_t \Sigma_t)$, where Σ_t is the covariance matrix estimated from the current particle set, and $\eta_t < 1$. $\tilde{\mathbf{v}}_{0:2t}^{(i)}$ is accepted with the probability $\min(1, p(\chi^{(i)}, \tilde{\mathbf{v}}_{0:2t}^{(i)}|\mathcal{I})/p(\chi^{(i)}, \mathbf{v}_{0:2t}^{(i)}|\mathcal{I}))$. Currently we have not implemented jump transition between different viewpoint models.

4.5 Experiments

4.5.1 Multi-view Data Collection

The first challenge in building a mixture of view-dependent models is to obtain realistic, multi-view training data. In the side-view case (Chapter 3), we solved this problem by fitting a hand-labeled template to indoor green-screen images using a uniform shape prior. The quality of the fits obtained is inevitably limited. Here, we choose to hand-label the data by combining interactive tracking and the presented localization method. Given a walking sequence, we first hand-labeled a number of key frames and used them to initialize an appearance-based body tracker. We then edited the tracking errors by hand and, if necessary, added more key frames. This procedure was repeated until all frames in the sequence were correctly labeled. We only labeled arms and legs on one side using interactive tracking, as their counterparts suffered from severe occlusion thus were very difficult to track. Instead, we fit the missing limbs using the presented shape model, which was learned from the partially labeled data.

We use the CMU Mobo database which contains 25 subjects walking on a treadmill [33]. The subjects perform four different activities: slow walk, fast walk, incline walk and walking with a ball. For each subject, six synchronous sequences were recorded by six cameras evenly distributed around the treadmill (Figure 4.5). We use 150 slow-walk sequences for the training purpose. For each sequence, we labeled around 50 frames, which covers more than a complete gait cycle. Figure 4.4 shows some training examples overlaid with the labeled body contours.

4.5.2 Virtual Examples by Viewpoint Interpolation

An interesting fact is that, since the cameras are synchronized, the labeling at six discrete views can be interpolated to generate virtual contours at an arbitrary angle (Figure 4.5). This potentially enables us to construct densely populated body shape models. However, there are two $\pi/2$ angular gaps in the Mobo camera setup that are too large to generate realistic interpolation. This problem can be fixed using the periodic and symmetric property of human walking. First we flip horizontally the images from the cameras opposite to the missing ones. Then we shift the flipped sequences by half of the gait cycle. Figure 4.6 shows two examples of virtual body contours generated by linear interpolation.

4.5.3 Test Result

We applied the mixture shape model to both indoor and outdoor cluttered scenes. First, we tested the model on CMU Mobo incline-walk and fast-walk sequences. For each sequence we randomly selected one frame, resulting in a test set of 300 images. These images were obtained from view angles similar to the training data, but with the subject performing different activities. Figure 4.7 shows some example results. Plotted are the output of a simple mode selection procedure, which was used to deal with the possible swapping between left and right limbs in the inference output. First, the sampled body shapes generated by each component model are split into two clusters based on hand and foot positions respectively. Then we select the cluster with the highest fitting score, and output its mean shape and associated component model index (plotted as a compass in the top left corner of each image). As can be observed in Figure 4.7, both the viewpoint and body boundary estimates are quite accurate. Considering the fact that the variations of body shape among these 25 subjects are quite large, the results do demonstrate the superior modeling ability of our shape model. Note that the target poses in some images are quite different from the slow-walk training data, but are nevertheless correctly handled. This is so because our model assumes independent joint motion without activity specific constraints.

We also applied the model to a widely tested outdoor video sequence of a person walking in a circle (available online at `http://www.nada.kth.se/~hedvig/data.html`). Sample results are shown in Figure 4.8. The video contains a total of 174 frames with the size of 320×240 pixels. Note that we did not use the sequential nature of the data to impose dynamic constraints on the body pose over time. Each frame is fit *independently*.

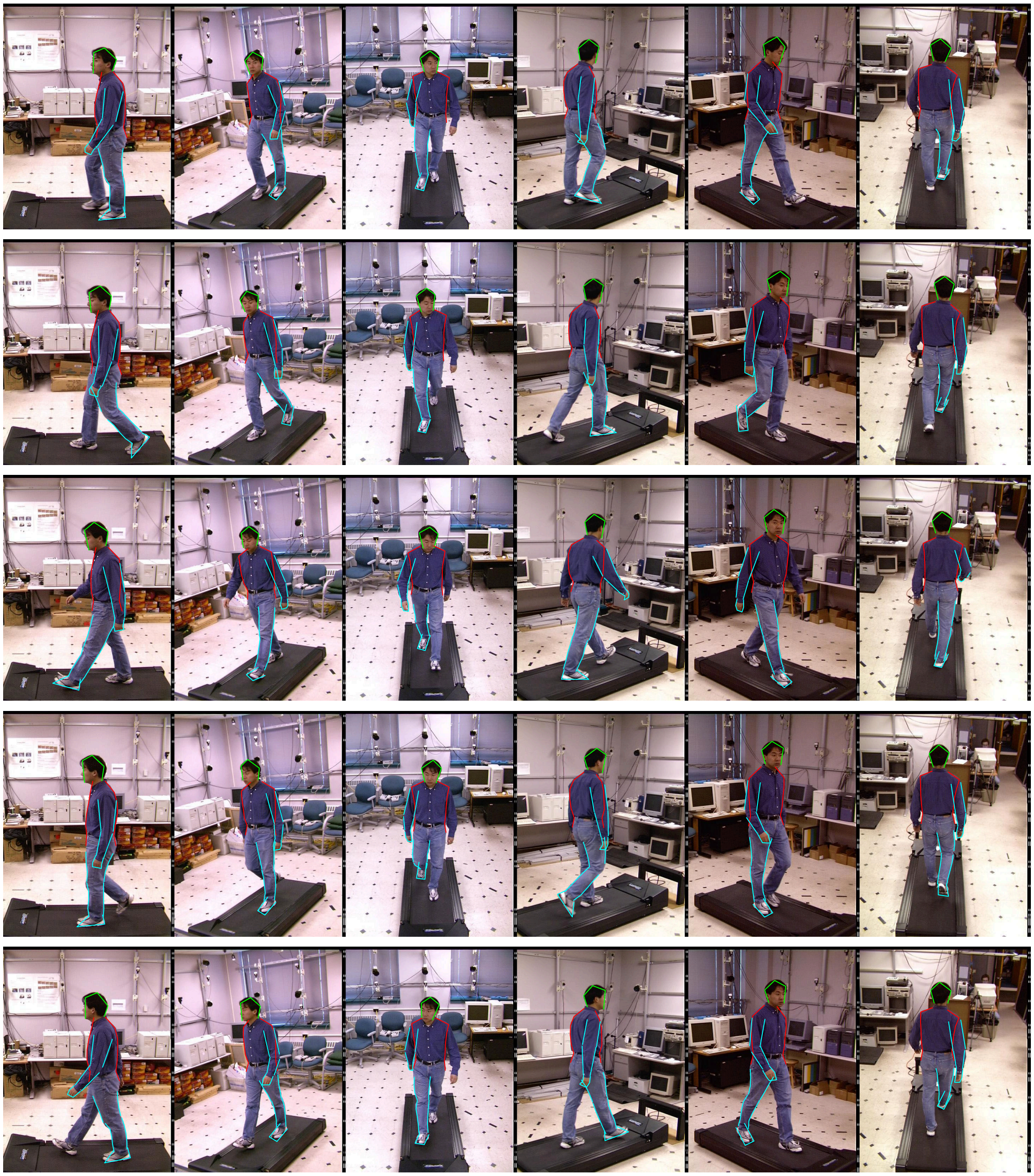


Figure 4.4: Example training images of one subject, overlaid with body contours obtained by interactive tracking. Rows top to bottom correspond to frames 1, 11, 24, 33, and 42 from six synchronous sequences.

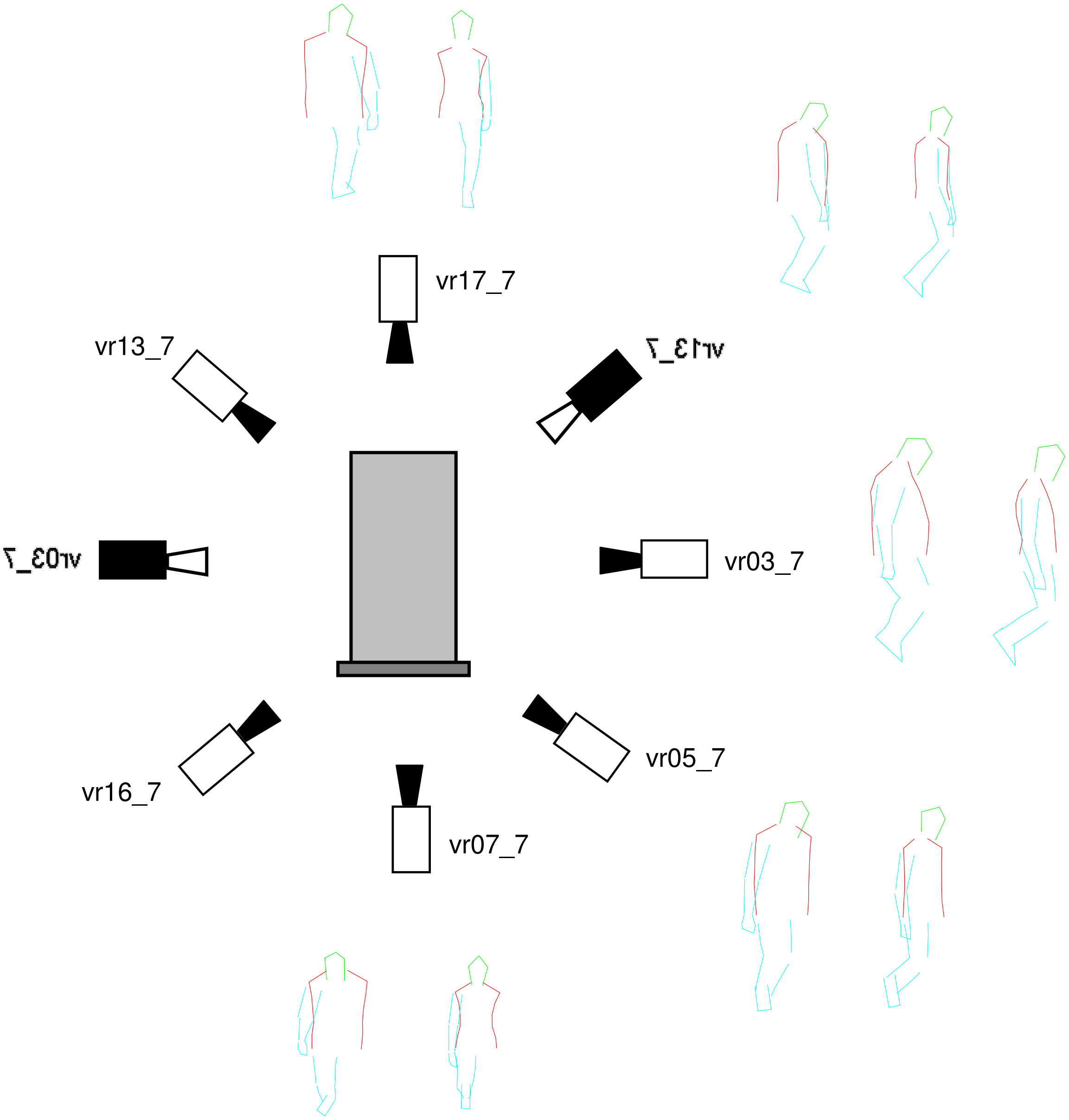
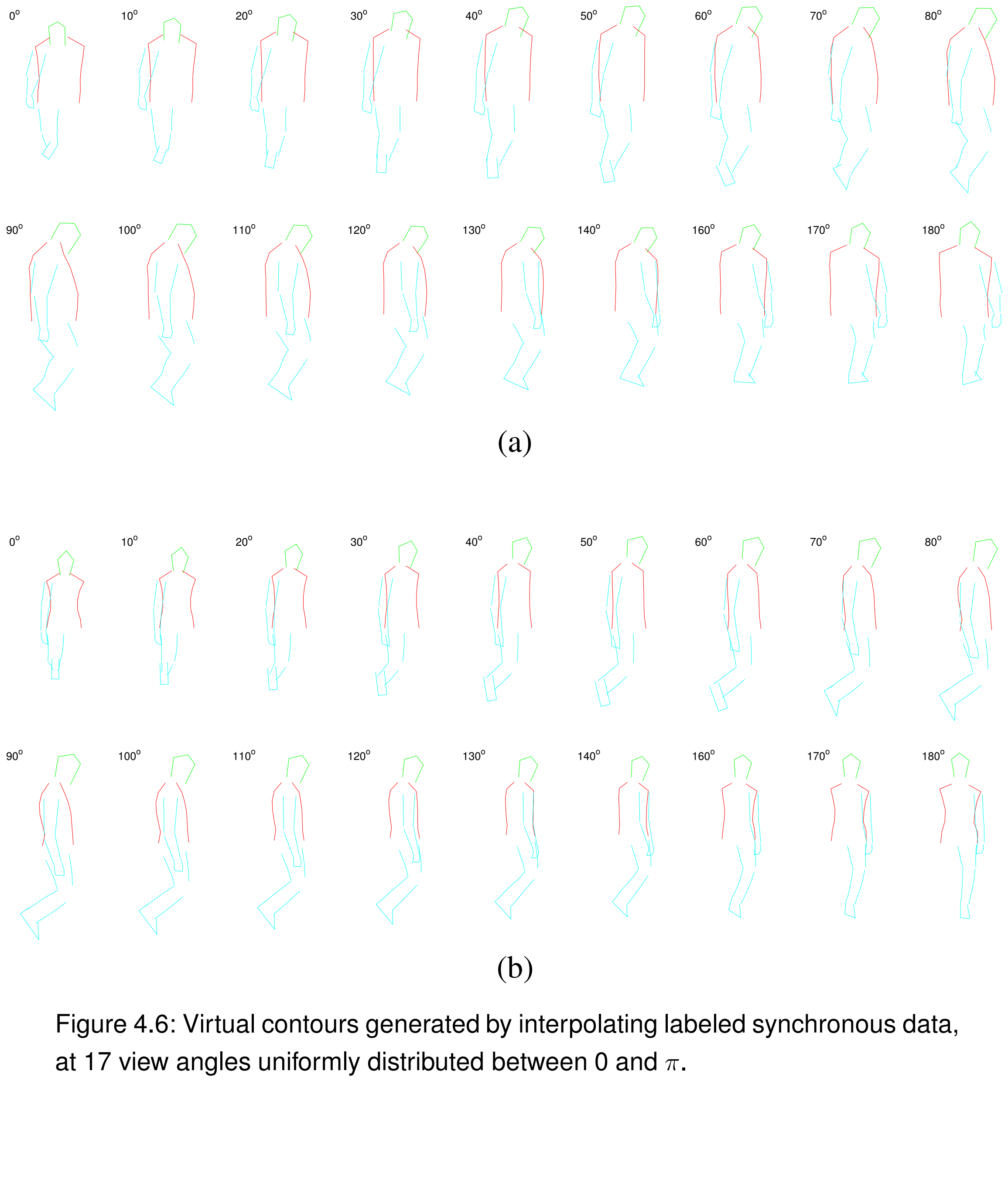


Figure 4.5: Camera setup of the CMU Mobo database. Two missing cameras (north-east and west) are simulated by flipping the images from their symmetric counterparts (north-west and east).



(a)

(b)

Figure 4.6: Virtual contours generated by interpolating labeled synchronous data, at 17 view angles uniformly distributed between 0 and π.

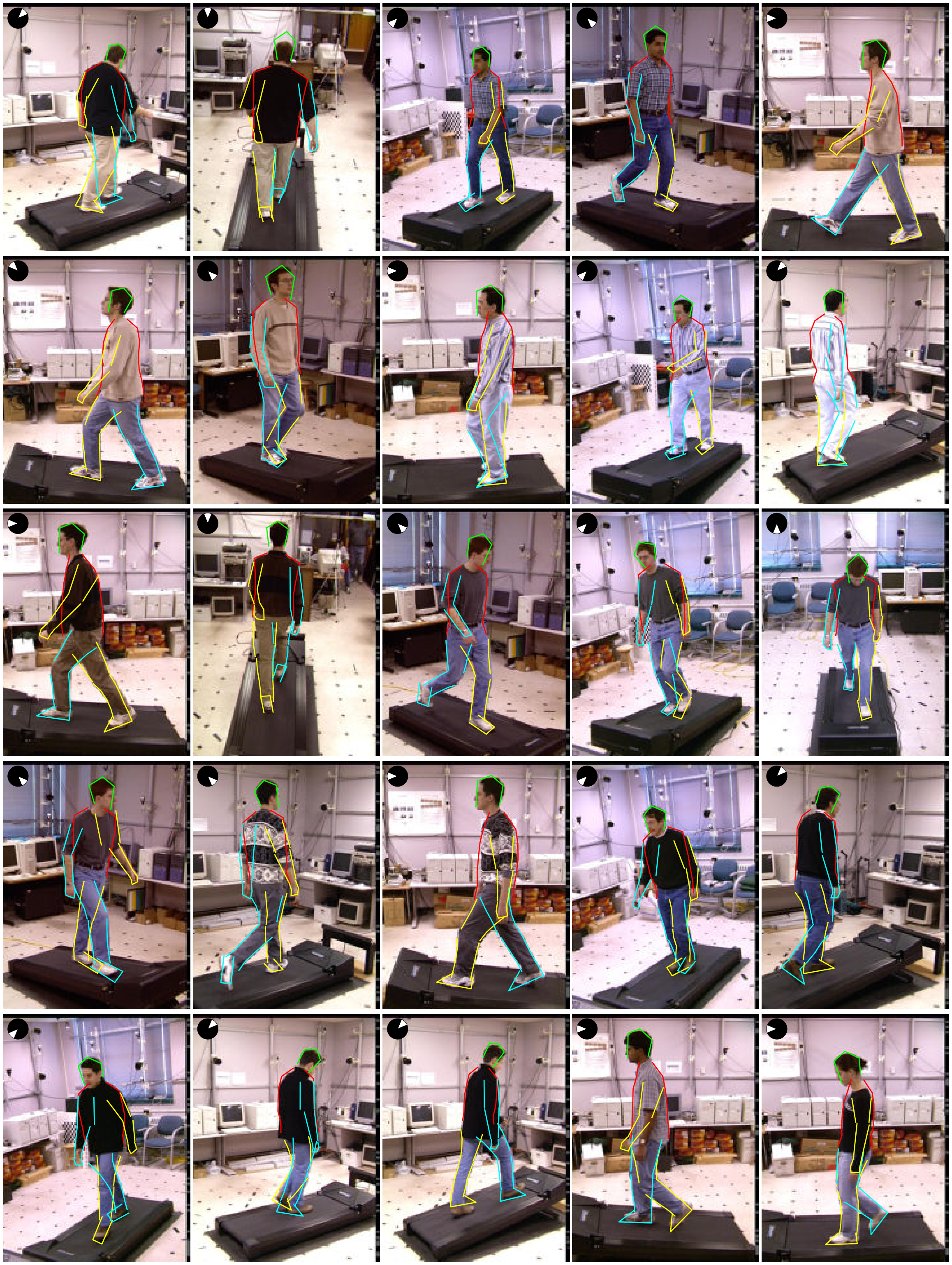
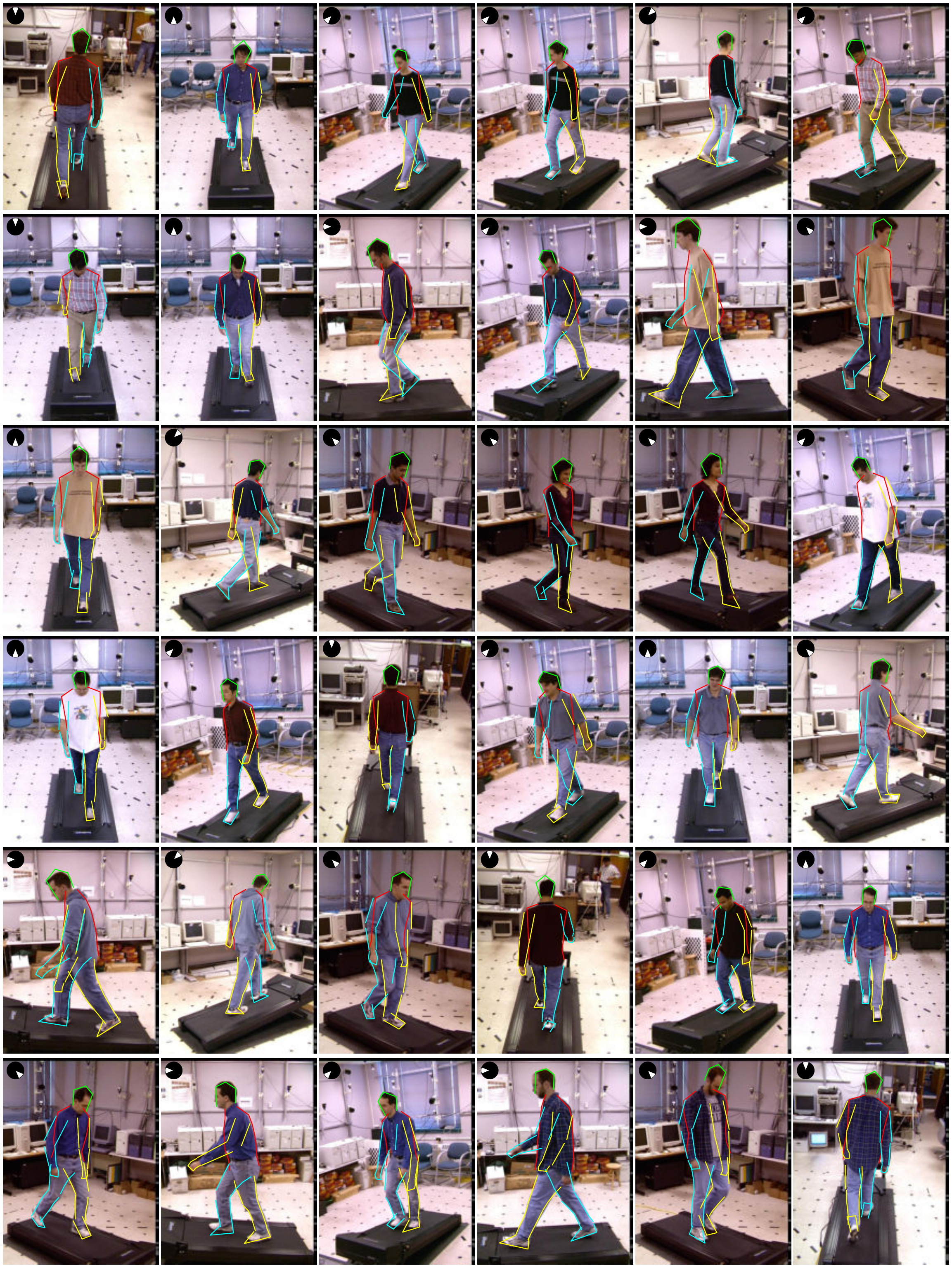


Figure 4.7: Sample test results on frames randomly drawn from incline-walk and fast-walk sequences of CMU Mobo dataset.

To be continued on pp. 52.



(Figure 4.7 on pp. 51 Continued) — Mobo Test — (1/1)

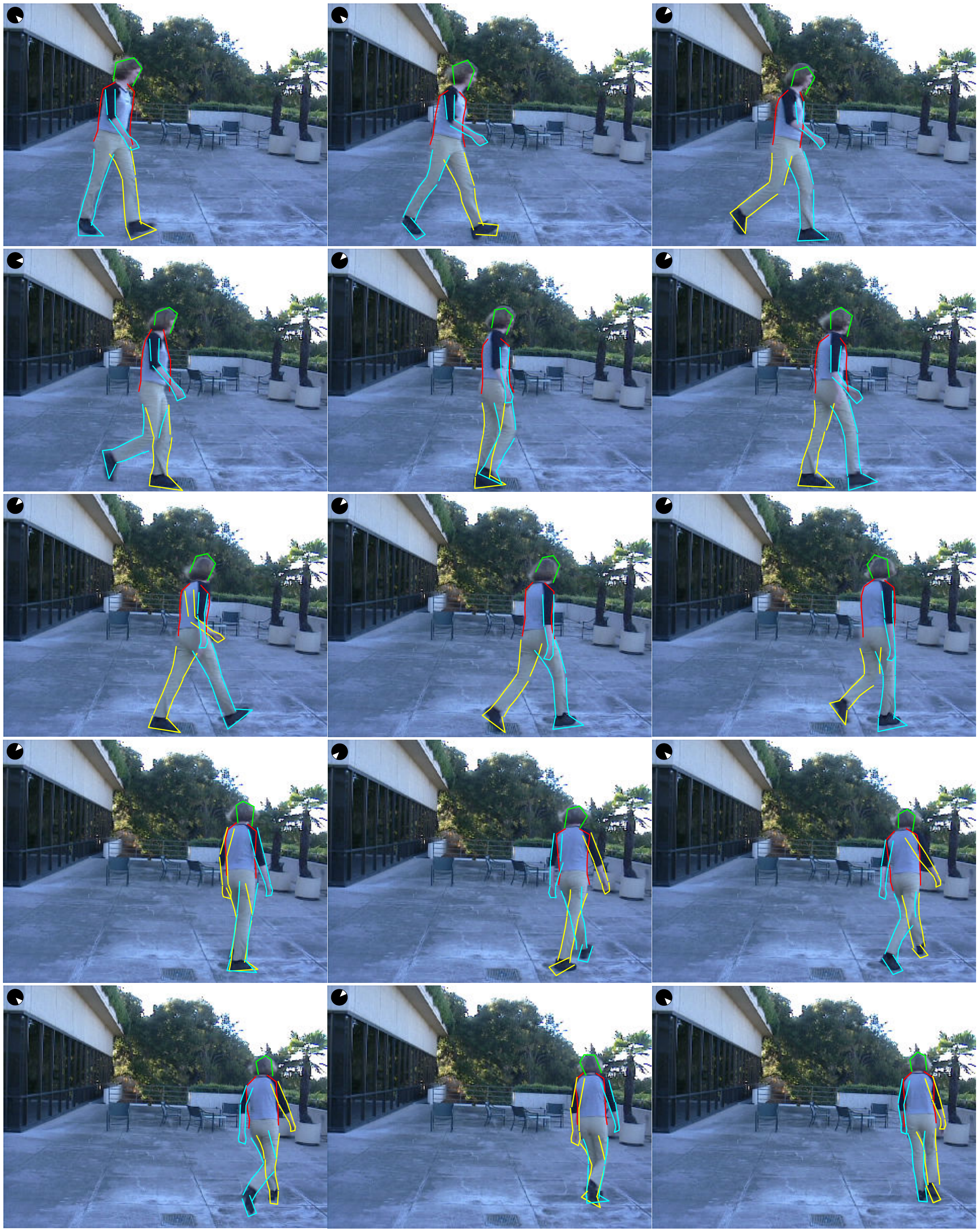
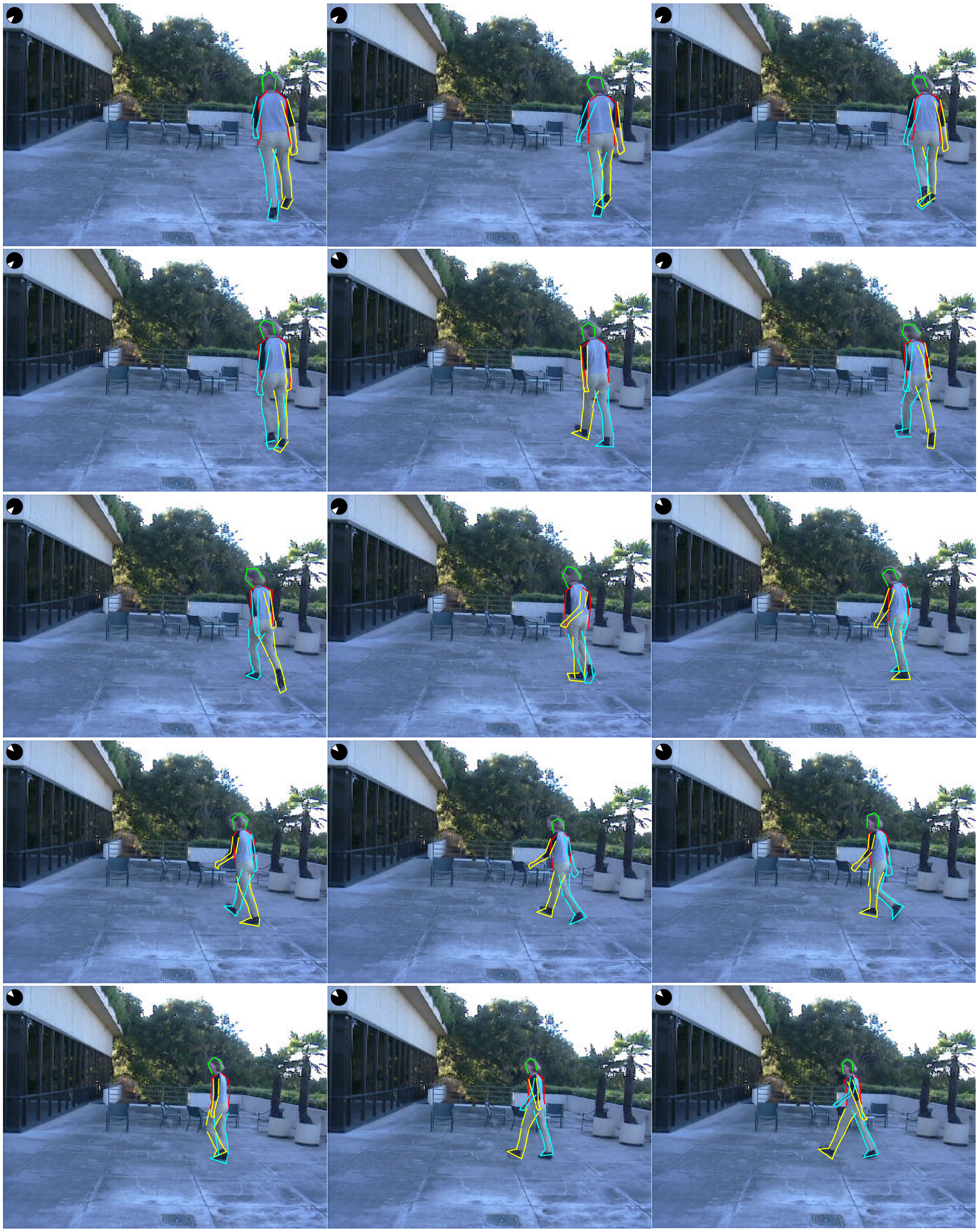
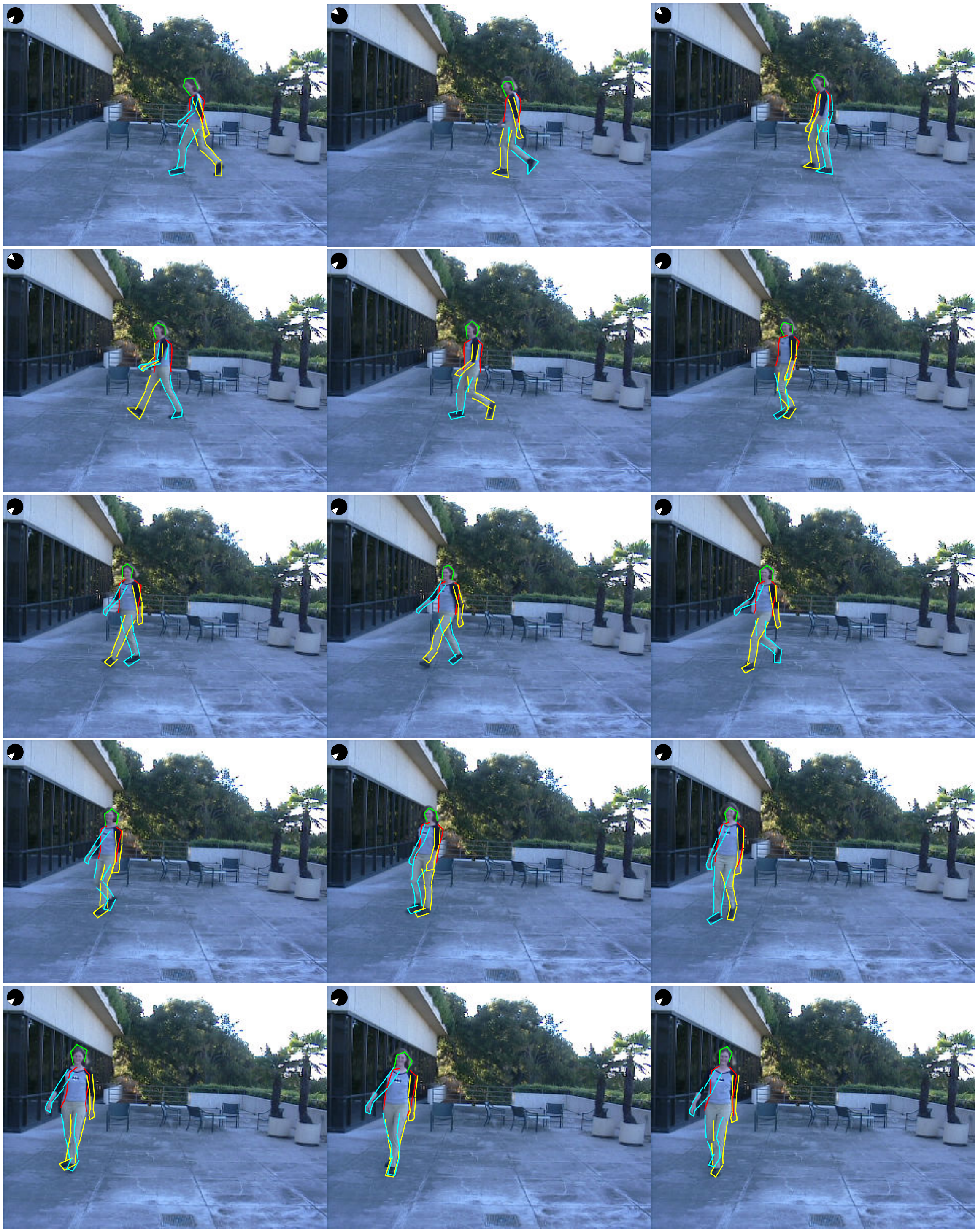


Figure 4.8: Sample test results on a 174 frame sequence of a person walking in a circle. Each frame is fit independently. Complete results are available at `http://www.cs.cmu.edu/~zhangjy/iccv05/`.

To be continued on pp. 54–55.



(Figure 4.8 on pp. 53 Continued) — Circle Test — (1/2)



(Figure 4.8 on pp. 53 Continued) — Circle Test — (2/2)

This test set is challenging in several ways. First, it contains continuous change of viewpoint, while the gap between our shape models is 45^o. Second, the circle radius is quite small. The head, torso, legs and feet of the target are almost never in the same direction. Third, the elevation angle of this test data differs from our training data by 10^o–15^o for side views, and 25^o for front and back views. The fitting algorithm shows reasonable performance on estimating the shape boundaries of body parts. However, we observed large noise in the viewpoint estimate. One obvious reason is the difference between training and testing viewpoints. Another reason is that the elevation angle of the test data is close to zero, in which case the inherent ambiguity between symmetric viewpoints becomes more evident.

4.6 Summary

We have extended our 2D model-based approach to localizing a human body in images viewed from arbitrary and unknown angles. The central component is a statistical shape representation of the nonrigid and articulated body contours, where a nonlinear deformation is decomposed based on the concept of parts. Several image cues are combined to relate the body configuration to the observed image, with self-occlusion explicitly treated. To accommodate large viewpoint changes, a mixture of view-dependent models is employed. Inference is done by direct sampling of the posterior mixture, using Sequential Monte Carlo (SMC) simulation enhanced with annealing and MCMC move. The fitting method is independent of the number of mixture components, and does not require the preselection of a "correct" viewpoint. The models were trained on a large number of interactively labeled gait images. Preliminary tests demonstrate the feasibility of the proposed approach.

Chapter 5

Hierarchical Models and Hybrid Search

5.1 Introduction

In the previous two chapters, we used a Bayesian top-down approach to take advantage of strong priors on body deformation, and to combine multiple image cues in a robust fashion. The body shape was parameterized by the positions of point landmarks densely sampled along the body contours. To accommodate large viewpoint changes, a mixture of view-dependent models was employed. Each model was decomposed based on the concept of parts, with anthropometric constraints and self-occlusion explicitly treated. Such a mixture model possesses the potential for high-accuracy localization, but has a complex form for which most inference algorithms do not apply. Thus we introduced a sequential structure so that inference could be done by Sequential Monte Carlo.

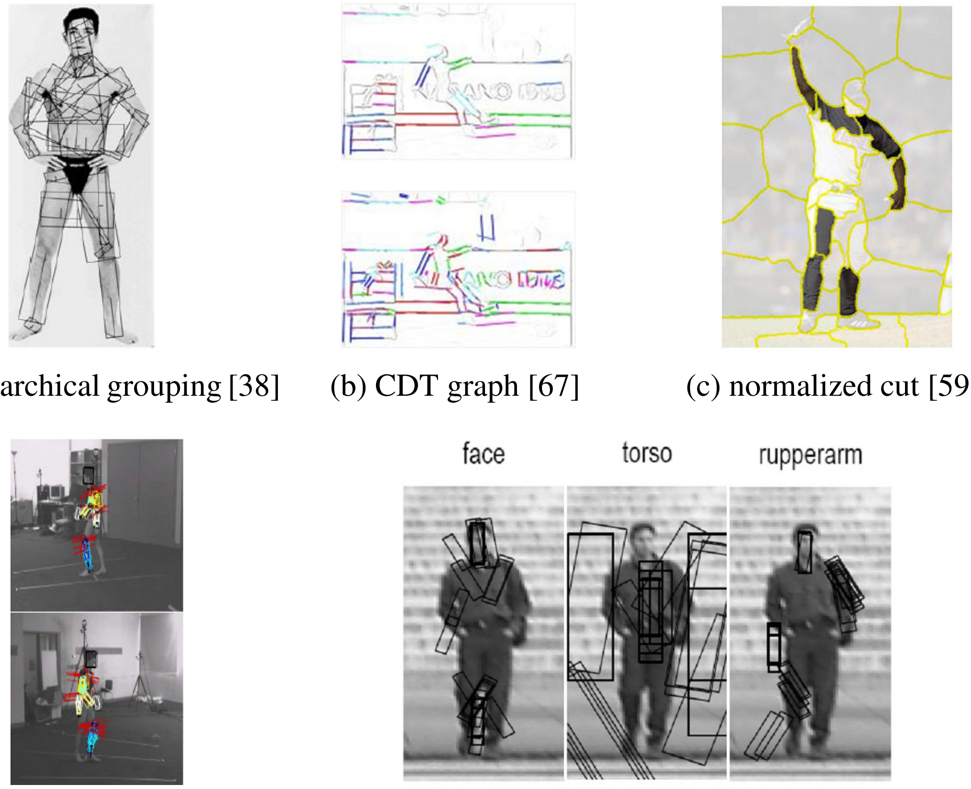
In this chapter, we study the problem of body localization in a generic setting: single image, arbitrary pose, and arbitrary viewpoint (Figure 1.4c). This is the ultimate goal of our thesis. We also wish to remove constraints on the body pose and background subtraction that have been used in previous sections. Three immediate questions are worthy of consideration:

Q1. What is the best strategy to expand the configuration space?

Instead of expanding the configuration space by one or two landmarks at each step, we may grow more landmarks (or even a whole body part) at a time. This may lead to improvements in both efficiency and reliability of the localization system. The best choice depends on the balance between how much uncertainty we can remove by collecting new image information, versus how much uncertainty will be introduced by expanding the configuration space. In this chapter, we provide a solution by employing a hierarchical decomposition of the configuration space, reflecting the idea of a coarse-to-fine search strategy.

Q2. Is background subtraction a must?

SMC is essentially a probabilistic version of beam search [71], and lacks a "look-ahead" mechanism. Given a limited number of particles, SMC is prone to diverge if no strong constraints, *e.g.,* from background subtraction, are available during the early stage of the search. This difficulty is compounded by Monte-Carlo variance [24]. For this reason, we initialized the shape from the face region, which is the most visually informative part of a human body. One feasible "look-ahead" mechanism is to use bottom-up proposals, which can be facilitated by body part detectors (see Figure 5.1 for some examples). However, building a robust part detector is difficult due to the simple structure and limited image support for each body part alone, especially in situations of self-occlusion and low resolution. A larger support region becomes available when detecting multiple parts as a



(a) hierarchical grouping [38] (b) CDT graph [67] (c) normalized cut [59]

(d) eigen-template [76] (e) Support Vector Machines [69]

Figure 5.1: Example body part detectors proposed in the literature.

whole [54, 92]. As an extreme case, global body models can be used [23, 98]. However, most work in this direction focuses on human detection rather than body part localization, and is limited to walking or standing poses. In this chapter, we propose to generate bottom-up proposals by finding *partial bodies* (*e.g.,* the whole leg, or all parts except the head) in arbitrary poses. These proposals are used to initialize and guide the top-down SMC inference. As a side product, foreground masks can be obtained from the intermediate output of the bottom-up process.

Q3. Can we train on walking data and test on an arbitrary pose?

We emphasize the statistical modeling of body shape deformation. To this end, the model prior was learned from a large number of labeled real gait images. This labeling is time consuming and requires considerable human interaction. The problem arises, therefore, of how to avoid acquiring new training images with arbitrary poses and minimize this tedious labeling process. Several methods are introduced:

- Relax models learned from gait images by, *e.g.,* expanding the allowed range of joint angles and part deformation;
- Use virtual shape examples (Section 4.5.2) to increase the deformation ability of the model;
- Compute marginal distribution of different parts, and assemble results from different viewpoints.

Using a combination of these methods, we are able to handle arbitrary pose without further data collection.

5.2 Overview of the Approach

We start from the dense body model introduced in Chapter 4 (Figure 5.2c or Figure 4.2). The body shape is parameterized by the positions of landmarks densely and uniformly sampled along the body contours. Given the weak assumptions of our problem setting, a direct use of such a detailed model is problematic. We adopt a coarse-to-fine strategy by introducing a 3-level hierarchical model decomposition (Figure 5.2). A compact set of landmarks are identified (Figure 5.2b) that characterize well the nonrigid and articulated body deformation with reduced complexity. The remaining landmarks are considered only after the inference on these key landmarks is completed.

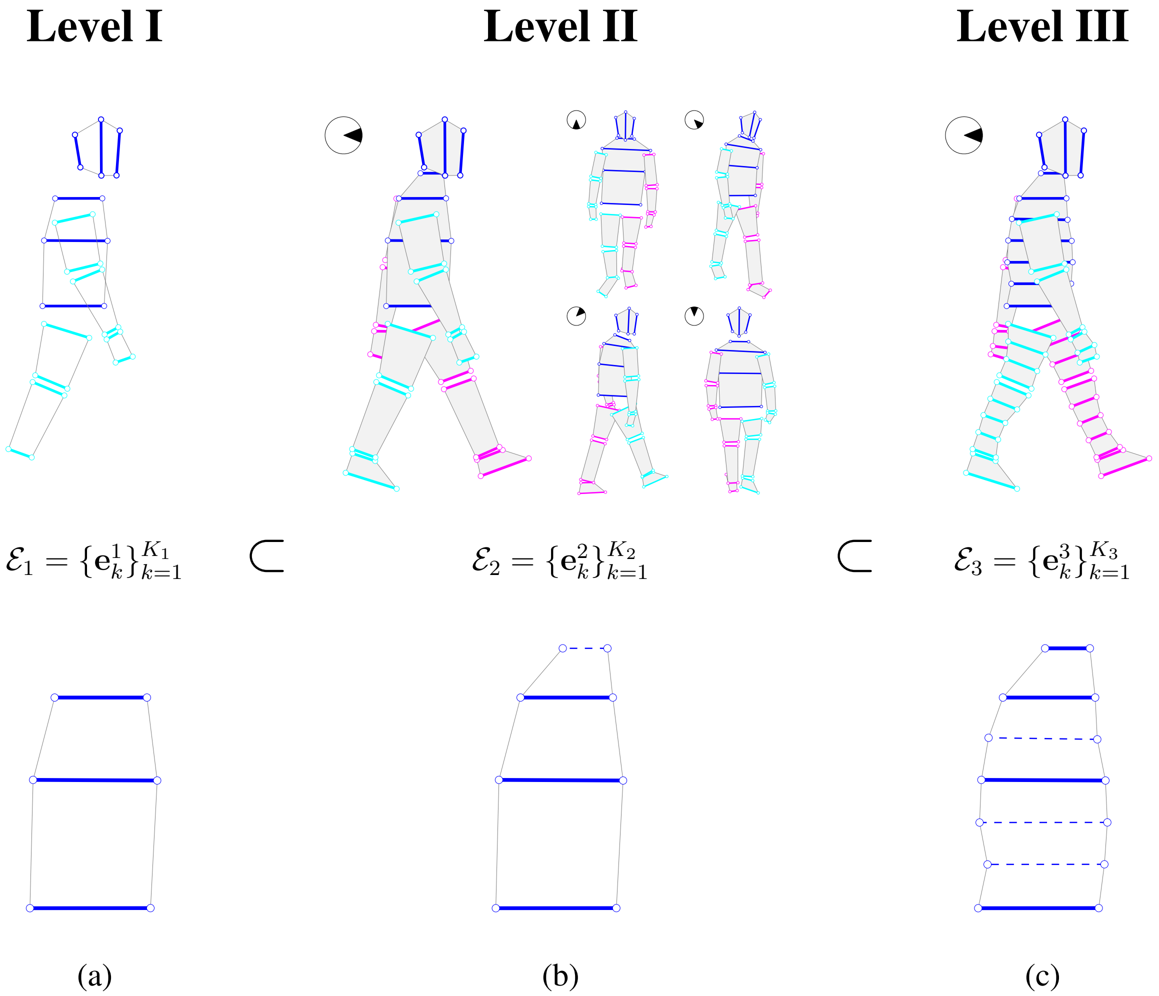

Level I
Level II
Level III
$\mathcal{E}_1 = \{\mathbf{e}_k^1\}_{k=1}^{K_1}$
$\subset$
$\mathcal{E}_2 = \{\mathbf{e}_k^2\}_{k=1}^{K_2}$
$\subset$
$\mathcal{E}_3 = \{\mathbf{e}_k^3\}_{k=1}^{K_3}$
(a)
(b)
(c)

Figure 5.2: A three-level hierarchy of body models: (a) View-independent tree-structured model. (b) Mixture model, with eight view-dependent components from angles uniformly distributed in $[0, 2\pi]$. Only five basic components are shown. (c) Boundary model, *i.e.,* a mixture model similar to (b) but with increased landmark density. The model at level m is defined on $\mathcal{E}_m$, a set of K_m line segments (drawn in bold and color) that divide the body shape into quadrilaterals. The three levels are designed with a nested hierarchical structure ($\mathcal{E}_1 \subset \mathcal{E}_2 \subset \mathcal{E}_3$) in order to facilitate a coarse-to-fine search. This hierarchical nature is best illustrated by the incremental refinement of the torso shape (see 2nd row).

To locate key landmarks, we employ the mixture model approach described in Chapter 4, which handles self-occlusion, anthropometric constraints, and large viewpoint changes. Inference is performed by a simple top-down stochastic search (SMC).

To facilitate the top-down search, we introduce a tree-structured model defined on a further reduced set of landmarks (Figure 5.2a). This model is fit by deterministic search using Dynamic Programming (DP). DP alone has been used for body localization in the past [28, 65]. In our work, DP and SMC complement each other by searching in opposite directions in the configuration space. At each step of SMC search, there is a "conjugate" proposal map from the DP output that encodes the probabilities of partial body configurations that have not been visited by SMC. The proposal maps effectively initialize and guide the SMC inference, similar to the use of heuristic functions in A* search [71]. This hybrid strategy of combining deterministic and stochastic search ensures both the robustness and efficiency of DP, and the accuracy of SMC.

The output of the hybrid search is a set of shape samples. Each sample is associated with a particular viewpoint. One strategy to summarize the output is "winner-take-all", *i.e.,* to first identify the viewpoint with the most samples or the highest fitting scores, and then apply mode analysis only on the samples associated with that viewpoint. There are several arguments against this choice in the situation of arbitrary pose:

1. In some cases body parts are in such a pose that they should be best explained by different viewpoints;

2. The samples labeled with a sub-optimal viewpoint may fit poorly on legs but fit well on the torso. Discarding them would be a waste of resources;

3. There is inherent ambiguity among different viewpoints (*e.g.,* front and back facing targets have very similar boundary shapes due to human body bilateral symmetry).

Based on these considerations, we compute the final output by *viewpoint combination* rather than viewpoint selection.

We train the model hierarchy on hand-labeled gait images from the CMU Mobo Database. A large number of virtual examples were also generated (Section 4.5.2) to increase the deformability of the shape prior. We obtain promising test results on over 100 cluttered, single images with varying poses including walking, dancing and various sports activities (see Figure 1.4c for a sample of images used).

5.3 Hierarchical Models

Our proposed system employs three models with increasing levels of complexity: the tree-structured model, the mixture of view-dependent models, and the boundary model (Figure 5.2). In this section, we discuss the definitions of prior and likelihood terms for each model. Section 5.4 presents the inference process by a hybrid strategy combining deterministic and stochastic search in a coarse-to-fine manner. A flowchart of the complete algorithm is depicted in Figure 5.3.

We represent the body shape by a set of piecewise linear boundary curves, or equivalently by a set of L landmarks $\mathbf{v}_{1:L} = \{\mathbf{v}_l\}_{l=1}^{L}$. Landmarks on the external boundary curves of each body part are paired into $K = L/2$ line segments, $\mathbf{e}_{1:K} = \{\mathbf{e}_k\}_{k=1}^{K}$, which divide the body shape into quadrilaterals (Figure 4.2). These line segments constitute the basic elements of the hierarchical model. We divide $\mathbf{e}_{1:K}$ into P sequentially ordered *parts*, $\mathcal{W} = \{W_p\}_{p=1}^{P}$, where $W_p = \{\mathbf{e}_{p,k}\}_{k=1}^{K_p}$ consists of K_p sequentially ordered line segments. W_p is virtually attached to a particular parent part $W_q (q < p)$ through two edges $(\mathbf{e}_{p,1}, \mathbf{e}_{q,K_q})$, which constitute a flexible *joint*. The P parts are connected into a "tree" structure by a total of $(P - 1)$ joints $\mathcal{J} = \{(p, q)\}$. The shape can be traversed sequentially by visiting the line segments $\{\mathbf{e}_{1,1} \cdots \mathbf{e}_{1,K_1}\}\{\mathbf{e}_{2,1} \cdots \mathbf{e}_{2,K_2}\} \cdots \{\mathbf{e}_{P,1} \cdots \mathbf{e}_{P,K_P}\}$.

We further divide line segments into three nested subsets $\mathcal{E}_1 \subset \mathcal{E}_2 \subset \mathcal{E}_3$, on which a 3-level hierarchical model is defined (Figure 5.2). Each level is indexed by a superscript. For example, the k-th segment in the p-th part of the m-th level is denoted as $\mathbf{e}_{p,k}^{m}$. Segments belonging to the p-th part of the m-th level but not to the previous level are denoted as $\mathbf{e}_{p,:}^{m\setminus(m-1)}$. The superscript will be dropped for simplicity when it can be easily determined from the context.

5.3.1 View-independent Tree-structured Model

The tree-structured model is defined on a small set of line segments that capture the basic body structure (Figure 5.2a or Figure 5.4a). These segments are grouped into 7 body parts $\{$head, torso, thigh, calf, upper arm, lower arm, hand$\}$. The head and torso contains three line segments each, while all other parts contains two each, which brings the total number of line segments to 16. Note that the topology is simplified to only one leg and one arm. Left and right legs/arms are mapped to the same line segments. The model is made view-independent by pooling together training data from all viewpoints.

We design the prior and likelihood functions in such a way that it becomes possible to obtain globally optimal solutions by deterministic search. These solutions are then used to

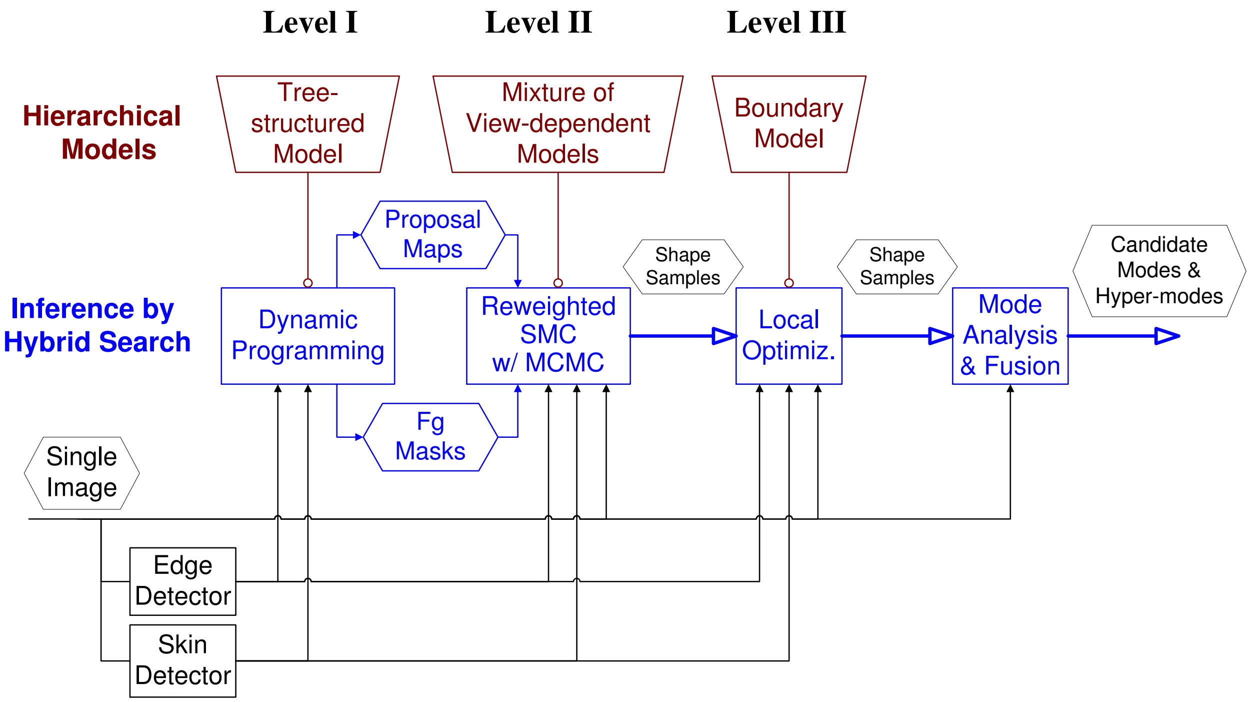


Figure 5.3: Algorithm flowchart. The input of the system is a single image, from which two image cues are extracted: edge gradient map, and skin/hair color mask (very lenient). A hybrid strategy combining deterministic (DP) and stochastic (SMC) search is conducted over a 3-level hierarchy. Inference is done in three steps: 1) A tree-structured model is fit to the input image by DP. The search starts from the bottom (feet) to the top (head). The output is a series of proposal maps, together with foreground masks for different body parts. 2) A mixture model is fit by SMC. The search starts from the top (head) to the bottom (feet). Proposal maps from DP are combined with the prior terms of the mixture model into an improved proposal function for the SMC search, while the foreground masks generated from DP outputs are utilized in the computation of SMC importance weights. The output is a set of shape samples. 3) A detailed boundary model is fit by local optimization, initialized by the SMC output. This 3-step hybrid search is followed by a mode analysis and fusion module to generate a number of candidate modes as the output of the localization system, which can be further clustered into a few compact hyper-modes.

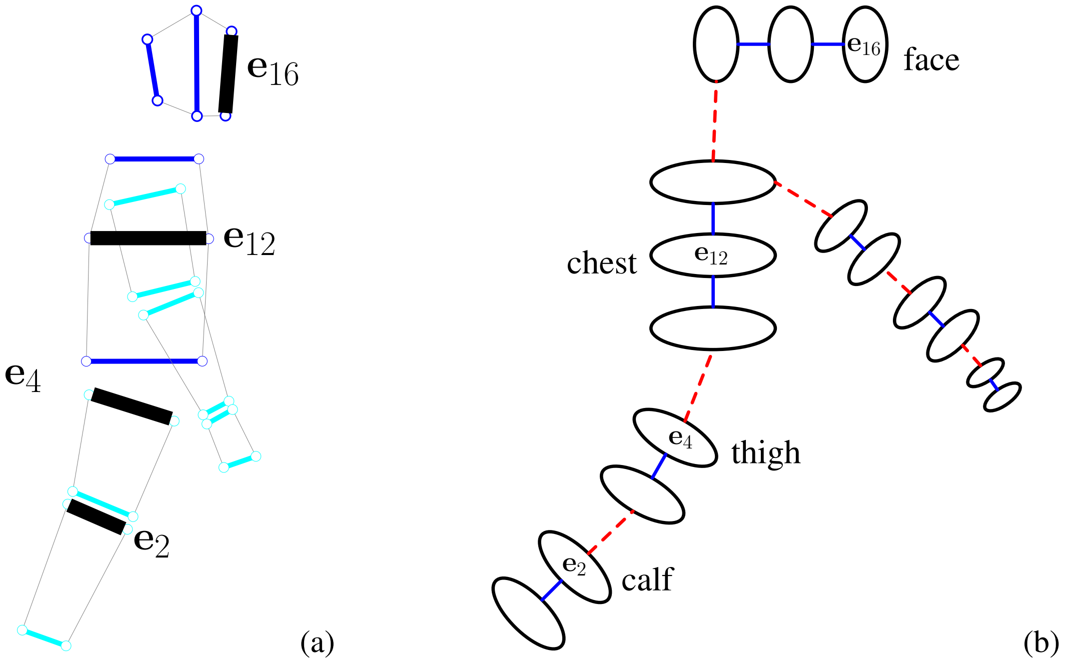


Figure 5.4: Tree-structured model. (a) Physical topology. (b) Probabilistic structure. Each node in (b) corresponds to an internal line segment in (a). Four example correspondences are labeled $\{\mathbf{e}_{16}, \mathbf{e}_{12}, \mathbf{e}_4, \mathbf{e}_2\}$. Blue (solid) links in (b) denote part connections while red (dotted) links denote joint connections.

initialize and guide the inference of more complex models.

Given two adjacent line segments $\mathbf{e}_{k-1}$ and $\mathbf{e}_k$, the deformation between them is parameterized by a similarity transform that maps $\mathbf{e}_{k-1}$ to $\mathbf{e}_k$ in the local coordinates of $\mathbf{e}_{k-1}$. The prior of the model is given by,

$$\begin{aligned} H(\mathbf{e}_{1:K}) &= \prod_k H(\mathbf{e}_{k-1:k}) \\ &= \prod_k p(x_k)p(y_k)p(\rho_k)p(\theta_k) \end{aligned} \tag{5.1}$$

where (x_k, y_k) is translational offset, ρ_k is relative scale and θ_k is rotation angle. Note that parts and joints are parameterized in the same way without any constraint on the form of deformation. $p(x)$, $p(y)$, $p(\rho)$ and $p(\theta)$ are modeled as histograms learned from multi-view training data.

The likelihood of the model is given by,

$$
\begin{aligned}
G(\mathbf{e}_{1:K}) &= \prod_k G(\mathbf{e}_{k-1:k}) \\
&= \prod_k \begin{cases} 1 & \text{if } \mathbf{e}_{k-1:k} \text{ is a joint} \\ \phi^s(\mathbf{e}_{k-1:k})\phi^e(\mathbf{e}_{k-1:k}) & \text{otherwise} \end{cases}
\end{aligned}
\tag{5.2}
$$

The skin potential ϕ^s is defined on head and hand segments, and is computed as the product of skin/hair probabilities at fixed positions in the quadrilateral $\mathbf{e}_{k-1:k}$. Note that the skin detector is very lenient due to the simultaneous detection of both skin and hair. The edge potential ϕ^e is computed as the average boundary probability along the two segments connecting $\mathbf{e}_{k-1}$ and $\mathbf{e}_k$.

This probabilistic model has a simple tree-structure, as depicted in Figure 5.4b. Given the partial body configuration $\mathbf{e}_{1:k}$, the marginal posterior $Q(\mathbf{e}_k)$ has a recursive form,

$$
Q(\mathbf{e}_k) = \sum_{\mathbf{e}_{1:k-1}} H(\mathbf{e}_{1:k})G(\mathbf{e}_{1:k}) = \sum_{\mathbf{e}_{k-1}} H(\mathbf{e}_{k-1:k})G(\mathbf{e}_{k-1:k})Q(\mathbf{e}_{k-1}). \tag{5.3}
$$

5.3.2 Mixture of View-dependent Models

The mixture model is defined on a compact set of line segments that well characterize the articulated body deformation (Figure 5.2b). These line segments are grouped into 14 parts {head, torso, left/right thigh, left/right calf, left/right foot, left/right upper arm, left/right lower arm, left/right hand}. Each part has the same topology as its counterpart in the tree-structured model of the previous level, except that one line segment is added to the torso to better model the shoulder. Note that left and right limbs here are mapped to the same limb in the tree-structured model. The mixture model consists of eight part-based component models. Each component works for a small range of view angles. Details of this mixture and part-based decomposition can be found in Chapter 4, and the main results are summarized below for completeness.

We design the prior and likelihood functions of the component model in such a way that self-occlusion and anthropometric constraints can be handled, while the model can still be fit via simple stochastic search. We define two deformation mechanisms:

1. Shape variation of the parts, which is parameterized by Procrustes residuals $\mathbf{r}_{p,:} = \{\mathbf{r}_{p,k}\}_{k=1}^{K_p}$ where $\mathbf{r}_{p,:}$ is modeled as a multivariate normal;
2. Articulated movement of the joints, which is parameterized similar to the tree-structured model.

Accordingly, the prior of a component model is decomposed as,

$$p(\mathbf{e}_{1:K}) = \prod_p p(x_p, y_p)p(\rho_p)p(\theta_p) \prod_i p(\mathbf{r}_{p,i}|\mathbf{r}_{p,1:i-1}).$$

To impose anthropometric constraints on the relative lengths of the limbs, we introduce conditioning variables $\gamma_p = \|\mathbf{e}_{p,1}\|/l_1$ for parts and $\gamma_q = \|\mathbf{e}_{q,K_q}\|/l_1$ for joints, where l_1 is the length of the face line segment, $\mathbf{e}_{p,1}$ and $\mathbf{e}_{q,K_q}$ are line segments through which two parts are virtually attached. The final form of the prior is,

$$\begin{aligned} p(\mathbf{e}_{1:K}) \propto & \prod_{(p,q)\in\mathcal{J}} p(x_p, y_p|\gamma_q)p(\rho_p|\gamma_q)p(\theta_p) \\ & \prod_i p(\mathbf{r}_{p,i}|\mathbf{r}_{p,1:i-1}, \gamma_p). \end{aligned} \tag{5.4}$$

We define potential functions on a set of clusters $\mathcal{C}$ that cover the body shape. Each cluster $C \in \mathcal{C}$ contains a small number of related line segments. Four types of potentials are defined based on edge, skin color, foreground mask, and region similarity. The foreground mask is generated from the intermediate output of the tree-structured model and will be described in Section 5.4.1. The likelihood is computed as the product of all types of potentials,

$$p(\mathcal{I}|\mathbf{e}_{1:K}) \propto \prod_z \prod_{C\in\mathcal{C}^z} \phi^z(\mathbf{e}_C), \tag{5.5}$$

where $z \in \{$edge, skin, foreground, region$\}$.

To accommodate the self-occlusion caused by viewpoint change, a depth ordering of parts is assigned to each view-dependent model. The depth order is considered in the computation of potentials, resulting in clusters that contain many line segments across different parts (*e.g.,* two overlapping legs).

The final posterior has a recursive form,

$$p(\mathbf{e}_{1:K}|\mathcal{I}) \propto \prod_k \Gamma_k \cdot \Phi_k, \tag{5.6}$$

where,

$$\Gamma_k = \begin{cases} p(x_p, y_p|\gamma_q)p(\rho_p|\gamma_q)p(\theta_p) & \text{if } \mathbf{e}_{k-1:k} \text{ is a joint} \\ p(\mathbf{r}_{p,i}|\mathbf{r}_{p,1:i-1}, \gamma_p) & \text{otherwise} \end{cases}$$

$$\Phi_k = \prod_z \prod_{C\in\mathcal{C}_k^z} \phi^z(\mathbf{e}_C)$$

and $\mathcal{C}_k^z$ is the set of clusters newly "activated" (*i.e.,* complete covered) at step k.

5.3.3 Mixture of Detailed View-dependent Models

The boundary model has the same component topology as the mixture model of the previous level, except that landmarks are more densely sampled along the body contours (Figure 5.2c). The number of landmarks is increased from 62 to 92. The new landmarks are introduced to characterize the detailed boundary deformation of each body part. We assume that this local deformation is conditionally independent. Thus the shape prior of the model can be written as,

$$p(\mathbf{e}^3) = p(\mathbf{e}^2)\, p(\mathbf{e}^{3\setminus 2}|\mathbf{e}^2) = p(\mathbf{e}^2) \prod_p p(\mathbf{r}_p^{3\setminus 2}|\mathbf{r}_p^2), \tag{5.7}$$

where $\mathbf{e}^{3\setminus 2}$ denotes those line segments belonging to the boundary model but not to the mixture model, and $\mathbf{r}_p$ denotes the Procrustes residuals of the p-th part. Note that $p(\mathbf{e}^2)$ is exactly the mixture model prior defined by Equation (5.4). In the current implementation, we simply model $p(\mathbf{r}_p^{3\setminus 2}|\mathbf{r}_p^2)$ as a multivariate normal.

5.4 Inference by Hybrid Search

We adopt a hybrid strategy that combines deterministic and stochastic search in a coarse-to-fine manner in the configuration space of 92 landmarks. The first two models (*i.e.,* tree-structured model and mixture model) are coupled via two "conjugate" sequential search processes (*i.e.,* DP and SMC). Proposal maps from DP inference of the first model are integrated into an improved proposal function for the SMC inference of the second model, while the foreground masks generated from DP are utilized in the computation of SMC importance weights. The output of SMC is a set of shape samples, which directly initialize the optimization of the third model (*i.e.,* boundary model). Finally, mode analysis and fusion are applied to generate a few candidate modes and hyper-modes as the final output. A flowchart of the complete algorithm is shown in Figure 5.3.

5.4.1 Dynamic Programming

We first fit the tree-structured model to the input image. The marginal posterior $Q(\mathbf{e}_k)$ defined in Equation (5.3) has a simple recursive form and can be computed by DP. The basic computation in $Q(\mathbf{e}_k)$ evaluation is a weighted sum over quantization bins of $\mathbf{e}_{k-1}$, which can be considered as a convolution. The complexity of DP is $O(N^2)$, where N is the number of bins. If we choose a quantization resolution of {32, 32, 16, 32} for $\{x$, y,

ρ, $\theta\}$ respectively, N becomes $32 \times 32 \times 16 \times 32 \sim 10^5$. Therefore, the cost of a naive DP implementation is unacceptable. Some fast algorithms for DP exist [28], but are not applicable here because: 1) H and G are modeled by non-Gaussian histograms, and 2) the convolution kernel ($H \cdot G$) is not homogeneous. Fortunately, there are two properties that can be used for acceleration:

1. Both H and G have limited spatial support, so that most bins can be pruned during the convolution;

2. H is decomposable, so that the algorithm can be modified to do four 1D convolutions (over x, y, ρ and θ respectively).

Note that the second property only works for joints because G is not decomposable. As a result, part likelihood evaluation becomes the bottleneck for the speed of our DP implementation, which is why we only use simple potential functions (*i.e.,* edge and skin mask) in the tree-structured model.

The output of DP, $\{Q(\mathrm{e}_k)\}_{k=1}^K$, constitutes a series of proposal maps that encode the probabilities of partial body configurations (see Figure 5.5 for some examples). We compute a large number of shape samples by sampling backwards from $Q(\mathrm{e}_K)$ to $Q(\mathrm{e}_1)$, and learn an appearance model for each of {head, torso, thigh, calf} respectively. First, a weighting mask is constructed, where the pixel value is the number of sample shapes covering that pixel. Next, weighted positive and negative training samples (with RGB value as features) are collected by thresholding the weighting mask. Finally, a discriminative quadratic classifier is trained (using the implementation in [21]). The classifier is applied to the original image to obtain a binary foreground mask. Several examples are displayed in Figure 5.6. Note that the foreground masks can be very noisy and contain large false positive areas. This is because

- The discrimination power of a pixel-based quadratic classifier is limited;

- Each foreground classifier is trained using a different subset of the background as negative samples.

However, the large false positive area is not a problem as long as the mask gives a good segmentation of the area around its targeted body part, which is actually explored by the following modules. We also perform a validity check on each scanline to alleviate this problem. A scanline is valid if the ratio between the number of its foreground and background pixels is less than a threshold (set to 4 as default in our experiments). Only valid

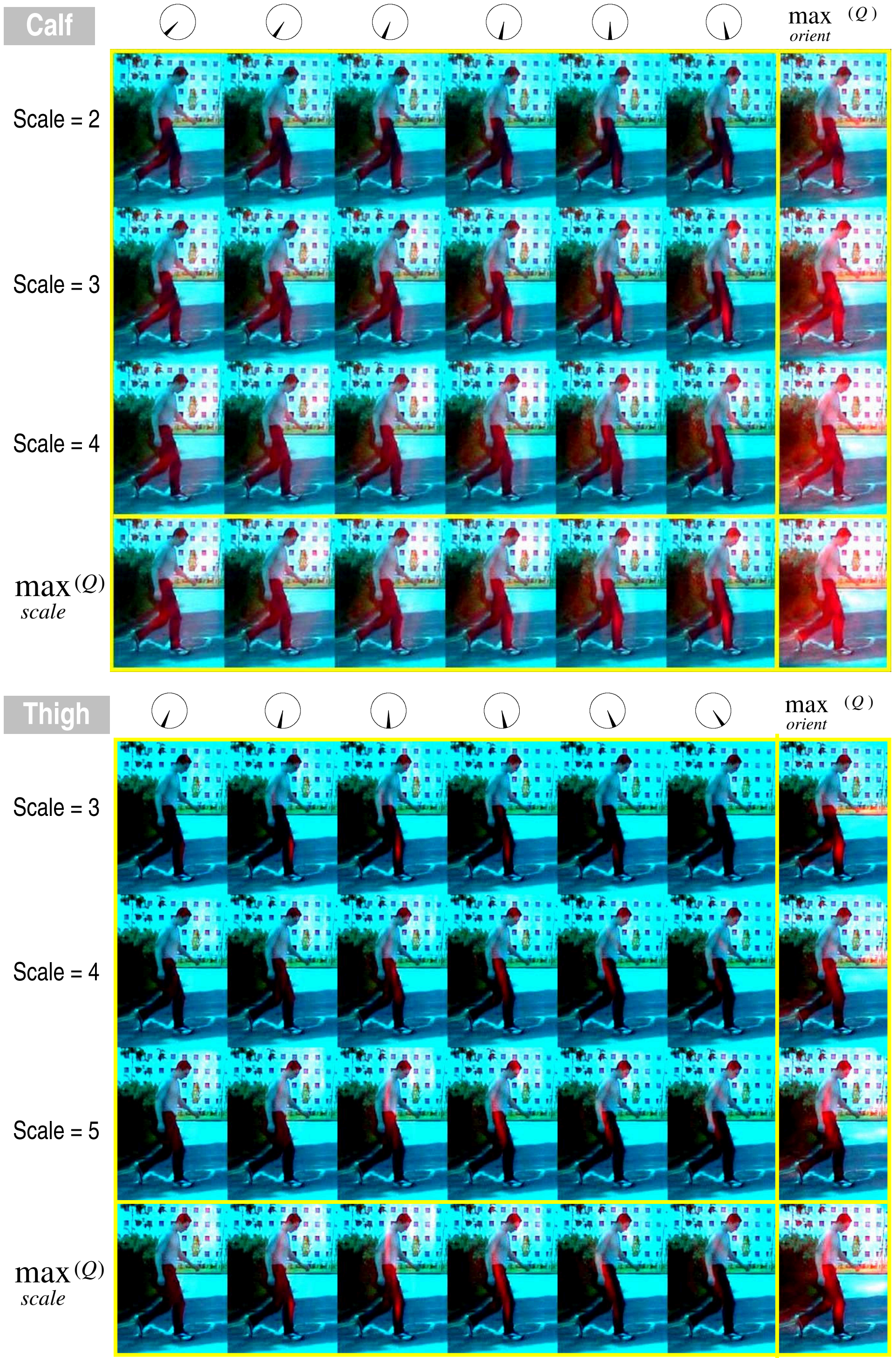
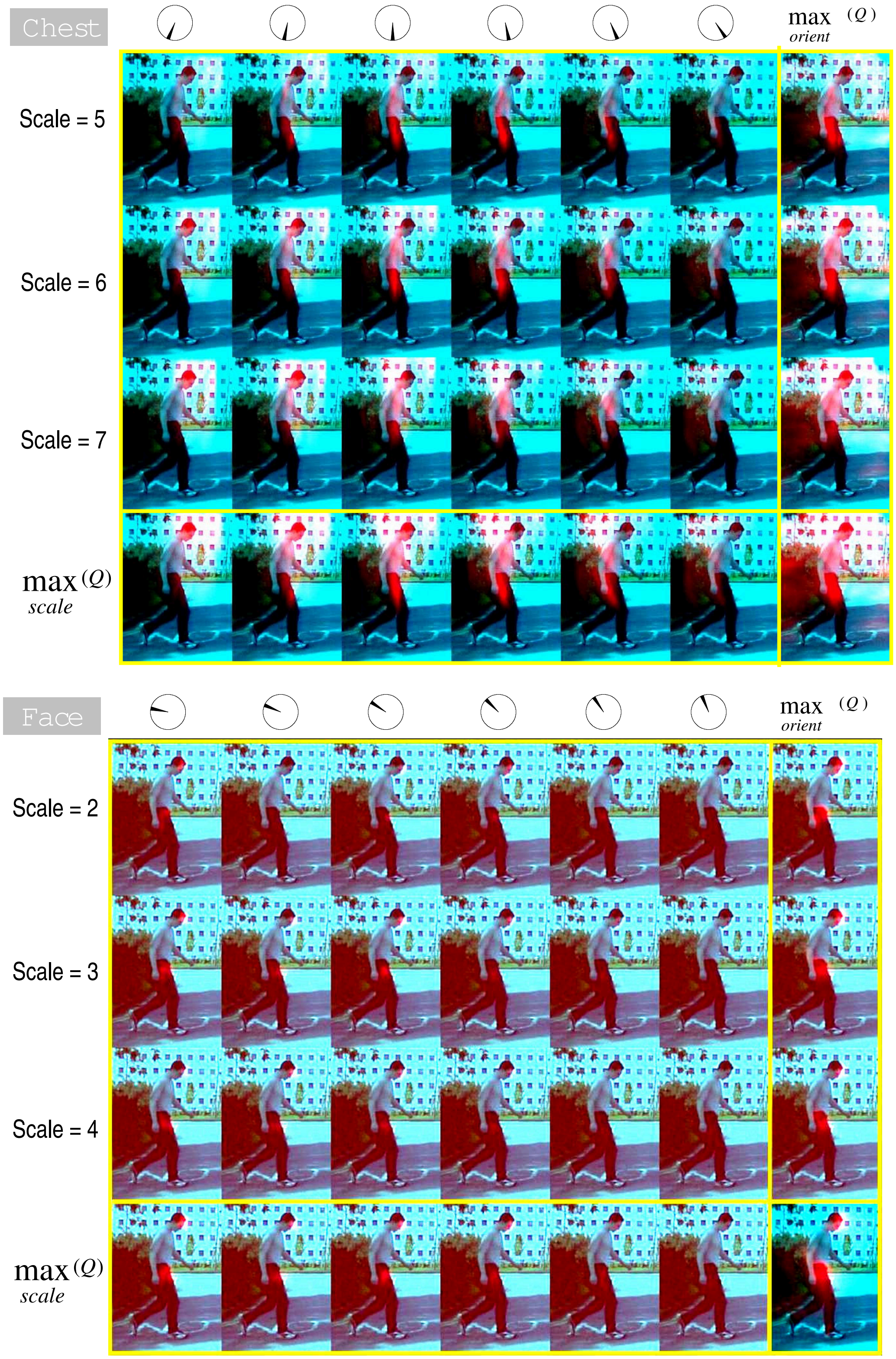


Figure 5.5: Visualizing the marginal posterior Q defined in Equation (5.3) for four line segments: calf, thigh, chest and face (defined in Figure 5.4). Each image shows the probability distribution of the center position of a line segment, at a particular scale and orientation. This 2D distribution is modulated onto the red channel of the original image.

To be continued on pp. 70.



(Figure 5.5 on pp. 69 Continued) — DP proposal maps — (1/1)

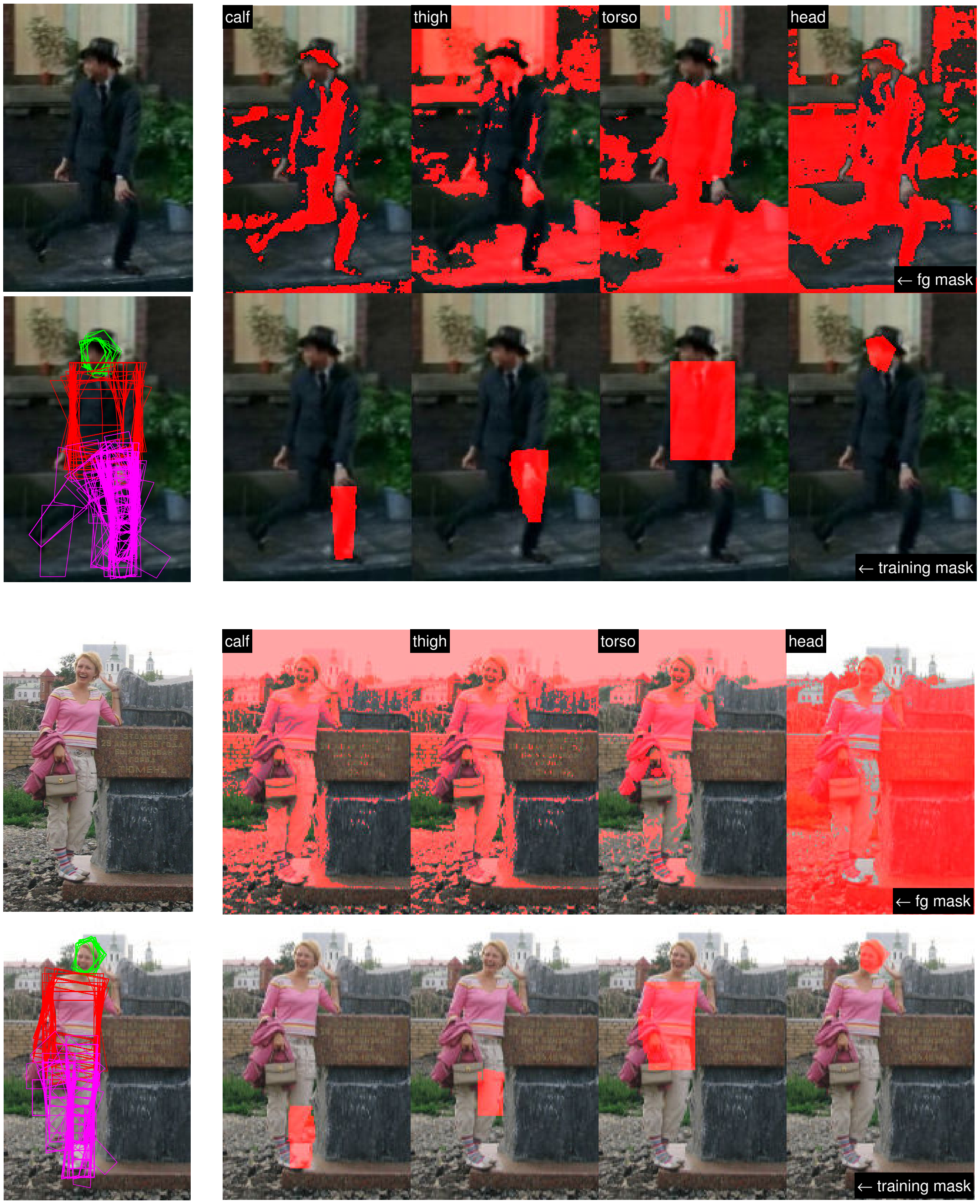
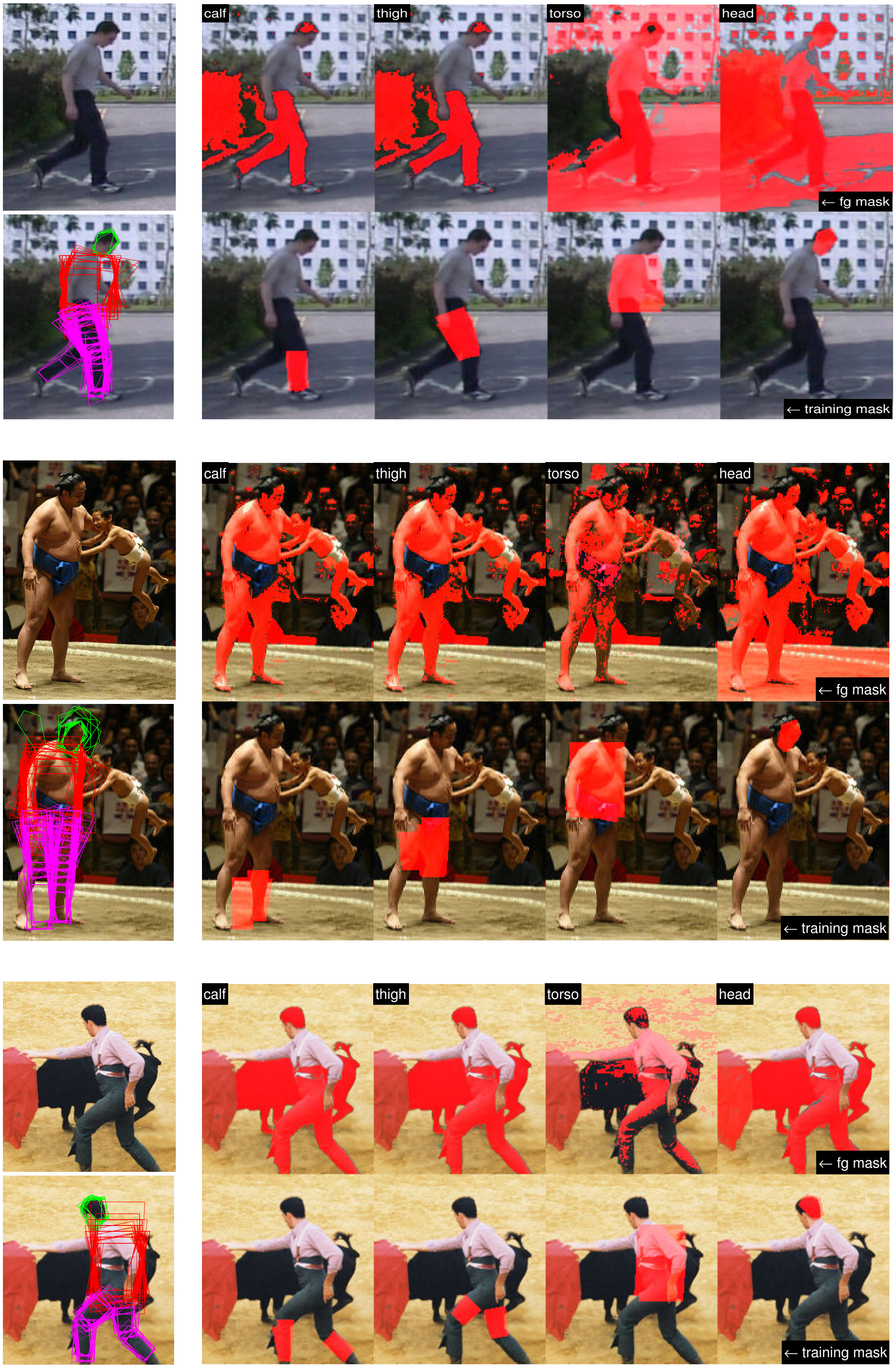


Figure 5.6: Example foreground classification based on DP outputs. First, a number of shape samples are computed by backward sampling (left, bottom). Then, weighting masks are constructed for four parts {calf, thigh, torso, head} respectively (right, bottom). Finally, four discriminative classifiers are learned using weighted training pixels, and applied to the original image to obtain binary foreground masks (right, top). Both weighting and foreground masks are modulated onto the red channel of the original image.

To be continued on pp. 72.



(Figure 5.6 on pp. 71 Continued) — Backward sampling, foreground masks — (1/1)

lines are fed to the following modules as a strong cue. A similar method has been used in [65] to construct appearance models for human tracking.

5.4.2 Reweighted SMC with MCMC and Annealing

We next fit the mixture model to the input image. Equation (5.6) shows that the posterior of the view-dependent model has a recursive form, thus can be fit via Sequential Monte Carlo. A naive implementation of SMC uses the shape prior Γ as the proposal, and the likelihood potential Φ as the importance function. However, without strong constraints such as background subtraction, the search is prone to diverge. Here we adopt the deterministic search using DP to reduce the effect of Monte Carlo variance, and to provide a "look-ahead" mechanism. DP and SMC are designed to search in opposite directions in the configuration space, such that at step k of SMC, there is a "conjugate" proposal map $Q(\mathbf{e}_k)$ which encodes the probabilities of partial body configurations $\mathbf{e}_{k+1:K}$ that SMC has not yet visited (see Figure 5.7 for an illustration). $Q(\mathbf{e}_k)$ plays a similar role in SMC as heuristic

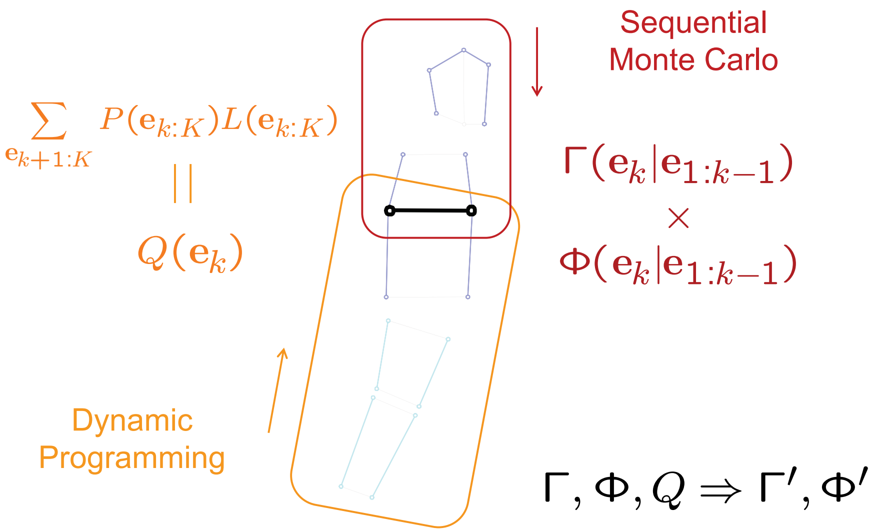

Sequential
Monte Carlo
$\sum_{\mathbf{e}_{k+1:K}} P(\mathbf{e}_{k:K})L(\mathbf{e}_{k:K})$
$||$
$Q(\mathbf{e}_k)$
$\Gamma(\mathbf{e}_k|\mathbf{e}_{1:k-1})$
$\times$
$\Phi(\mathbf{e}_k|\mathbf{e}_{1:k-1})$
Dynamic
Programming
$\Gamma, \Phi, Q \Rightarrow \Gamma', \Phi'$

Figure 5.7: Illustration of the conjugate inference processes of DP and SMC. At step k, the proposal map $Q(\mathbf{e}_k)$ from DP is combined with the prior term Γ_k and likelihood term Φ_k of the mixture model into an improved proposal function Γ'_k and importance weight Φ'_k as a guidance of the SMC search. Note that $Q(\mathbf{e}_k)$ is computed from line segments in the lower part of the body that have not been visited by SMC (*i.e.*, $\mathbf{e}_{k+1:K}$), while Γ_k and Φ_k are computed from line segments in the upper part that have already been visited (*i.e.*, $\mathbf{e}_{1:k-1}$).

lookahead functions in A* search.

At the beginning, we initialize the SMC procedure by sampling $\mathbf{e}_1$ from $Q(\mathbf{e}_1)+U(\mathbf{e}_1)$, where U is a regularization term defined as a uniform distribution over a specified image region. At step k, we use a reweighted importance sampling [82] to draw samples from a distribution closer to the true posterior. The proposal is modified as,

$$\Gamma_k' = \Gamma_k \cdot Q(\mathbf{e}_k), \tag{5.8}$$

while the importance function is the same as regular SMC. Another reweighting procedure is applied after resampling, which multiplies the weights of all particles by $1/Q(\mathbf{e}_k)$, to keep the objective function unchanged.

To sample from Γ_k', we reformulate it as,

$$\Gamma_k' = \frac{\Gamma_k Q(\mathbf{e}_k)}{\int \Gamma_k Q(\mathbf{e}_k) d\mathbf{e}_k} \int \Gamma_k Q(\mathbf{e}_k) d\mathbf{e}_k. \tag{5.9}$$

The first term has an irregular distribution that can not be directly sampled. We employ MCMC to solve this problem, with Γ_k as the transition kernel. The second integration term $\int \Gamma_k Q(\mathbf{e}_k) d\mathbf{e}_k$ is difficult to compute, so we approximate it as $Q(\tilde{\mathbf{e}}_k)$, where $\tilde{\mathbf{e}}_k$ is the MCMC output. This term is used as a weight associated with $\tilde{\mathbf{e}}_k$.

Besides the guidance of DP, we also use an annealing procedure similar to that in Chapter 4 that gradually increases the peakiness of the likelihood term in order to avoid being trapped in local maxima during the early stage of the search.

We search in parallel through all the view-dependent models. The number of samples N_i associated with a particular viewpoint χ_i is proportional to its posterior probability, reflecting a mechanism of dynamic resource allocation. In practice, however, we often observed large fluctuation of N_i during the search, which negatively affects the quality of the estimate. In addition, there are many reasons (stated in Section 5.2) to maintain multiple models instead of a single "correct" one. Therefore, we divide all view-dependent models into three groups of ambiguous viewpoints: {front and back-facing}, {left-facing}, {right-facing}. A regularized resource allocation scheme is employed such that,

1. Resources of the viewpoints in the same group are always kept the same.
2. Resource reallocation for the three groups is applied only at selected steps, when enough discriminative information has been accumulated.

The robust allocation is achieved by maintaining a buffer of discrete distributions, which, multiplied by the number of samples (N_i), keep track of the posterior estimate of the viewpoint.

5.4.3 Local Optimization

Given the inference output of the mixture model, the boundary model can be fit by local optimization techniques. In the current implementation, a one-step importance sampling is employed. More complex and accurate segmentation techniques can be used, such as level sets [14], Markov Random Field and Graph Cut [47].

5.4.4 Mode Analysis and Fusion

The output of the hybrid search described in the previous three sections is a set of shape samples (on the order of 10^4 in our experiments) with prior and likelihood scores. Each sample is associated with a particular viewpoint (component model). Instead of choosing a single "correct" viewpoint, we employ a marginalization and combination scheme (Figure 5.8). First, sample shapes of the whole body are broken into parts. Next, the marginal distribution of each part is computed for each viewpoint or viewpoint combination, and the candidate modes are found. Finally, the part candidates from different viewpoints and viewpoint combinations are reassembled into the final output. This scheme is based on the following considerations:

1. Our model was trained on gait images. It is possible to increase the model's deformability by marginalization and reassembly.

2. Sometimes body parts are in such a pose that they should be best explained by different viewpoints.

3. Mode analysis in the subspace of parts instead of the high dimensional body configuration space requires fewer samples.

4. Combining ambiguous viewpoints together may increase the number of effective samples.

Instead of a full body part decomposition, we conduct mode analysis for three body groups $\{$torso/head, legs/feet, arms/hands$\}$. Each body group contains those parts that are strongly correlated. This leads to body group candidates that satisfy geometric constraints well, and thus greatly alleviate the problem of group assembly.

We first apply cluster analysis to get M modes for each body group. Each mode is summarized by the mean and covariance of a multivariate Gaussian distribution. M is typically set to 2 due to the flipping symmetry of left and right limbs. The likelihood surface of arms/hands is expected to be more complex, for which a larger mode number can

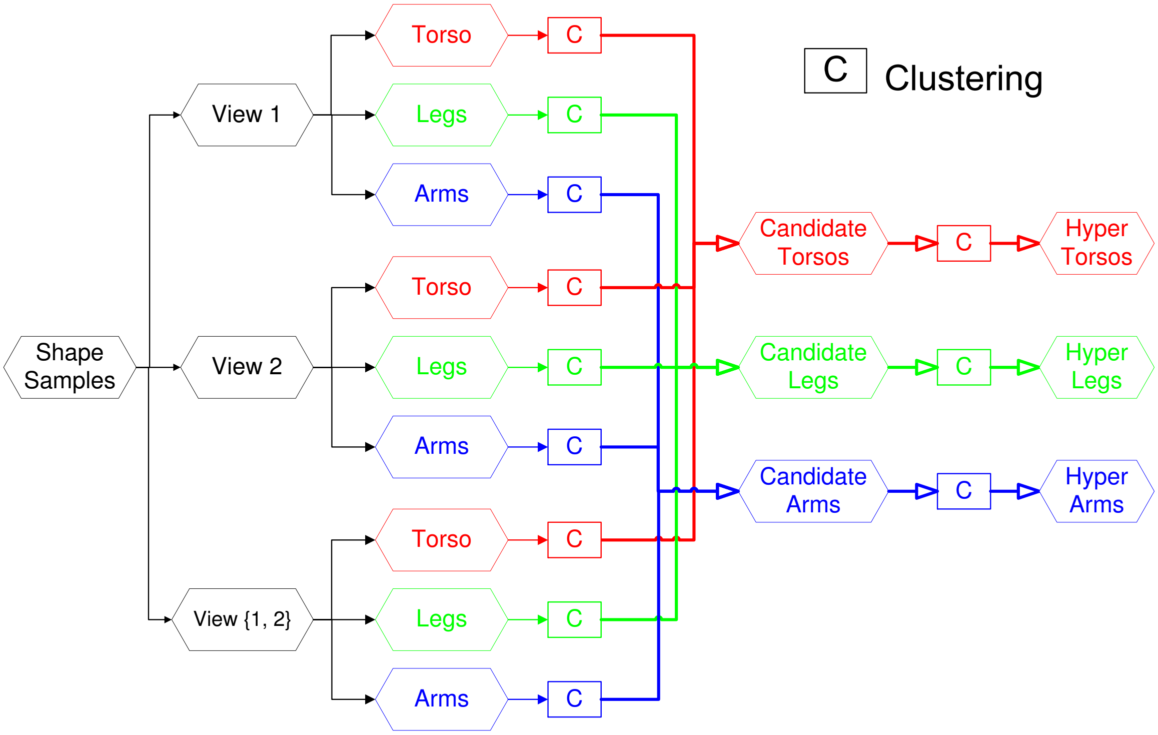


Figure 5.8: Flowchart of the mode analysis & fusion module. Only two viewpoints are used for illustration purpose. First, sample shapes of the whole body are broken into three body groups {torso/head, legs/feet, arms/hands}. Next, each body group is clustered into a number of candidate modes. This is done to each viewpoint respectively, and also to each ambiguous viewpoint combination. Finally, candidates from different viewpoints and viewpoint combinations are pooled, sorted, and clustered into visually compact hyper-modes.

be chosen ($M_{arm} = 6$ in our experiments). The clustering is done to the samples of each viewpoint respectively, and also to the samples pooled from each ambiguous viewpoint combination. Accordingly, we get on average around 10–30 candidate modes for each body group. The exact number varies depending on the number of component models that survive the search.

We then extract ridge and blob features for each candidate. Ridge/blob scores are used to prune arms/hands group. Those candidates whose scores are below the maximum score among all candidates for more than a threshold are discarded. Likelihood scores are used to prune the other two body groups in a similar way.

Finally we sort the remaining candidates by a linear combination of likelihood and ridge/blob scores. The resultant three sets of ranked candidates constitute the output of our localization system. Note that this output of body group candidates is different from that of commonly used part detectors. Experiments demonstrate that our group candidates well satisfy the geometric constraints between different groups (see Figure 5.11 for some examples). Therefore, given an ideal ranking function, the final output can be generated by combining the top-scoring candidate from each group. In contrast, bottom-up part detectors produce unorganized output that must be assembled with the help of some top-down model.

Because many candidates are very similar, we can also cluster them into visually compact hyper-modes. Details of this approach will be discussed in the Experiments section (Section 5.5.3).

5.5 Experiments

5.5.1 Training Hierarchical Models

We trained the 3-level model hierarchy using a set of hand-labeled gait images from the CMU Mobo Database. The images were captured using six synchronized cameras distributed evenly around a treadmill. As a result, virtual contours from an arbitrary viewpoint could be generated by interpolating the labeling of different views. We sampled the training shapes at 5^o intervals. Details of this data collection process can be found in Section 4.5.

The tree-structured model was trained by pooling together samples from all view angles. Each view-dependent model was trained with (both real and virtual) samples within a 90^o view range. This range was chosen deliberately large in order to increase the deformation ability of the shape prior. In addition, models trained on gait images were relaxed by expanding the allowed range of joint angles and part deformation in order to handle arbitrary poses.

5.5.2 Test Dataset

A total of 340 images were used to demonstrate the feasibility of the proposed approach. They are divided into three types:

1. 90 walking images (50 from USH outdoor gait database and 40 from CMU Mobo gait database). The USH data features an outdoor street scene and a single side viewpoint. The lighting condition is challenging due to serious shadowing and image compression. CMU Mobo data features an indoor environment. Background clutter and skin-like objects constitute major distractions. Details of these two databases can

be found in Sections 3.5 and 4.5. Note that background subtraction, though available, is not used in the following experiments.

2. 150 break dancing images from a Volkswagen TV advertisement (available online at `http://image.guardian.co.uk/sys-video/Media/video/2005/01/27/golfgti.mov`). The video is captured from a moving camera on a rainy night. The original frame is 428×240, from which the human target is roughly cropped out by hand. 150 frames are uniformly sampled for test purposes. Note that, although this is a video sequence, we did not use any motion cues.

3. 100 images collected from the web. Poses vary from walking to various sports activities (Figure 5.14).

5.5.3 Test Result

The output of our localization system contains three sets of candidates, categorized as torso, legs and arms. The average number of candidates per category is 10 for torso, 15 for legs, and 30 for arms. We are conservative in candidate pruning to insure a high true positive rate on the diverse and challenging test data.

Because many candidates are very similar, we cluster them into visually compact hyper-modes for better interpretation. We first use a linkage-based clustering method without specifying the number of clusters. As a result, the average number of hyper-modes reaches 2.9 for torso, 5.5 for legs and 10.6 for arms.

We then explore the possibility of less hyper-modes by visualizing the clustering results. The intuition is that the number of hyper-modes can be further reduced as long as they remain visually compact. We use the index $\sigma_{compact}$ to evaluate the compactness of a hyper-mode, which is defined as,

$$\sigma_{compact} = \frac{1}{(\text{torso height})}\sqrt{\frac{1}{J}\frac{1}{N}\sum_{j,n}\|\mathbf{v}_j^{(n)} - \bar{\mathbf{v}}_j\|^2}, \tag{5.10}$$

where $\mathbf{v}_:^{(n)}$ is the n-th candidate shape in the hyper-mode, $\bar{\mathbf{v}}_:$ is the average shape of all candidates, and j is the vertex index. $\sigma_{compact}$ computes the average standard deviation of all candidate positions of a vertex, normalized by the length of the torso. Shown in Figure 5.9 is a plot of the compactness of the "preferred" hyper-mode (*i.e.,* the one that contains the candidate that best matches human perception), versus the number of hyper-modes specified. Results at five numbers $\{2,4,5,6,10\}$ are also listed in Table 5.1. We observe that the accuracy is still satisfactory even when clustering the output into 2 (torso), 4 (leg) and 5

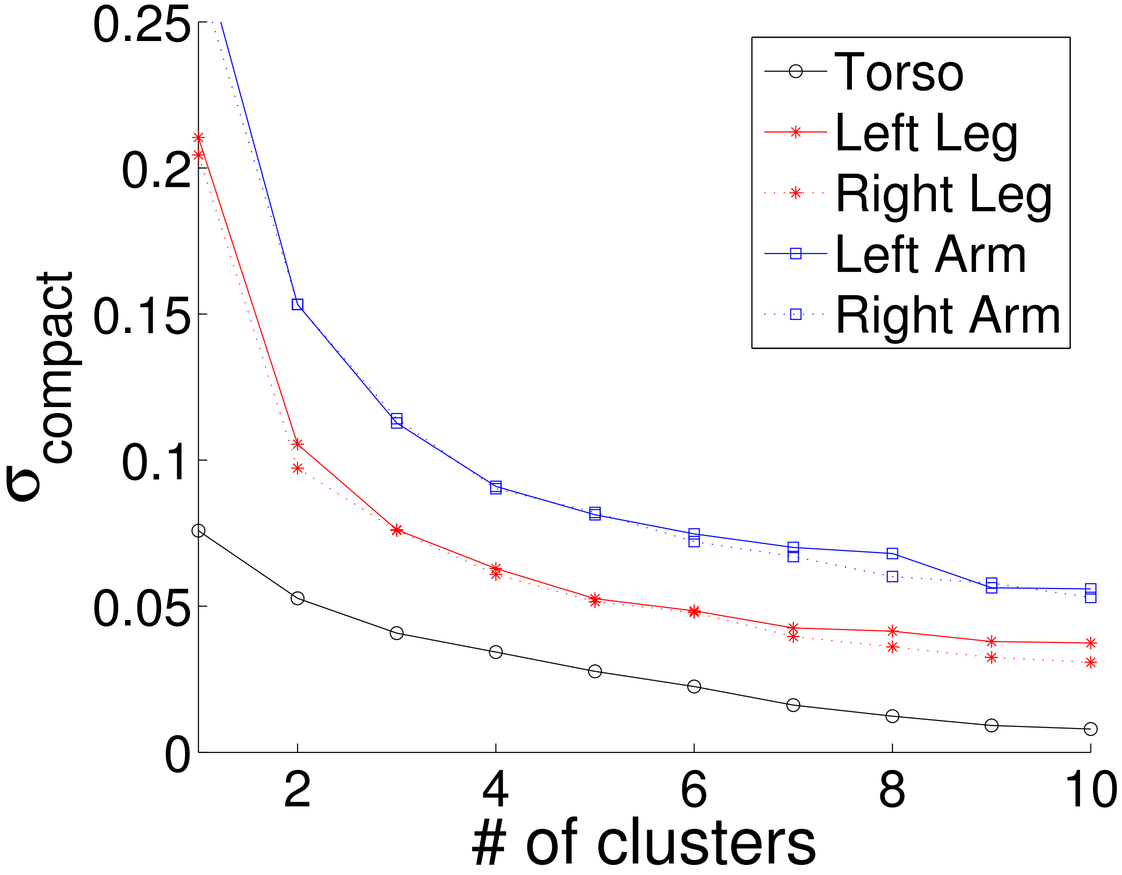


Figure 5.9: Compactness ($\sigma_{compact}$) of the "preferred" hyper-mode, as defined in Equation (5.10), versus the number of hyper-modes.

Table 5.1: Compactness ($\sigma_{compact}$) of the "preferred" hyper-mode, as defined in Equation (5.10), versus the number of hyper-modes. The settings marked as bold were chosen to compute the hyper-mode results in Figure 5.11.

# clusters	torso	lleg	rleg	larm	rarm
2	**0.0527**	0.1054	0.0972	0.1533	0.1533
4	0.0343	**0.0629**	**0.0609**	0.0910	0.0902
5	0.0277	0.0526	0.0515	**0.0813**	**0.0821**
6	0.0225	0.0484	0.0479	0.0748	0.0721
10	0.0080	0.0374	0.0308	0.0560	0.0530

(arm) hyper-modes, with corresponding compactness indices as 0.0527, 0.0619 and 0.0817. Figure 5.10 gives a visual interpretation of these numbers. And Figure 5.11 shows the complete candidate sets on 20 example images that are organized in this way. These candidates are sorted based on a combination of likelihood scores and blob/ridge features, as described in Section 5.4.4. "Preferred" modes are outlined with yellow frames. The ideal automated scoring function would have the top-scoring mode coinciding with the "preferred" mode. We observe from Figure 5.11 that,

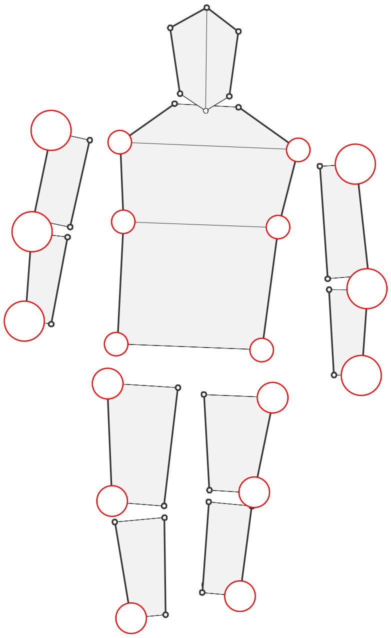


Figure 5.10: A visual interpretation of the compactness indices chosen in Table 5.1. Each red circle radius is set to the average standard deviation of the associated vertex.

1. The candidates satisfy the geometric constraints well, and a final body assembly can be constructed by simply picking a (top-scoring) candidate from each body group;

2. Each candidate set also contains similar neighbors of the "preferred" solution, and samples around other local maxima of the likelihood surface;

3. In most cases, good candidates of torso and legs can be found in the two top-ranked hyper-modes;

4. In many cases, candidates of higher ranks than the "preferred" ones are not significantly worse.

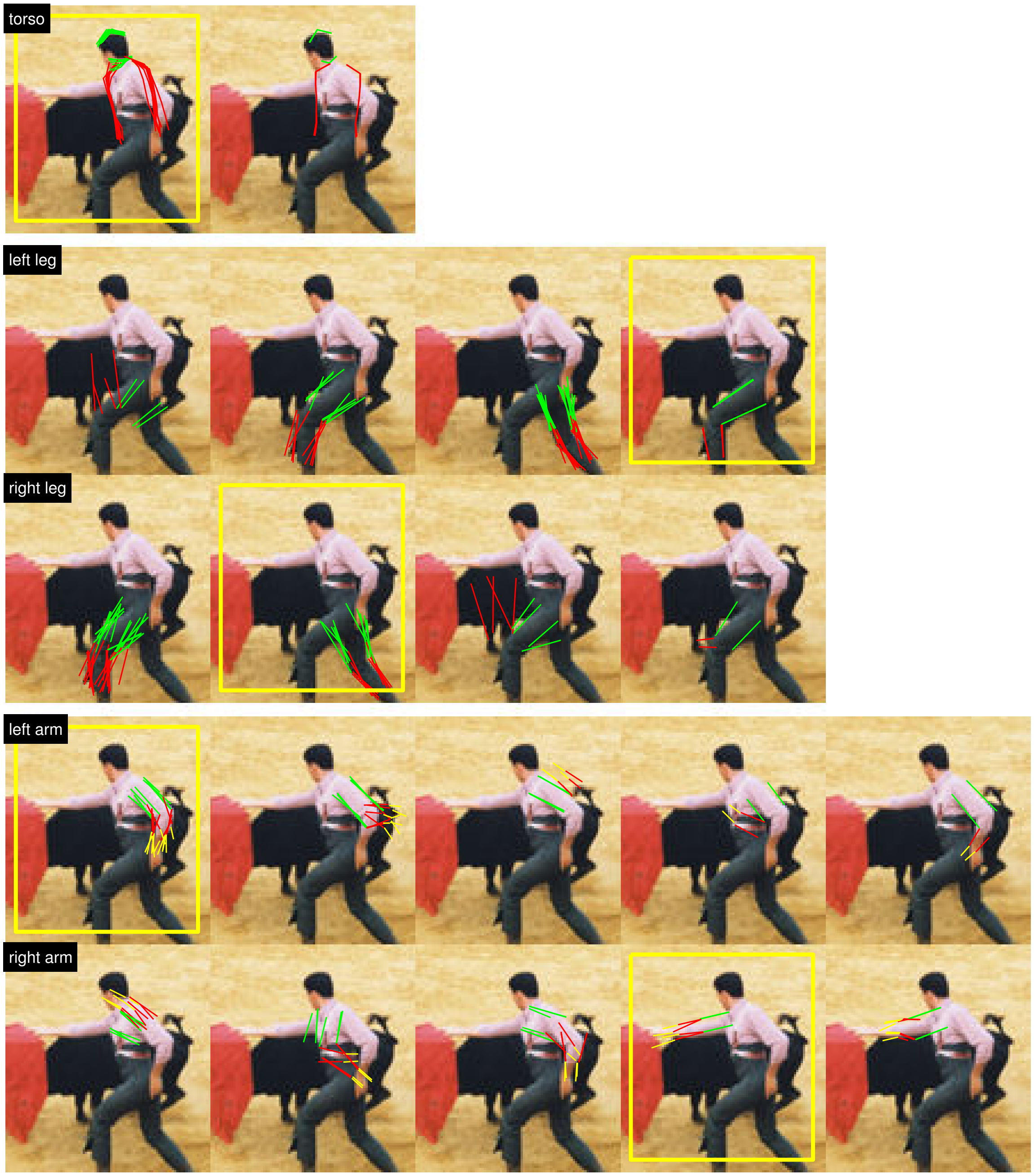
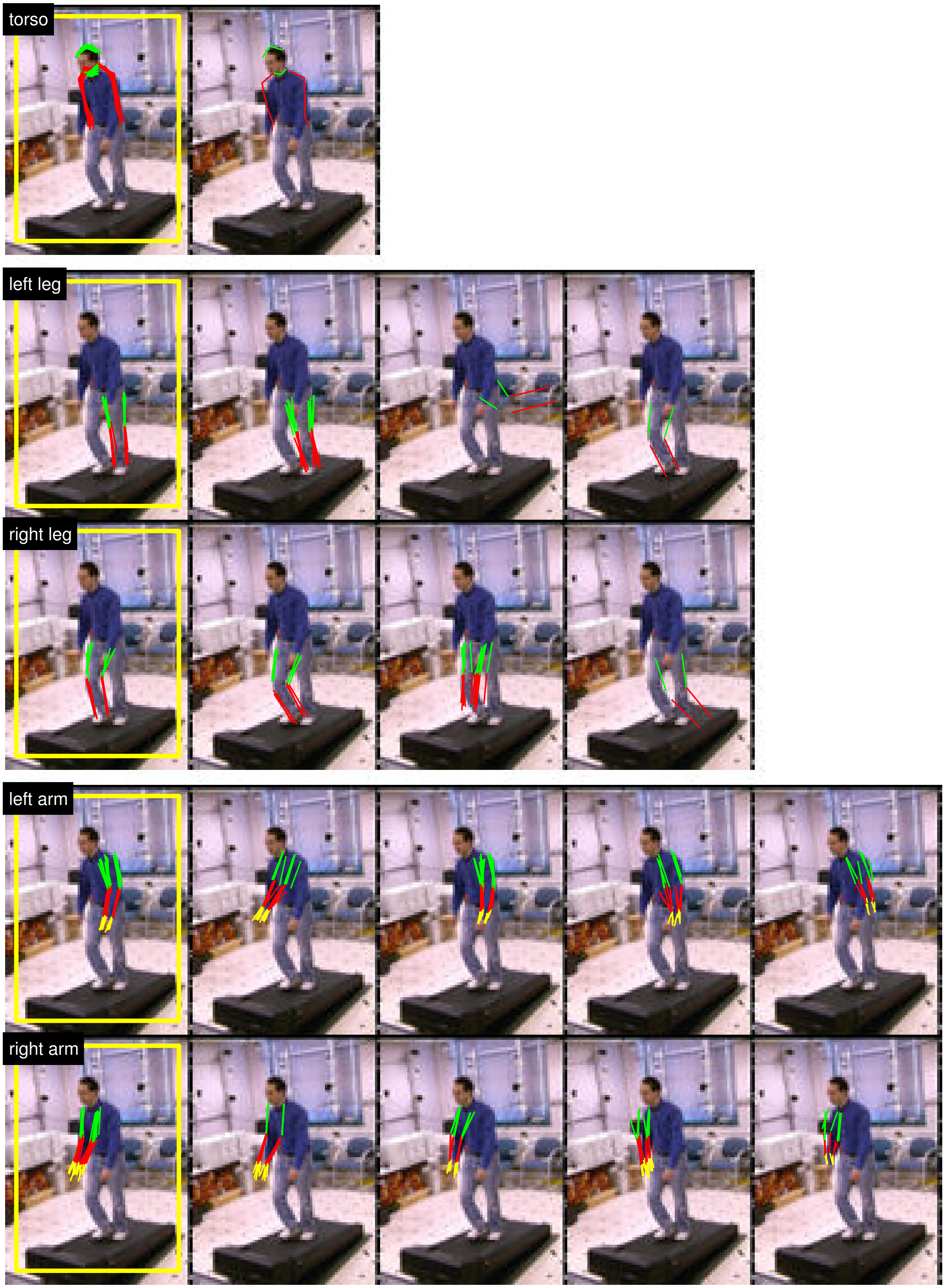
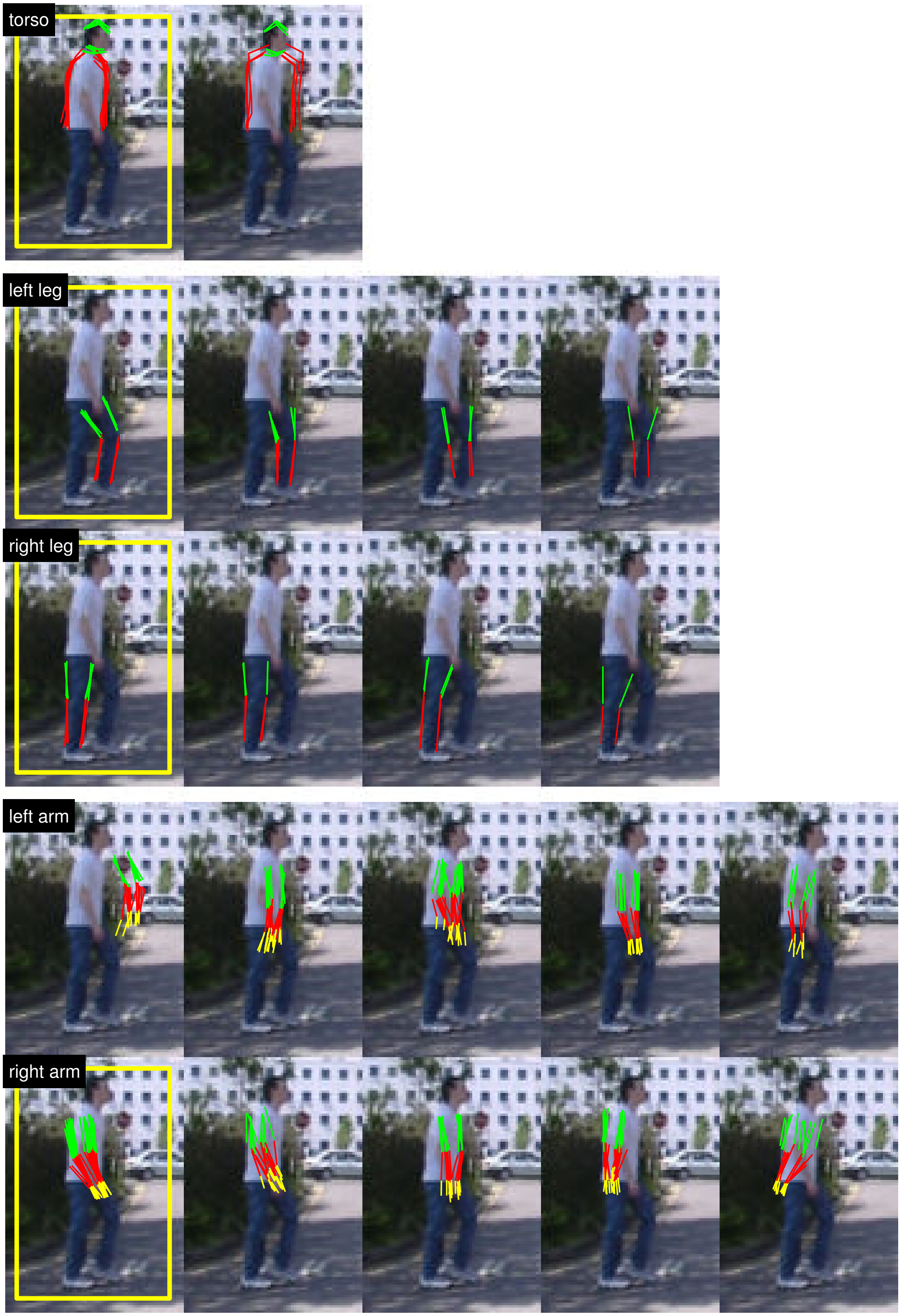


Figure 5.11: Complete candidate sets on 20 example images. Candidates are clustered into {2, 4, 4, 5, 5} hyper-modes for {torso, left leg, right leg, left arm, right arm} respectively. These hyper-modes are sorted based on a combination of likelihood and blob/ridge scores. “Preferred” hyper-modes are marked by yellow frames. The ideal automated scoring function would have the top-ranked hyper-mode (leftmost in each candidate set) coinciding with the preferred hyper-mode.

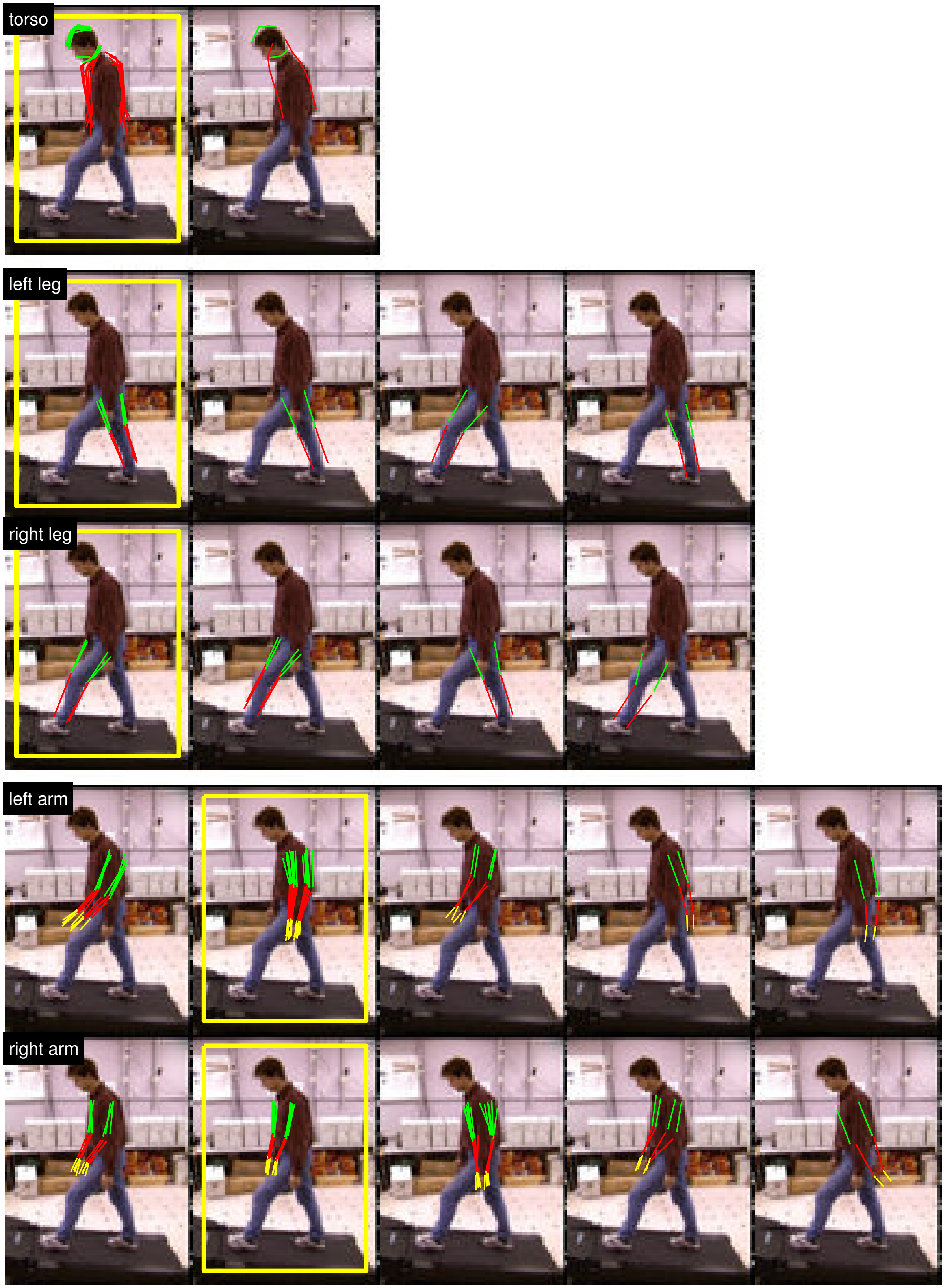
To be continued on pp. 88–100.



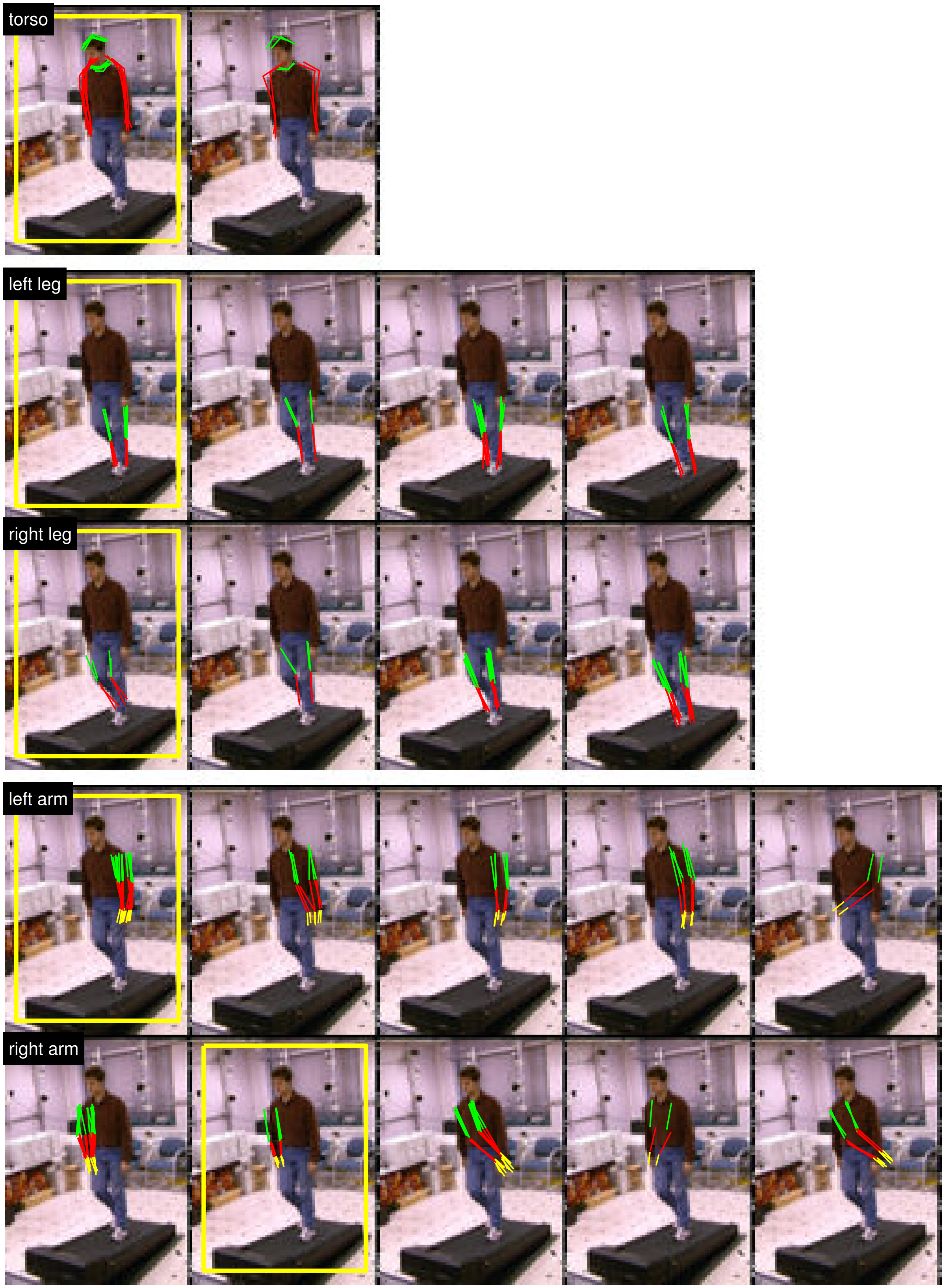
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (1/19)



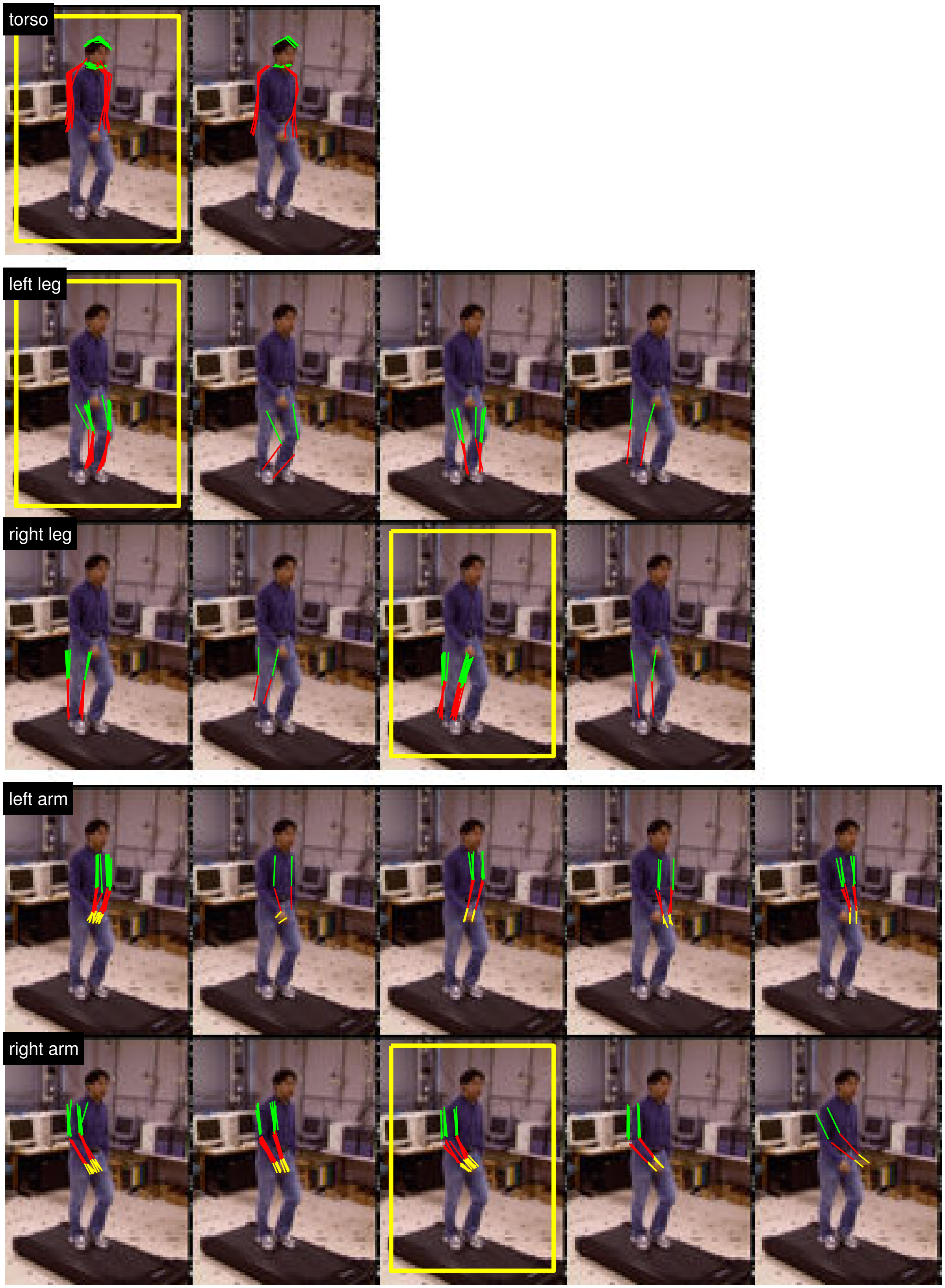
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (2/19)



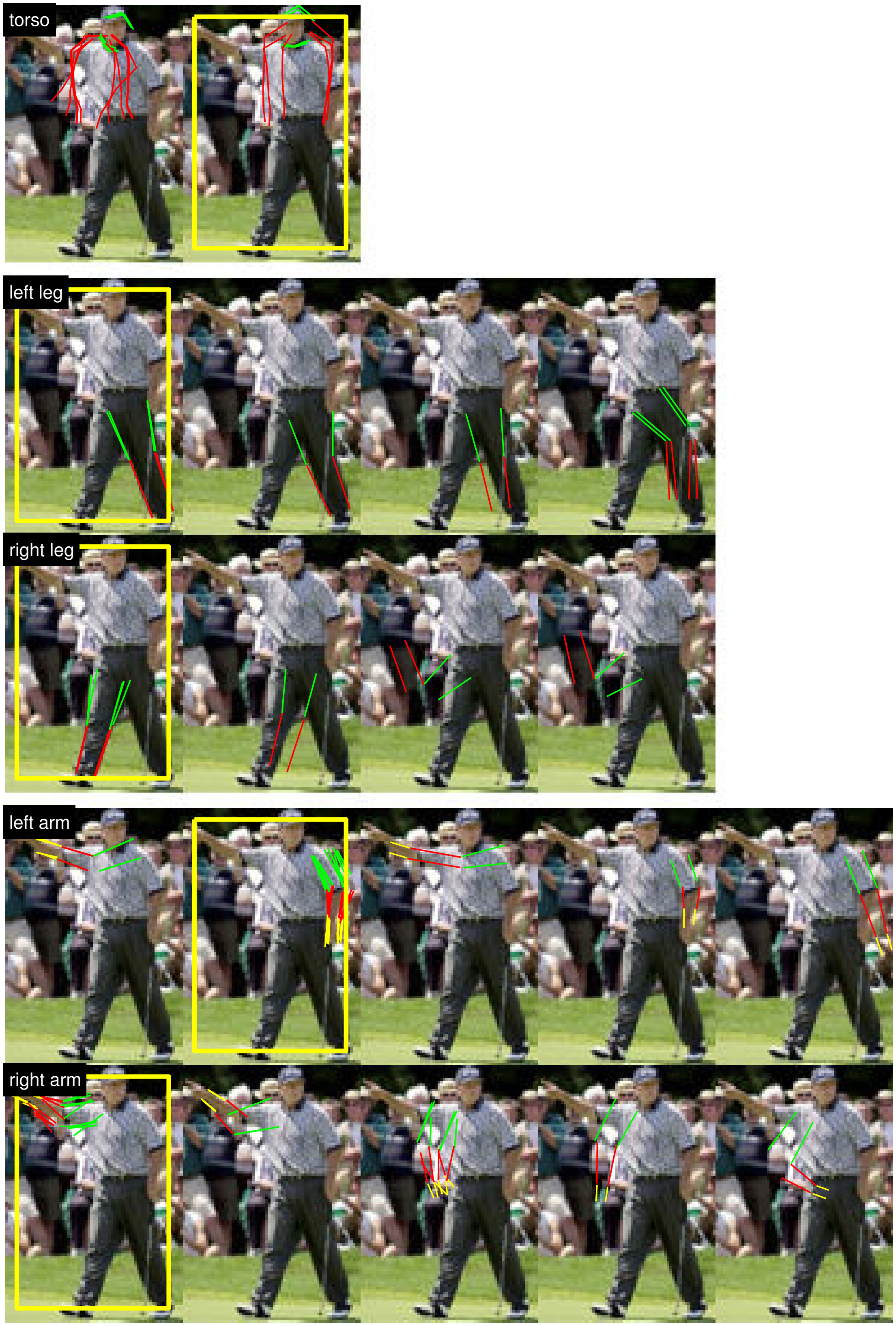
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (3/19)



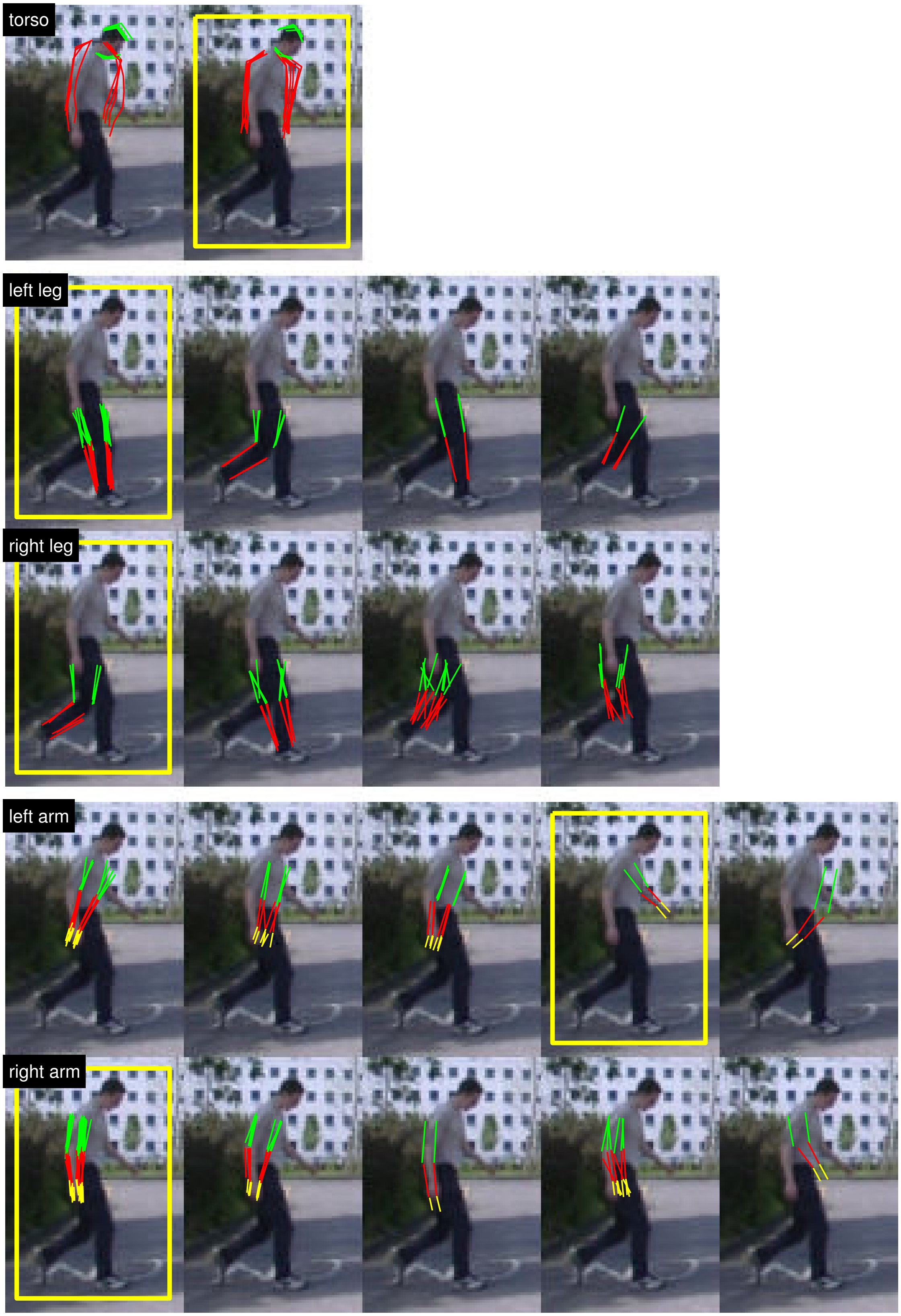
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (4/19)



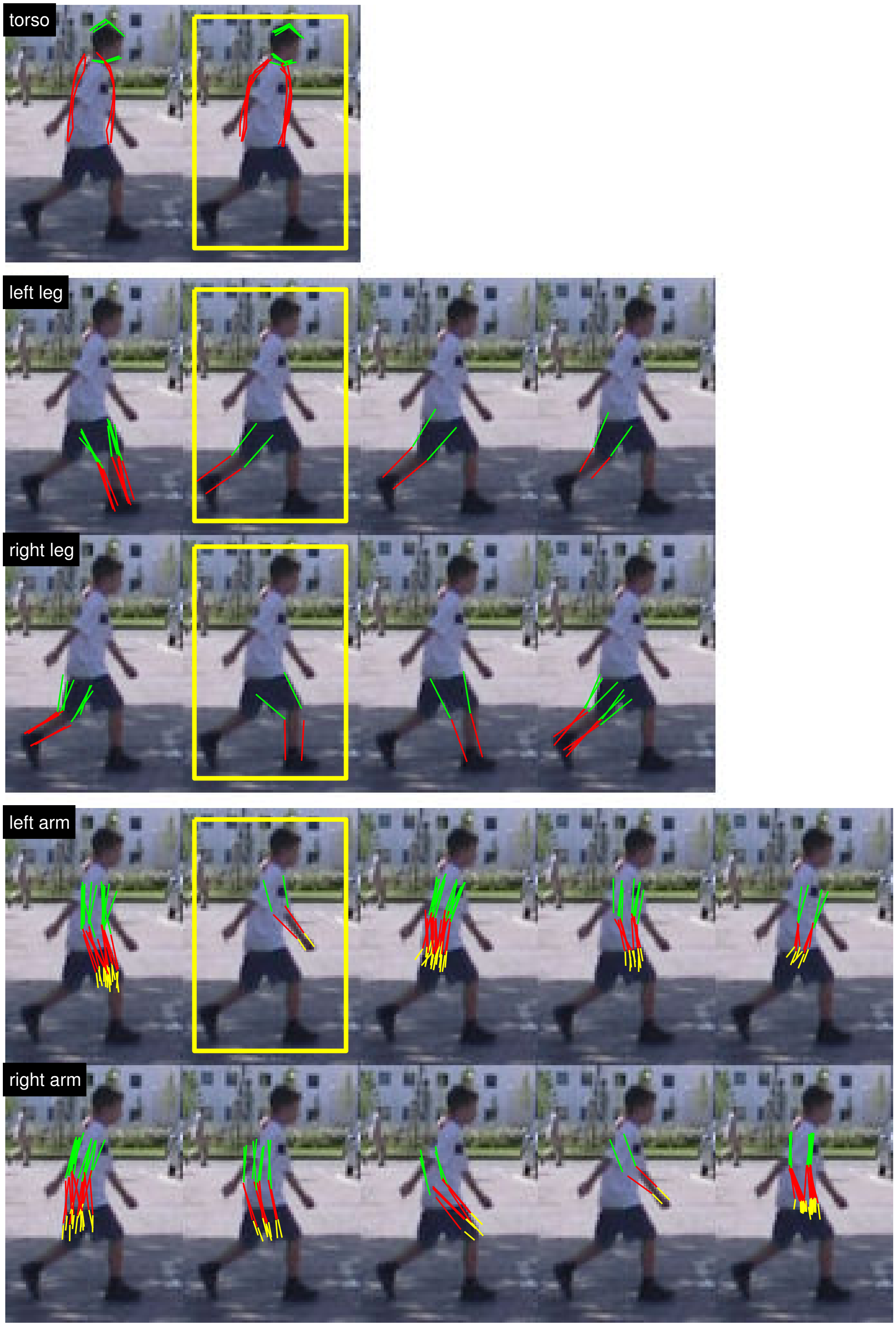
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (5/19)



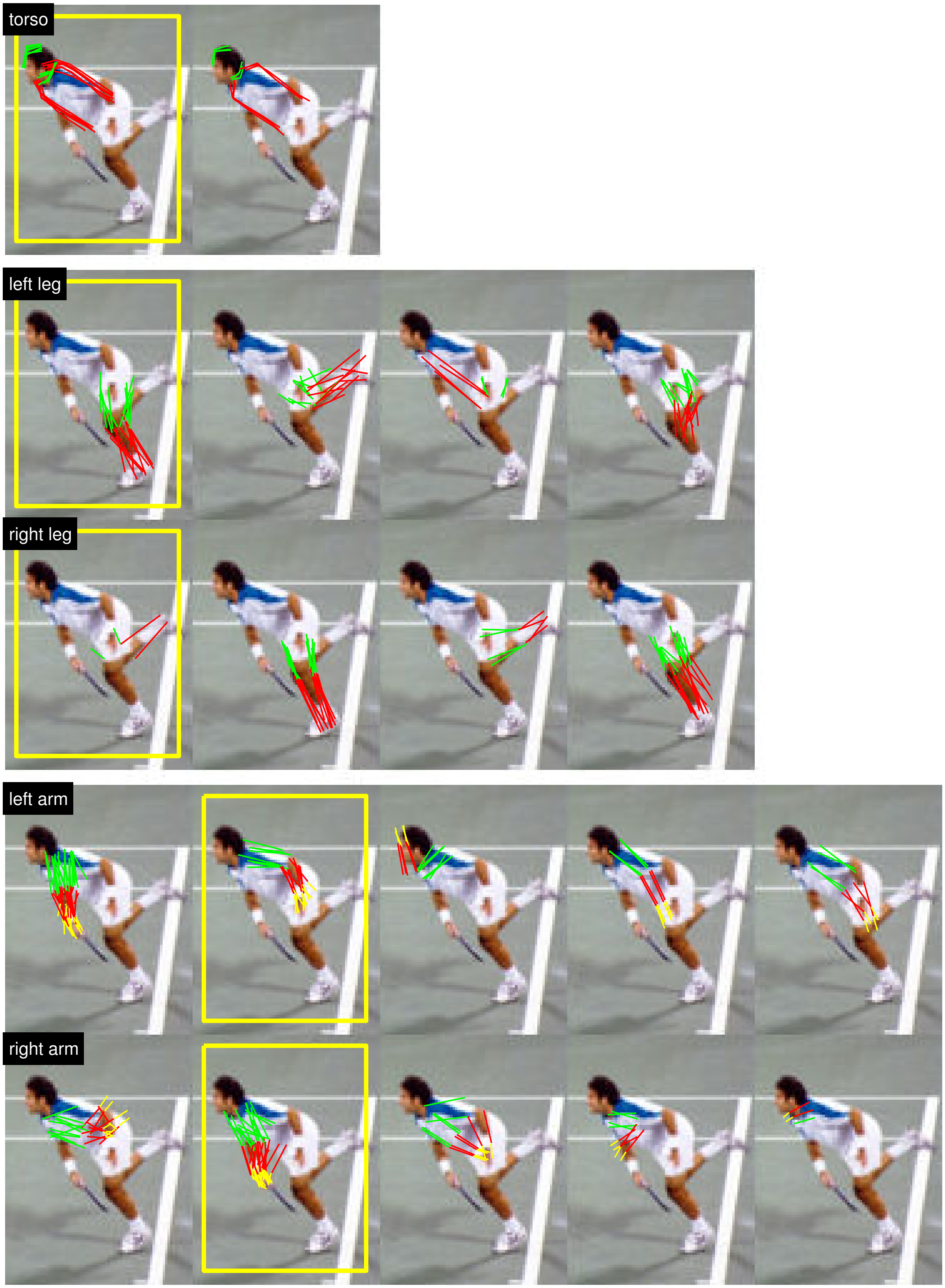
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (6/19)



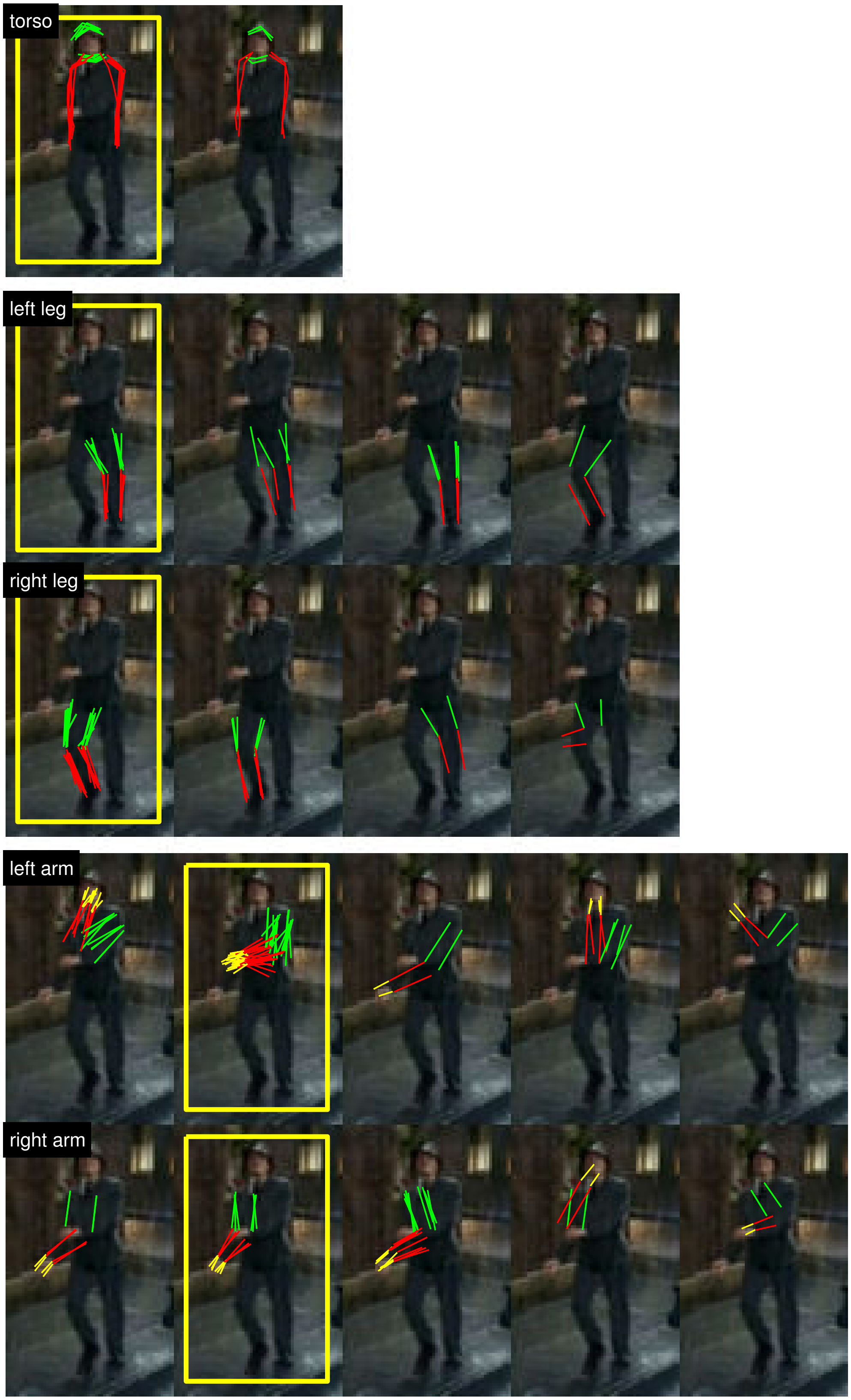
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (7/19)



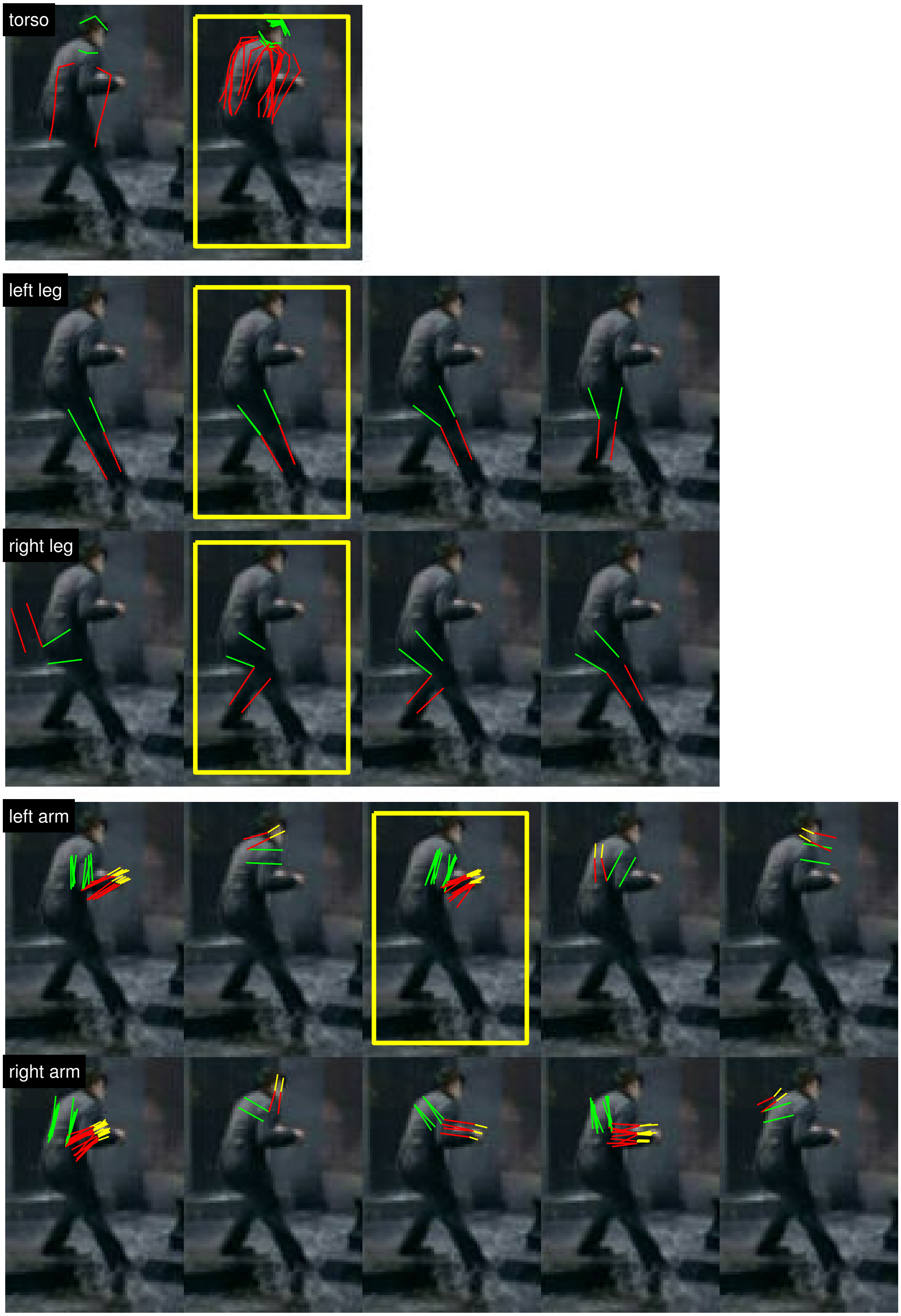
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (8/19)



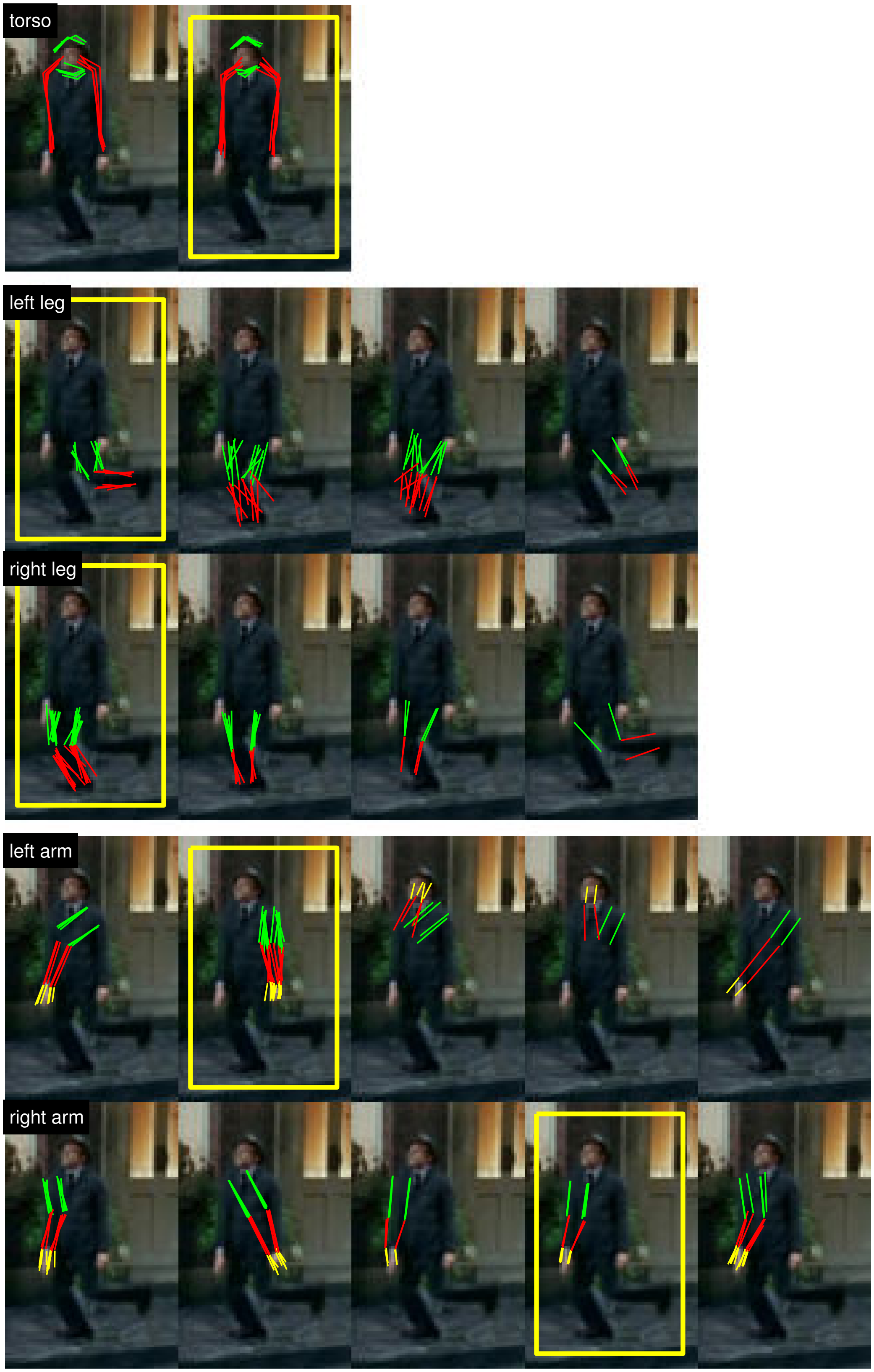
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (9/19)



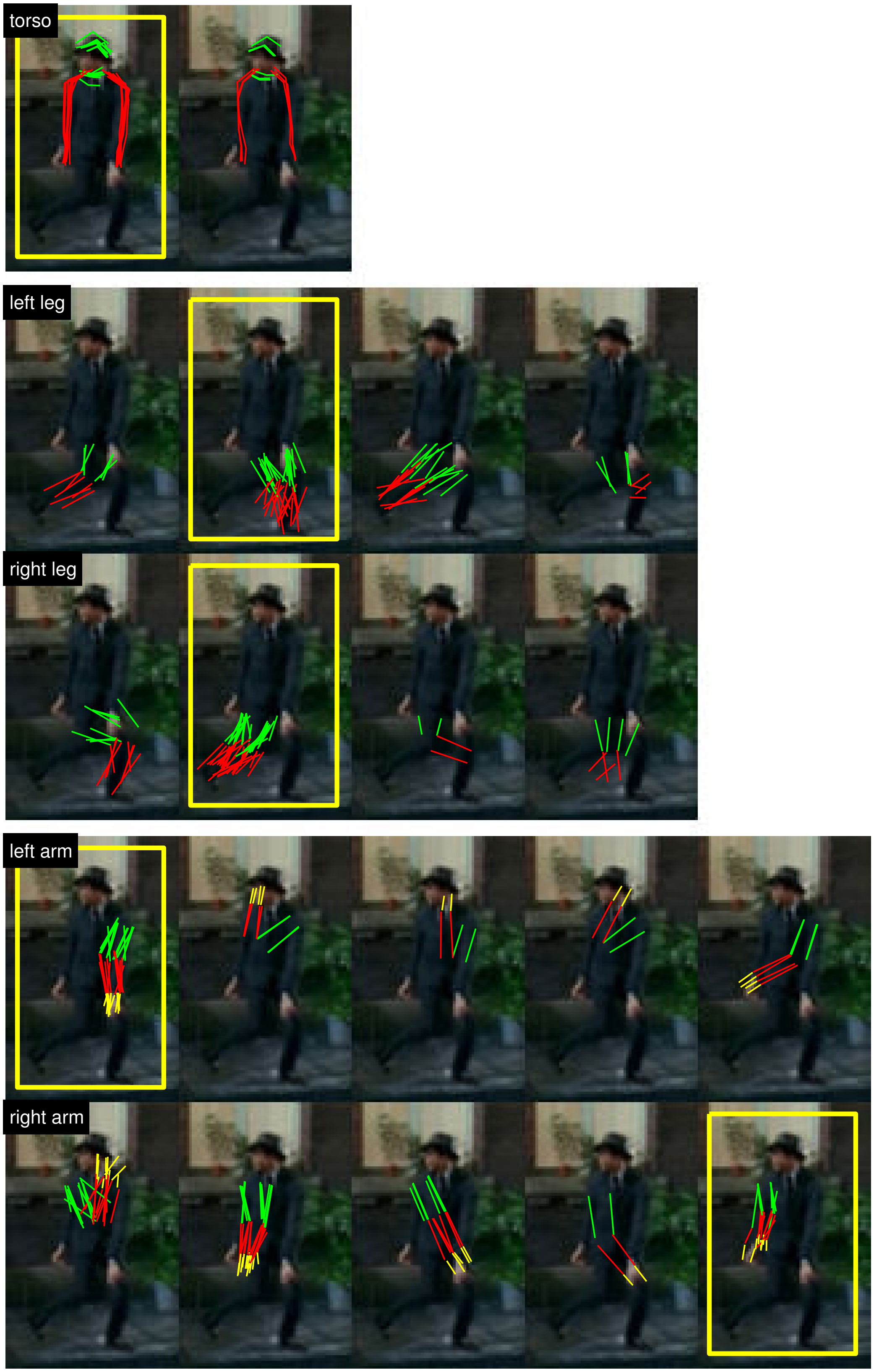
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (10/19)



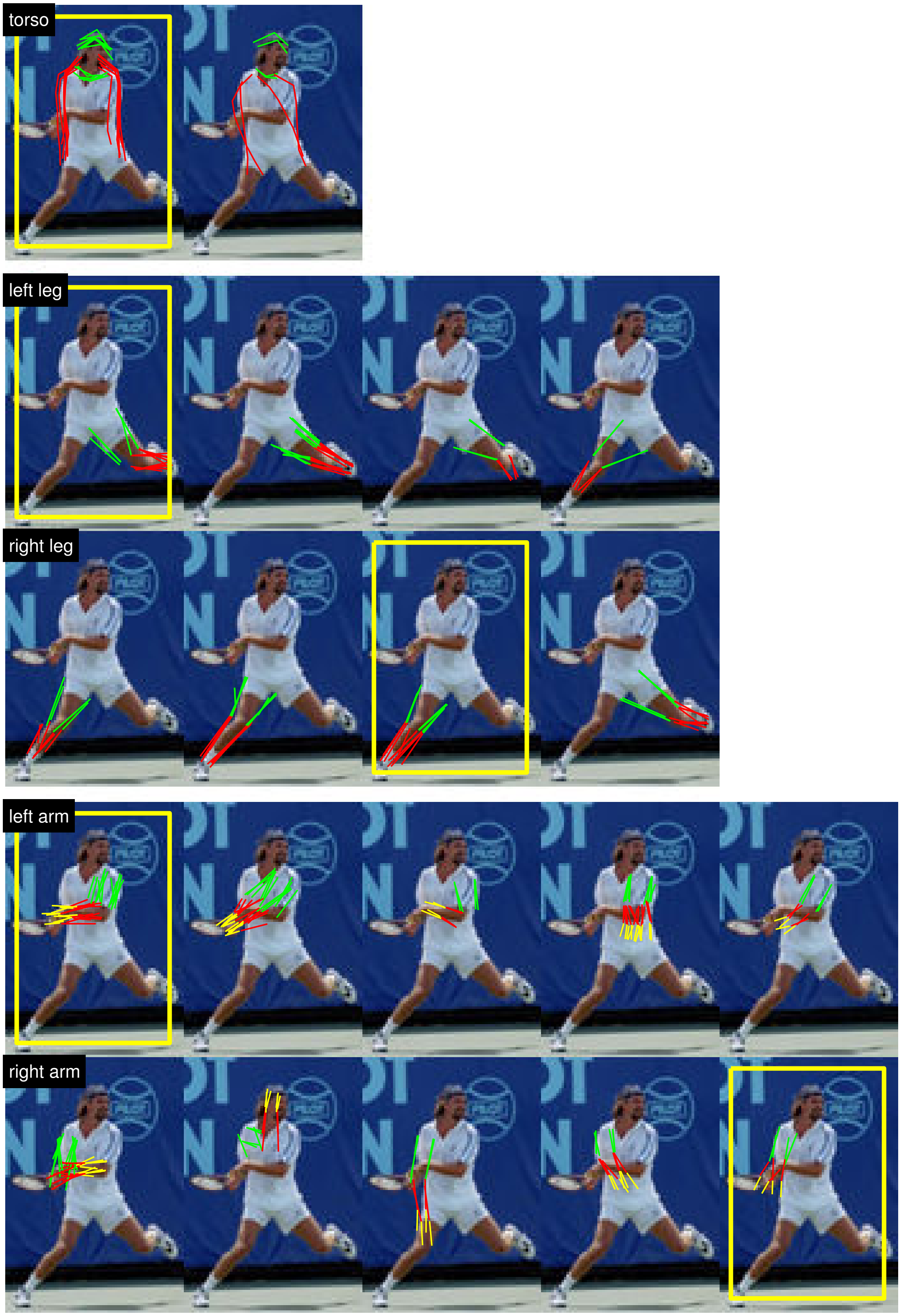
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (11/19)



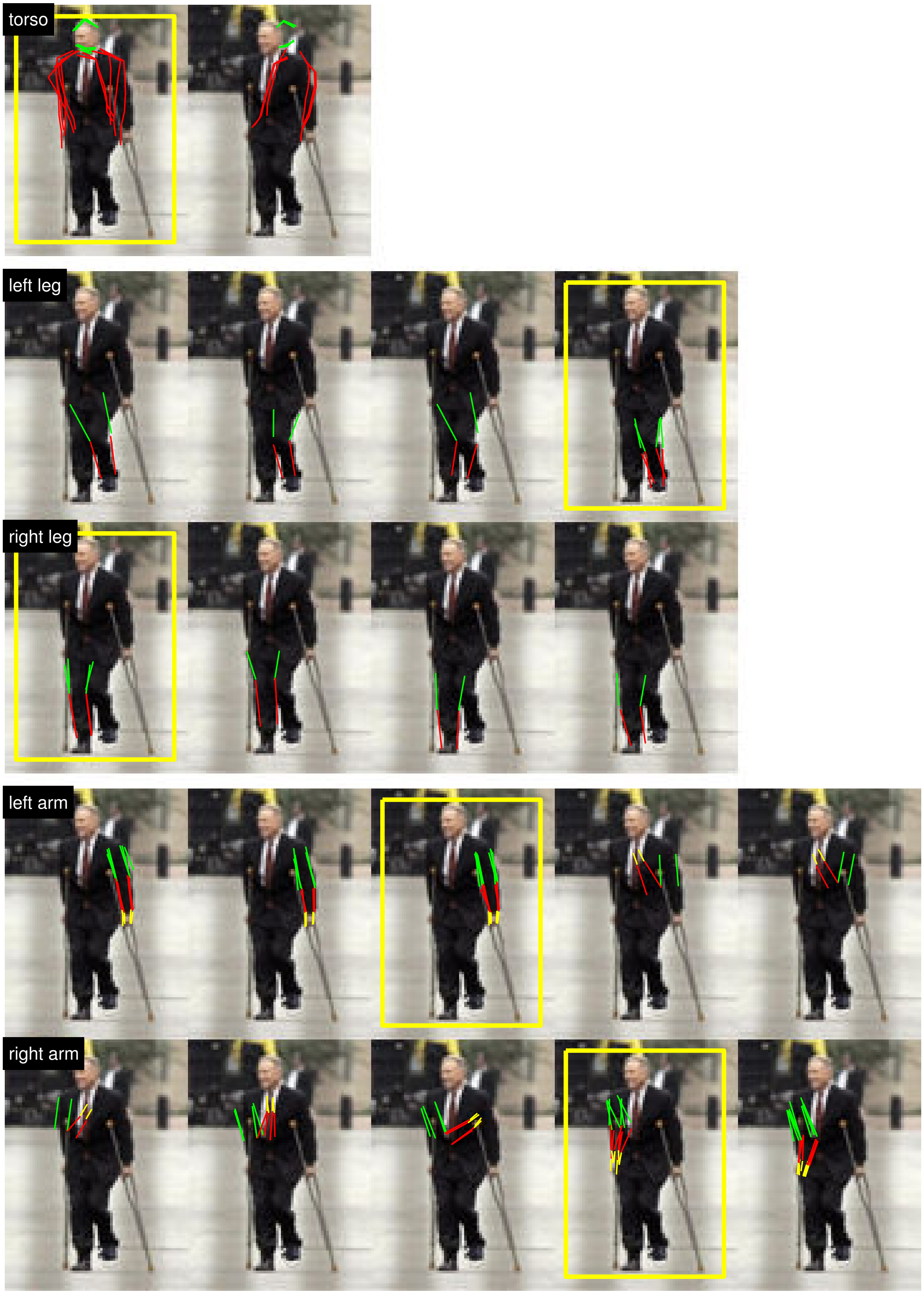
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (12/19)



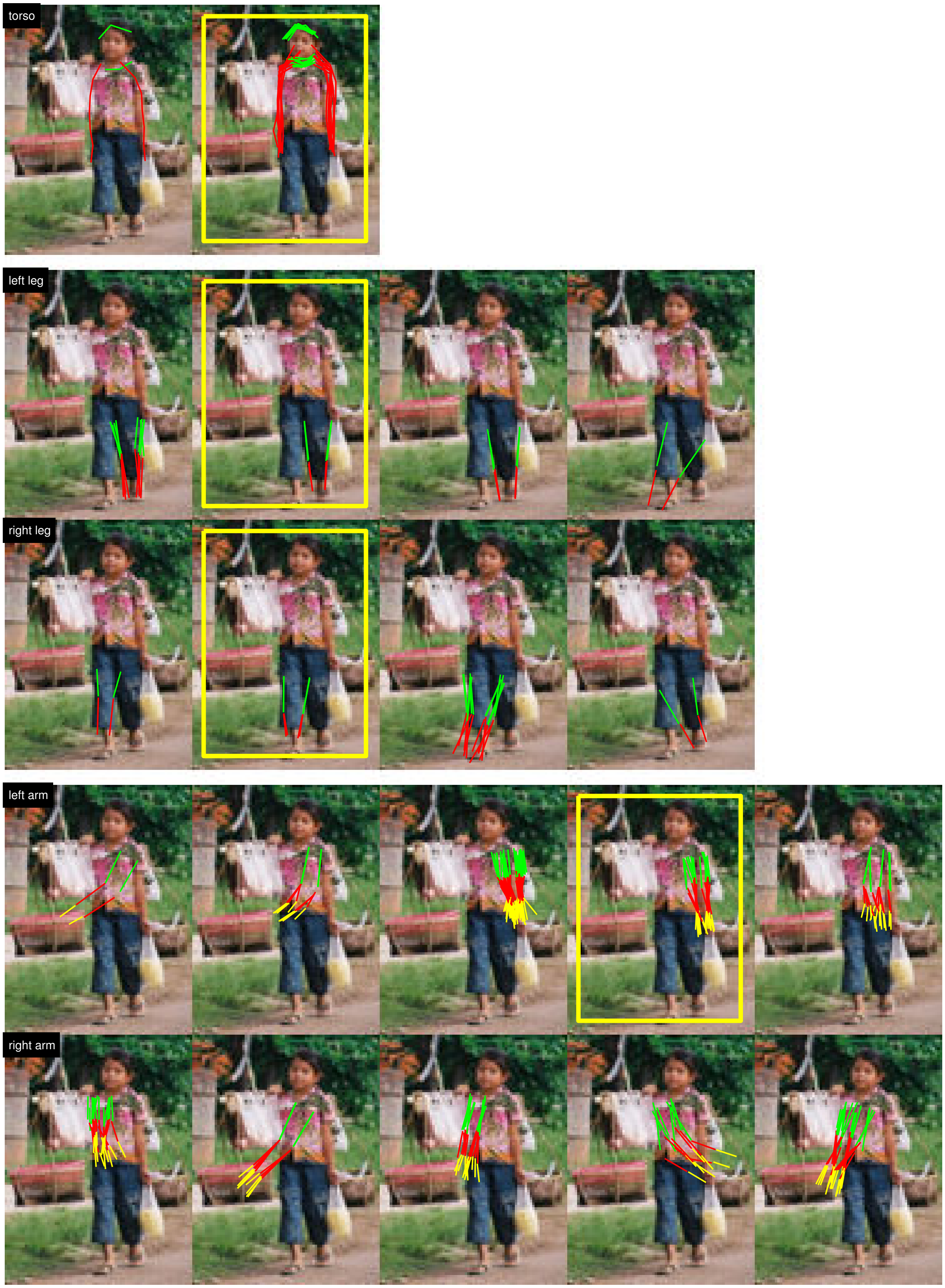
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (13/19)



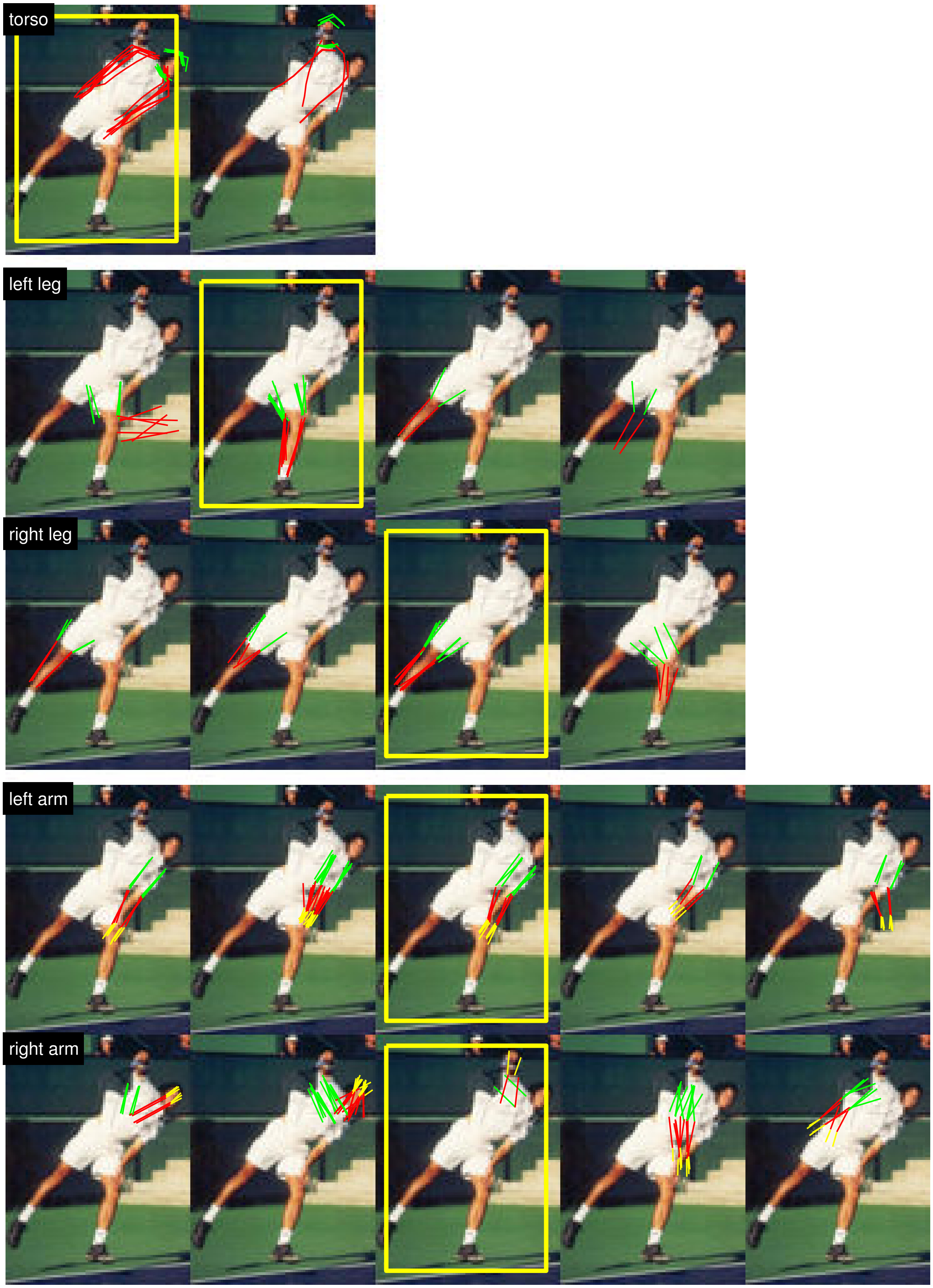
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (14/19)



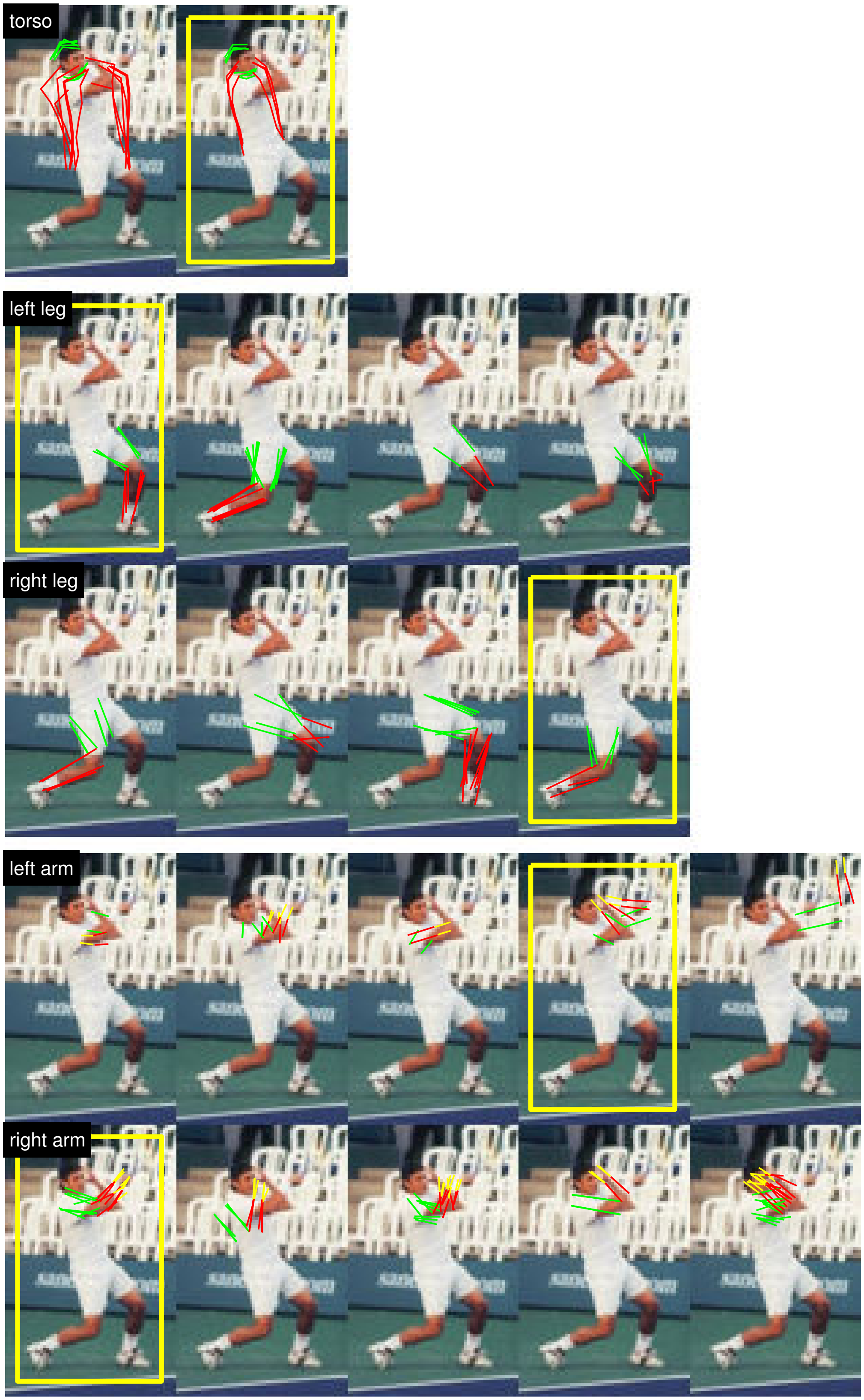
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (15/19)



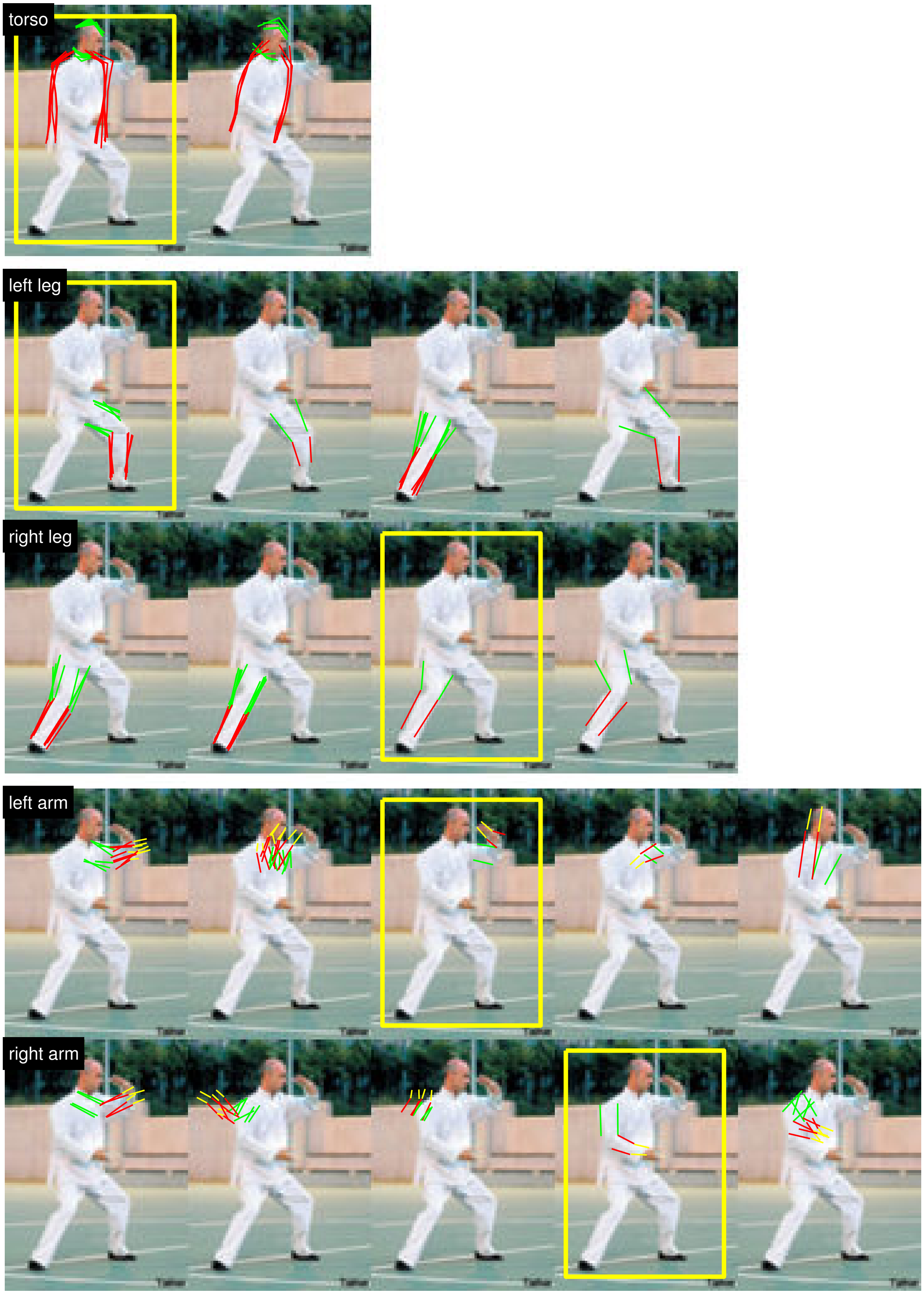
(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (16/19)



(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (17/19)



(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (18/19)



(Figure 5.11 on pp. 81 Continued) — Complete candidate sets — (19/19)

We consider fitting to have failed if the correct torso or leg position is missing from the output. On the 340 test images, our success rate is around 40%. Figures 5.12, 5.13, and 5.14 shows a sample of 113 successful cases (40 gait, 27 break dancing, and 46 web), where a "preferred" mode is manually selected from each candidate set. The average rank of these "preferred" modes is 4.7 out of 10 for torso, 5.2 out of 15 for legs, and 12.8 out of 30 for arms. These ranks vary with image/activity types. We have not done any special tuning of our algorithm for each type of activity. Also note that the existence of similar candidates and the flipping symmetry of left/right limbs should be considered in interpreting these ranks. Images with typical errors are hand picked and outlined with red frames. Most of these are missing or mislabeled arms. It is observed that fittings on the tennis players are least precise. This is because the athletes wear shorts and socks while our training subjects all wear long trousers. The assembled results have two implications:

1. They demonstrate the ability of our system to generate compact candidate sets that contain good candidates.

2. They are accurate enough to provide a good starting point for ground truth labeling.

The main reasons for fitting failure are the presence of clutter and unusual poses, and a few examples are shown in Figure 6.1. The reasons and possible solutions will be discussed in more detail in Section 6.2.

5.6 Summary

We have presented a 2D model-based algorithm for human body localization in still images. A hybrid search is conducted, combining stochastic and deterministic strategies in a coarse-to-fine manner, and facilitated by a hierarchical decomposition of the model. Experiments show that the number of particles can be drastically reduced while still achieving the same performance. Improvements in both speed and accuracy have been achieved compared to using only top-down SMC. The time to fit one image is about 5 minutes on a 2GHz PC, 5 times faster than using SMC alone.

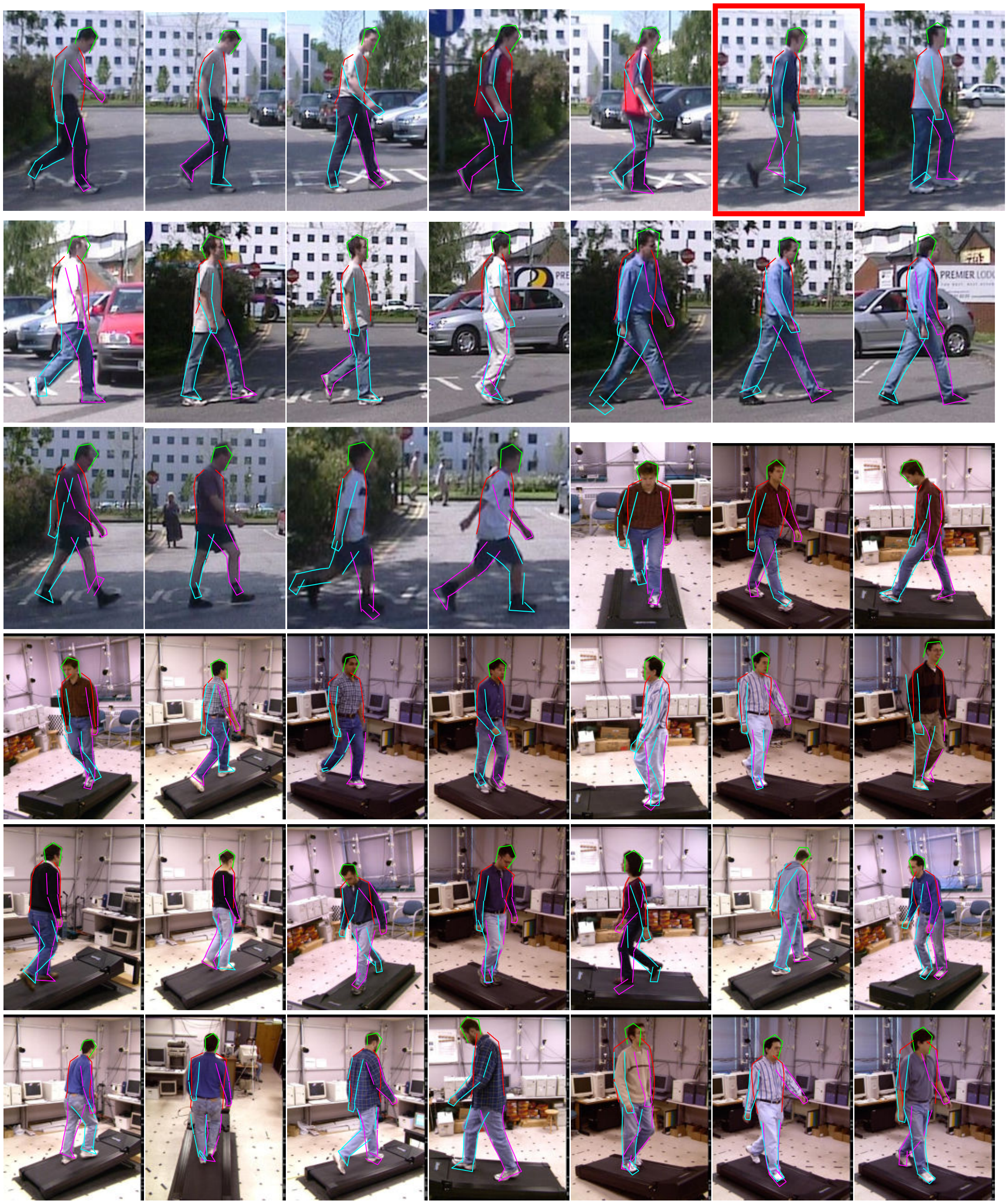


Figure 5.12: Example results on gait images assembled from “preferred” modes manually specified in each candidate set. Rows 1–3 are from the USH Gait Database, and rows 4–6 are from the CMU Mobo Database. No background subtraction is used. Images with typical errors are marked with red frames.

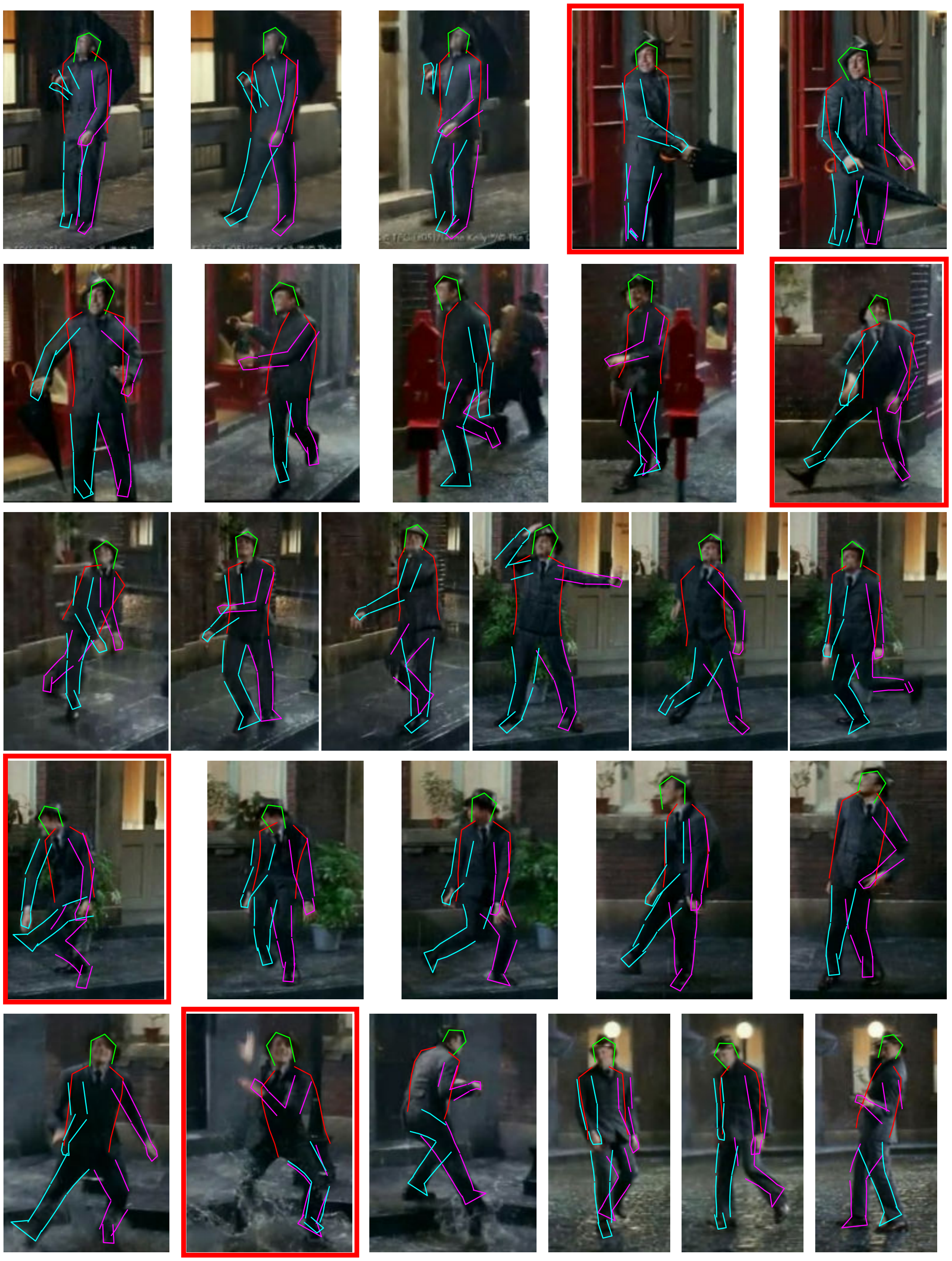


Figure 5.13: Example results on break dancing assembled from "preferred" modes manually specified in each candidate set. Shown are frames from a publicly available Volkswagen TV ad captured by a moving camera on a rainy night. Each frame is fit independently, and no motion cues are used. Images with typical errors are marked with red frames.

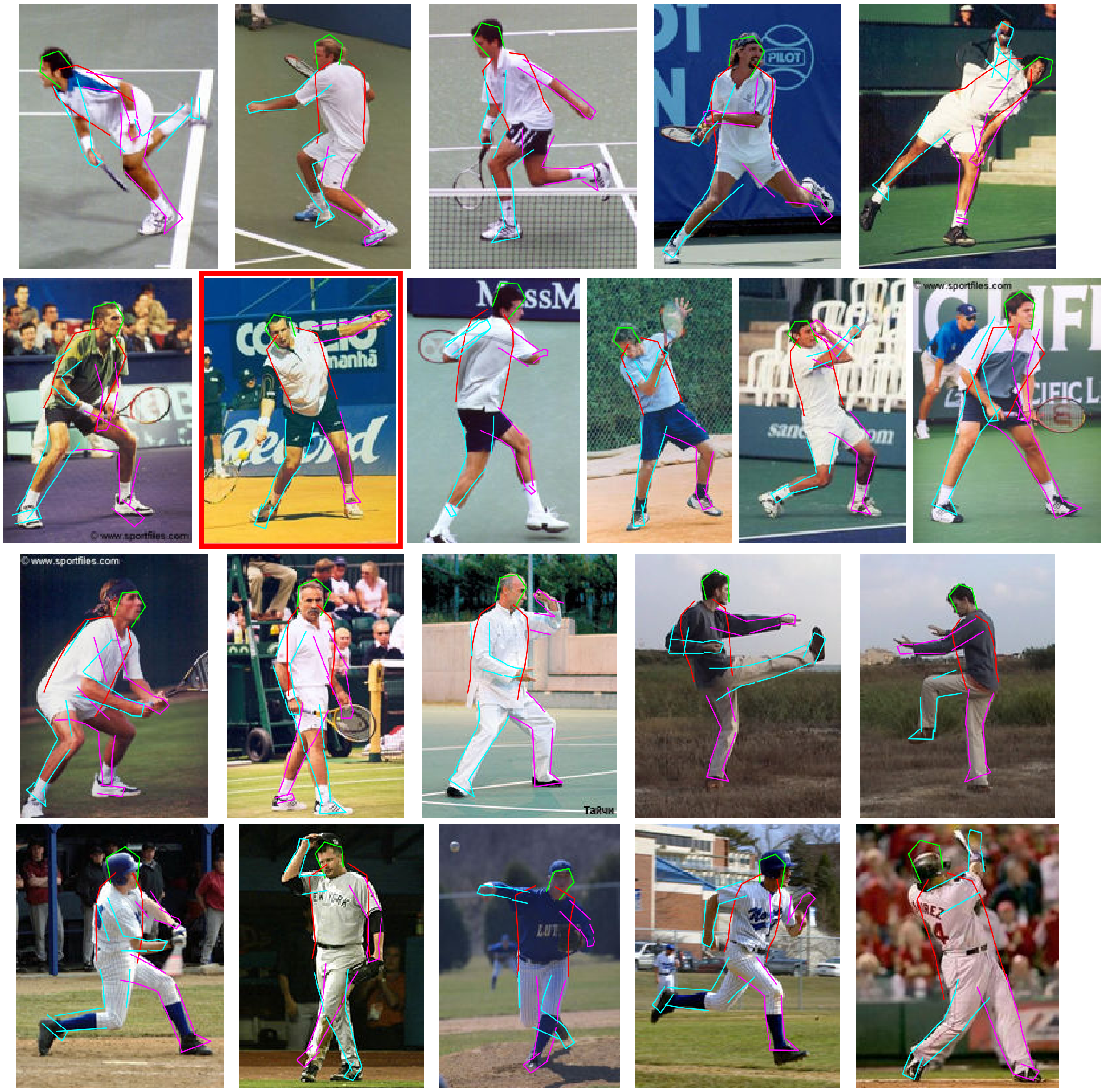
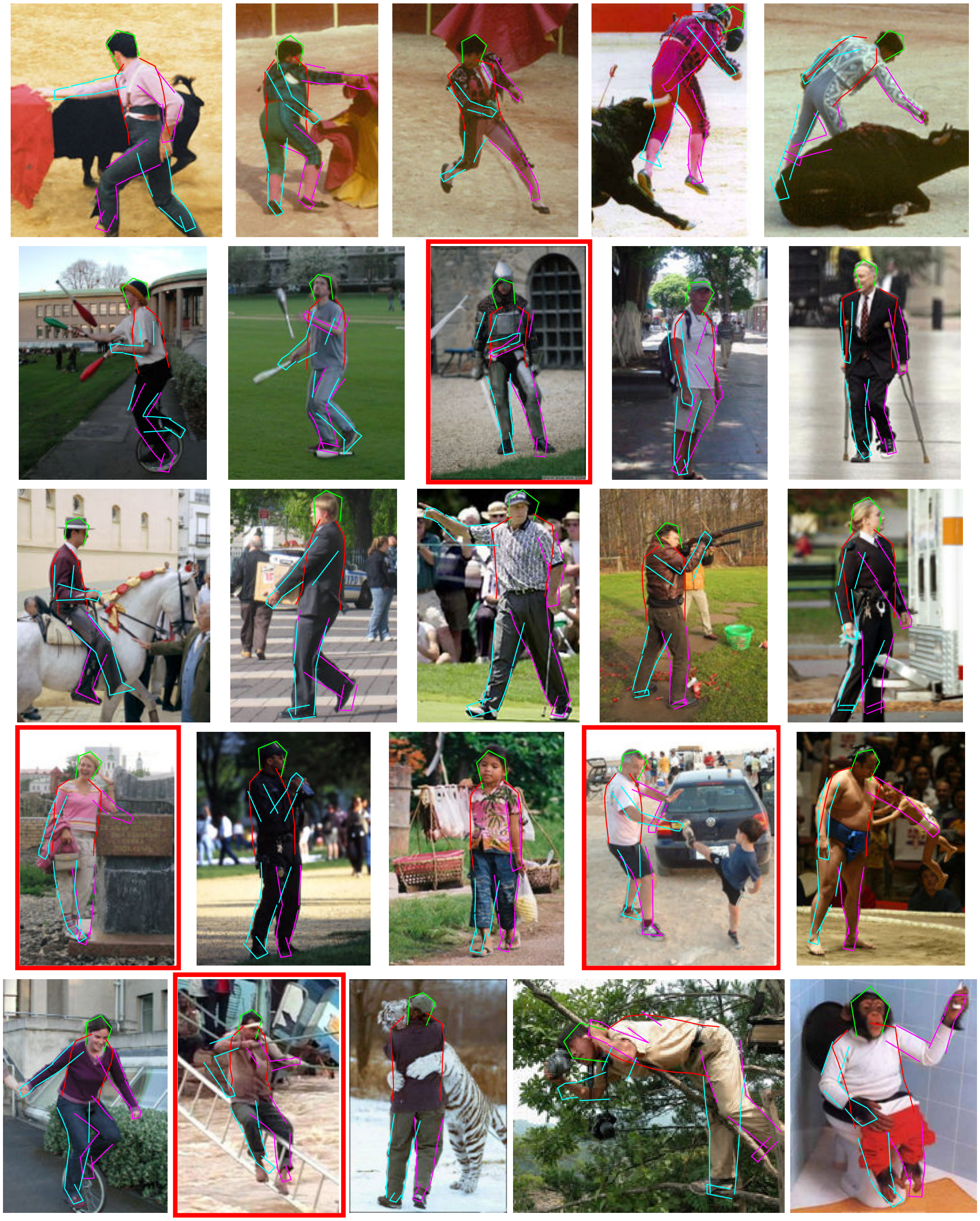


Figure 5.14: Example results on web images assembled from “preferred” modes manually specified in each candidate set. The poses vary from standing and walking to various sport activities, including tennis, baseball, bullfighting, taichi, and juggling. Images with typical errors are marked with red frames.

To be continued on pp. 105.



(Figure 5.14 on pp. 104 Continued) — Web Test — (1/1)

Chapter 6

Conclusion

6.1 Contribution

This thesis explores a model-based approach for body part localization in 2D still images. It has three main technical contributions:

- A landmark-based statistical representation of the nonrigid and articulated body shape is proposed and tested. A sequential structure is imposed on the landmark set such that inference can be done by simple yet effective stochastic sampling.

- A mixture of part-based 2D models is proposed and tested. The mixture model is fit by a parallel search that dynamically allocate resources to accommodate large number of mixture components.

- A 3-level hierarchy of body models is proposed and tested. The hierarchical model is fit by a hybrid search combining stochastic and deterministic strategies in a coarse-to-fine manner.

A body localization system is implemented that works in a generic setting: single image, arbitrary pose, and arbitrary viewpoint. The system is tested on a diverse and challenging dataset. The results obtained are favorable compared to the state of the art [1, 28, 36, 50, 59, 67]. The training data (ground truth labeling of around 7500 images from the CMU Mobo Dataset) has been made publicly available at `http://www.cs.cmu.edu/~zhangjy/thesis/`.

6.2 Discussion and Future Work

D1. Success Rate of the Search

The contour-based model used in this thesis is much more sophisticated than that in the previous work (*e.g.,* [28]), as is the inference algorithm. The major concern regarding such a system is whether or not it can handle high dimensional search spaces. We consider fitting to have failed if the correct torso or leg position is missing from the output. Our experimental results suggest that,

1. Given the side-walking viewpoint/pose constraint and the availability of background subtraction, our system almost always succeeds (Section 3.5).

2. When the viewpoint constraint is removed from Case 1, the success rate of fitting is still over 95% (Section 4.5).

3. When both viewpoint and pose constraints are removed from Case 1, and no background subtraction is used, the success rate drops to 40%–50% (Section 5.5).

Most failures in Case 3 result from DP failures in the presence of clutter (see Figure 6.1a for some examples). Possible reasons for such failures include:

1. The shape prior of the tree-structured model is weak, viewpoint independent and contains only one leg and one arm.

2. The likelihood function of the tree-structured model only uses edge and skin cues, which are prone to background distraction.

3. The output of backward sampling is not properly pruned.

Even when DP succeeds, the SMC inference might still fail (see Figure 6.1b for some examples). We note two major reasons:

1. There are nearby objects that are similar in appearance to the foreground human target (*e.g.,* the left two images in Fig. 6.1b). In this case, a more sophisticated background model seems to be necessary, with which the foreground body model has to compete in order to get a reasonable segmentation [86].

2. The human target is in an unusual pose (*e.g.,* the right two images in Fig. 6.1b) that is very different from any of the training data.

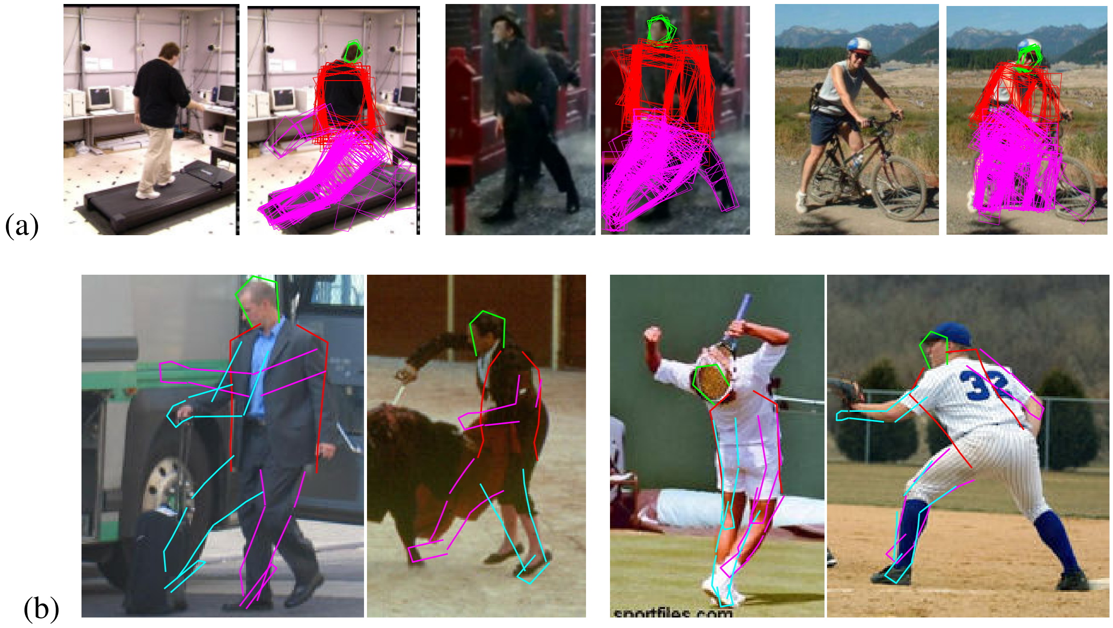


Figure 6.1: Examples of fitting failure. (a) DP fails due to background clutter. 3 pairs of examples with original image on the left and backward sampling results on the right. (b) SMC fails due to background clutter or unusual pose. Results overlaid are assembled from the best modes selected from each candidate set.

D2. Mode Selection and Human Perception

The output of our system is an entire population of samples resembling the posterior distribution over the configuration space. These samples are summarized by a few candidate modes or hyper-modes. Experimental results demonstrate the ability of our system to generate compact candidate sets that contain good candidates. However, ideally a single "optimal" solution should be found that best matches human perception.

Preserving multiple candidate solutions has been a common practice in state-of-the-art pose estimation systems (*e.g.,* [50, 59]). This common practice reflects the difficulty of designing an objective criterion that perfectly matches human perception, particularly when given a generic problem setting and challenging data as is the case in our work.

We have made a preliminary attempt to design scoring functions to automatically select the "preferred" modes depicted in Figures 5.12 through 5.14. However, there is still a gap between our result (see Figure 5.11 for some examples) and the ideal one, where the "preferred" mode would be ranked 1. This gap is especially obvious for the category of arms. To alleviate this problem, strong likelihood models, visual perception rules, and background image understanding might help. It is also possible to prune false alarms by exploiting extra information, *e.g.,* stereo depth maps or temporal consistency. Building a fully automatic mode selection scheme in the single-image scenario remains an interesting open problem.

D3. Temporal Tracking

Statistical pose estimation methods that work on a single image can be readily extended to multiple images or video sequences (*e.g.,* [51, 99]). Here we briefly discuss the possibility of a naive extension of our system to *on-line* body tracking. Consider the following Bayesian formulation. Let $\{\mathcal{I}^t, \mathcal{I}^{t-1}\}$ be two adjacent frames, and $\{\Omega^t, \Omega^{t-1}\}$ be the configurations at time t and $(t-1)$. With minor assumptions, the joint posterior can be written as

$$p(\Omega^t, \Omega^{t-1}|\mathcal{I}^t, \mathcal{I}^{t-1}) \propto p(\mathcal{I}^t|\Omega^t, \Omega^{t-1}, \mathcal{I}^{t-1})\, p(\Omega^t|\Omega^{t-1}) p(\Omega^{t-1}|\mathcal{I}^{t-1}). \tag{6.1}$$

The right side of Equation (6.1) consists of three terms. We show that they can be decomposed in a similar way to the case of single frame, such that the SMC inference framework (Section 3.4) can still be applied.

The term $p(\Omega^{t-1}|\mathcal{I}^{t-1})$ is the posterior computed before time t, and is represented by a set of samples $\left\{\mathbf{v}_{0:K}^{t-1,(i)}\right\}_{i=1}^{N}$. Assuming a non-adaptive appearance model, *i.e.,*

$$p(\mathcal{I}^t|\Omega^t, \Omega^{t-1}, \mathcal{I}^{t-1}) = p(\mathcal{I}^t|\Omega^t), \tag{6.2}$$

the likelihood term can be decomposed in exactly the same way as in Equation (3.11). We further assume that the prediction term $p(\Omega^t|\Omega^{t-1})$ can be factored into two components,

$$p(\Omega^t|\Omega^{t-1}) \approx p_1(\Omega^t) \cdot p_2(\|\Omega^t - \hat{\Omega}^t\|^2), \tag{6.3}$$

where $\hat{\Omega}^t$ is the prediction from previous frames using any dynamic model (*e.g.,* random walk or constant velocity), p_1 is the single-frame shape prior as defined in Equation (4.8), and p_2 is a decomposable term (*e.g.,* white Gaussian) that prefers smooth motion. Taking all into consideration, Equation (6.1) can be formulated into a similar sequential structure as Equation (3.19). The only difference is the introduction of local smoothness terms $p(\|\mathbf{v}_k^t - \hat{\mathbf{v}}_k^t\|^2)$, which do not affect the use of spatial SMC inference. We start from a set of particles from the previous frame, and interpret them as intermediate partial fits of a search in an augmented configuration space $\underline{\Omega} = \{\Omega^t, \Omega^{t-1}\}$.

D4. 3D Model

The current system does not work well with serious torso/limb foreshortening, which is a common weakness to all 2D based methods. Introducing a 3D model might help in this case, with two additional potential advantages:

1. To enable direct 3D pose recovery;

2. To prune 2D deformations that are physically impossible [93].

One way to combine 2D and 3D models is to use exemplars of 3D projection to model 2D deformation (*e.g.,* [23, 61]), by which 3D information is inherently encoded into the 2D model.

D5. Other Articulated Shapes

Besides the human body, people are sometimes interested in other nonrigid and articulated objects such as pets and domestic animals (see Figure 6.2 for some examples). These animals usually possess textured appearances. However, there are two difficulties in applying our contour models to them:

1. Their torsos are horizontally oriented, resulting in considerable nonlinear part deformation when viewpoint changes [20, 91];

2. We have to repeat the tedious acquisition process of the training data. It would be desirable to be able to learn the model automatically from images or videos [19, 46].

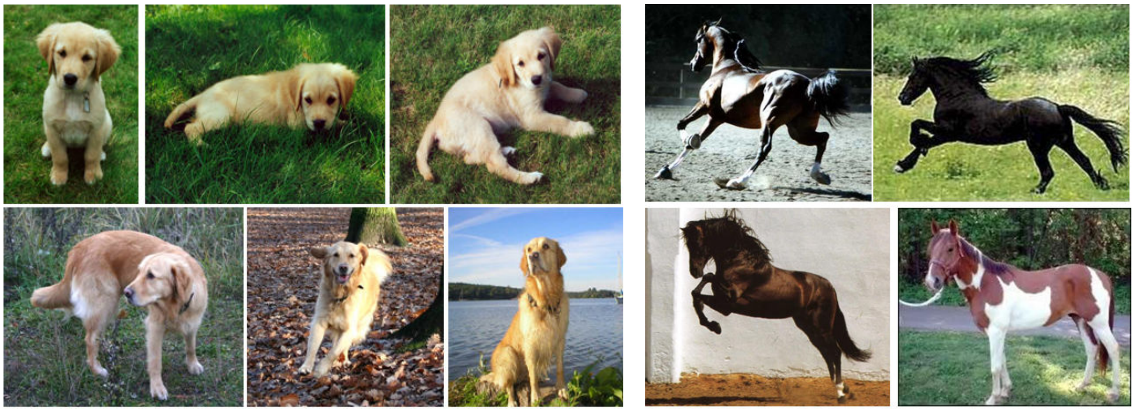


Figure 6.2: Two examples of articulated animals. *Left:* Photos from the album of a dog. *Right:* Images from the Weizmann Horse Database [9].

D6. Computational Considerations

Our current system takes minutes to process one image. The strategy of its main inference algorithm (SMC) is to sample the posterior by tens of thousands of particles. In this sense, our approach is brute-force, and thus simple. Computational efficiency may be improved, for example, by integrating deterministic optimization techniques, or by taking advantage of its parallelizability. Obviously, such a brute-force method can always benefit considerably from the Moore's Law prediction of rapid improvement in future computer performance.

D7. Towards Real-world Applications

In this thesis, we implemented a body localization system that shows promising results on a challenging dataset. Such a system may be applied to real-world applications by: 1) imposing constraints from the application context, 2) incorporating new image/motion cues, and 3) combining complementary detection/localization methods.

To conclude, we would like to step back and reflect on the general problem of people image analysis. Figure 6.3 shows the architecture of the People Image Analysis (PIA) Consortium at CMU, which develops and distributes technologies that process images and videos to detect, track, and understand people's face, body, and activities. The goal is to develop a comprehensive set of imaging and processing tools, systems, or subsystems that work in the real-world environment. As can be observed from Figure 6.3, only a limited portion of the architecture has been covered (to various degrees) by this thesis, and many interesting problems remain to be explored.

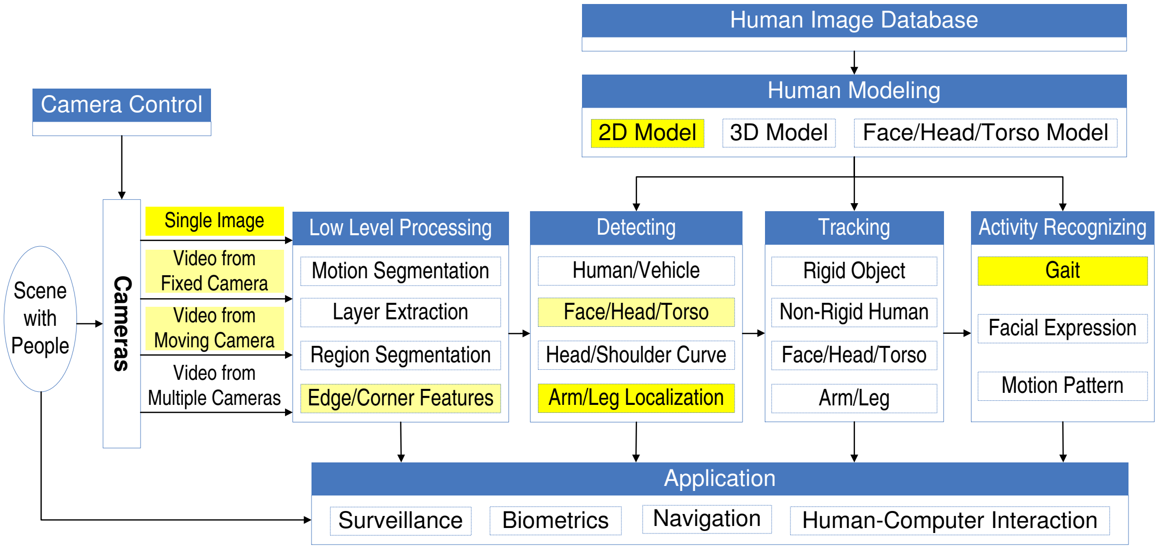


Figure 6.3: The architecture of the People Image Analysis (PIA) Consortium at CMU (`http://www.consortium.ri.cmu.edu`). Modules with yellow backgrounds have been covered to various degrees by this thesis.

Bibliography

[1] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *Proc. CVPR*, volume 2, pages 882–888, 2004.

[2] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *Proc. ECCV*, volume 3, pages 54–65, 2004.

[3] J. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[4] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. BoostMap: A method for efficient approximate similarity rankings. In *Proc. CVPR*, volume 2, pages 268–275, 2004.

[5] A. Barr. Global and local deformations of solid primitives. *Computer Graphics*, 18:21–30, 1984.

[6] J. Berger. *Statistical Decision Theory and Bayesian analysis*. Springer-Verlag, 1985.

[7] M. Bern and D. Eppstein. Mesh generation and optimal triangulation. In D. Du and F. Hwang, editors, *Computing in Euclidean Geometry*, number 4 in Lecture Notes Series on Computing, pages 47–123. World Scientific, 2nd edition, 1995.

[8] F. Bookstein. Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science*, 1(2):181–242, 1986.

[9] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proc. ECCV*, pages 109–122, 2002.

[10] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. CVPR*, pages 8–15, 1998.

[11] C. Cedras and M. Shah. A survey of motion analysis from Moving Light Displays. In *Proc. CVPR*, pages 214–221, 1994.

[12] C. Cedras and M. Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, 1995.

[13] T. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. CVPR*, volume 1, pages 239–245, 1999.

[14] T. Chan and W. Zhu. Level set based shape prior segmentation. In *Proc. CVPR*, pages 1164–1170, 2005.

[15] H. Chen, Z. Xu, and S. Zhu. Context sensitive grammar and composite templates for cloth modeling. In *Proc. CVPR*, 2006.

[16] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time: Part II: Applications to human modeling and markerless motion tracking. *Int. J. Comp. Vision*, 63(3):225–245, 2005.

[17] T. Cootes, G. Edwards, and C. Taylor. Active appearance model. In *Proc. ECCV*, volume 2, pages 484–498, 1998.

[18] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models: Their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[19] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proc. ECCV*, 2006.

[20] D. Cremers, T. Kohlberger, and C. Schnörr. Nonlinear shape statistics in Mumford-Shah based segmentation. In *Proc. ECCV*, pages 93–108, 2002.

[21] H. Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Implementation available at `http://www.isi.edu/~hdaume/megam/`, August 2004.

[22] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proc. CVPR*, volume 2, pages 126–133, 2000.

[23] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose recognition using spatio-temporal templates. In *ICCV Workshop on Modeling People and Human Interaction, Beijing, China*, October 2005.

[24] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[25] I. Dryden and K. Mardia. *Statistical Shape Analysis*. John Wiley and Sons, 1998.

[26] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. ICCV*, pages 726–733, 2003.

[27] P. Felzenszwalb. Representation and detection of deformable shapes. In *Proc. CVPR*, volume 1, pages 102–108, 2003.

[28] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comp. Vision*, 61(1):55–79, 2005.

[29] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[30] D. Gavrila and L. Davis. 3D model-based tracking of humans in action: A multi-view approach. In *Proc. CVPR*, pages 73–80, 1996.

[31] W. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[32] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3D structure with a statistical image-based shape model. In *Proc. ICCV*, volume 2, pages 641–648, 2003.

[33] R. Gross and J. Shi. The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.

[34] J. Hammersley and K. Morton. Poor man's Monte Carlo. *J. Roy. Stat. Soc. B*, 16:23–38, 1954.

[35] D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[36] G. Hua, M. Yang, , and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *Proc. CVPR*, pages 747–754, 2005.

[37] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. In *Proc. ICCV*, volume 1, pages 690–695, 2001.

[38] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *Int. J. Comp. Vision*, 43(1):45–68, 2001.

[39] M. Isard. Pampas: Real-valued graphical models for computer vision. In *Proc. CVPR*, volume 1, pages 613–620, 2003.

[40] A. Jain, Y. Zhong, and S. Lakshmanan. Object matching using deformable templates. *IEEE Trans. PAMI*, 18(3):267–278, 1996.

[41] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1973.

[42] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *Proc. Int. Conf. on Autom. Face and Gesture Recog.*, pages 38–44, 1996.

[43] D. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.*, 16:81–121, 1984.

[44] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comp. Graph. Stat.*, 5(1):1–25, 1996.

[45] K. Kremer and K. Binder. Monte Carlo simulation of lattice models for macromolecules. *Computer Physics Reports*, 7:259–310, 1988.

[46] M. Kumar, P. Torr, and A. Zisserman. Learning layered motion segmentation of video. In *Proc. ICCV*, pages 33–40, 2005.

[47] M. Kumar, P. Torr, and A. Zisserman. OBJ CUT. In *Proc. CVPR*, pages 18–25, 2005.

[48] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *Proc. CVPR*, volume 1, pages 722–729, 2004.

[49] M. Lee and I. Cohen. Human body tracking with auxiliary measurements. In *Proc. Workshop on Anal. and Model. of Faces and Gestures*, pages 112–119, 2003.

[50] M. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *Proc. CVPR*, volume 2, pages 334–341, 2004.

[51] M. Lee and R. Nevatia. Dynamic human pose estimation using markov chain monte carlo approach. In *WACV/MOTION*, pages 168–175, 2005.

[52] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proc. ECCV*, pages 3–19, 2000.

[53] D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. In *Proc. Roy. Soc. Lond. B*, volume 200, pages 269–294, 1978.

[54] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*, volume 1, pages 69–81, 2004.

[55] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[56] T. Moeslund and E. Granum. Sequential Monte Carlo tracking of body parameters in a sub-space. In *Proc. Workshop on Anal. and Model. of Faces and Gestures*, pages 84–91, 2003.

[57] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. PAMI*, 23(4):349–361, 2001.

[58] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Proc. ECCV*, volume 3, pages 666–680, 2002.

[59] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proc. CVPR*, volume 2, pages 326–333, 2004.

[60] D. Morris and J. Rehg. Sigularity analysis for articulated object tracking. In *Proc. CVPR*, pages 289–296, 1998.

[61] R. Navaratnam, A. Thayananthan, P. Torr, and R. Cipolla. Hierarchical part-based human body pose estimation. In *BMVC*, 2005.

[62] V. Pavlovic, J. Rehg, T. Cham, and K. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proc. ICCV*, volume 1, pages 94–101, 1999.

[63] P. Perez, A. Blake, and M. Gangnet. JetStream: Probabilistic contour extraction with particles. In *Proc. ICCV*, pages 524–531, 2001.

[64] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *Proc. CVPR*, volume 2, pages 467–474, 2003.

[65] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, pages 271–278, 2005.

[66] L. Ren, G. Shakhnarovich, J. Hodgins, H. Pfister, and P. Viola. Learning silhouette features for control of human motion. *ACM Transactions on Graphics*, 24(4):1303–1331, 2005.

[67] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. ICCV*, pages 824–831, 2005.

[68] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *Proc. ECCV*, volume 4, pages 291–303, 2004.

[69] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *Proc. ECCV*, pages 700–714, 2002.

[70] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Proc. CVPR*, volume 2, pages 721–727, 2000.

[71] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2002.

[72] M. Ruzon and C. Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. PAMI*, 23(11):1281–1295, 2001.

[73] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Proc. ICCV*, volume 2, pages 750–757, 2003.

[74] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proc. ECCV*, pages 702–718, 2000.

[75] H. Sidenbladh, M. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *Proc. ECCV*, volume 1, pages 784–800, 2002.

[76] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *Proc. CVPR*, volume 1, pages 421–428, 2004.

[77] C. Sminchisescu and B. Triggs. Building roadmaps of local minima of visual models. In *Proc. ECCV*, volume 1, pages 566–582, 2002.

[78] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proc. CVPR*, volume 1, pages 69–76, 2003.

[79] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Trans. PAMI*, 25(7):814–827, 2003.

[80] N. Sprague and J. Luo. Clothed people detection in still images. In *Proc. ICPR*, pages 585–589, 2002.

[81] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *Proc. CVPR*, volume 1, pages 605–612, 2003.

[82] J. Sullivan, A. Blake, M. Isard, and J. Maccormick. Bayesian object localization in images. *Int. J. Comp. Vision*, 44(2):111–135, 2001.

[83] M. Swain and D. Ballard. Color indexing. *Int. J. Comp. Vision*, 7(1):11–32, 1991.

[84] M. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

[85] K. Toyama and A. Blake. Probabilistic exemplar-based tracking in a metric space. In *Proc. ICCV*, volume 2, pages 50–57, 2001.

[86] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *Int. J. Comp. Vision*, 63(2):113–140, 2005.

[87] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *Proc. ICCV*, pages 403–410, 2005.

[88] R. Urtasun and P. Fua. 3D human body tracking using deterministic temporal motion models. In *Proc. ECCV*, pages 92–106, 2004.

[89] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa. Role of shape and kinematics in human movement analysis. In *Proc. CVPR*, volume 1, pages 730–737, 2004.

[90] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, volume 1, pages 511–518, 2001.

[91] Q. Wang, G. Xu, and H. Ai. Learning object intrinsic structure for robust visual tracking. In *Proc. CVPR*, pages 227–233, 2003.

[92] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. ICCV*, pages 90–97, 2005.

[93] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D Active Appearance Models. In *Proc. CVPR*, pages 535–542, 2004.

[94] M. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. PAMI*, 24(1):34–58, 2002.

[95] J. Zhang, R. Collins, and Y. Liu. Representation and matching of articulated shapes. In *Proc. CVPR*, volume 2, pages 342–349, 2004.

[96] J. Zhang, R. Collins, and Y. Liu. Bayesian body localization using mixture of nonlinear shape models. In *Proc. ICCV*, volume 1, pages 725–732, 2005.

[97] J. Zhang, J. Luo, R. Collins, and Y. Liu. Body localization in still images using hierarchical models and hybrid search. In *Proc. CVPR*, volume 2, pages 1536–1543, 2006.

[98] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *Proc. CVPR*, pages 459–466, 2003.

[99] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *Proc. CVPR*, pages 406–413, 2004.