Efficient Human Pose Estimation from Real World Images

Timothy J. Roberts

February 2005

Abstract

Reliable, efficient human pose estimation from images is a precursor to many useful applications including advanced human computer interfaces, surveillance systems, image archive analysis and smart environments. Whilst progress has been made on human pose estimation, the research often makes strong assumptions about the appearance. In particular, assumptions are often made regarding the background, foreground, self and other object occlusion and number of viewpoints. These assumptions limit the application of computer human pose estimation systems. In contrast, the focus of this thesis is pose estimation from single real world images or monocular sequences of poorly constrained scenes. Furthermore, it aims to accomplish this efficiently. The body of the thesis is structured into three layers: formulation, likelihood and estimation.

First, the popular, generative, part based approach is extended to allow pose hypotheses that have different numbers of parts to be compared. This *partial configuration* formulation allows pose estimation in the presence of other object occlusion, enables efficient estimation and automatic (re)initialisation and gives robustness to body parts with a non-contrasting or poorly modelled appearance. The problem of comparing partial configurations is stated as a Bayesian decision problem of discriminating between the class of people and of backgrounds. To describe the body part shape a probabilistic model is learnt from manually segmented and aligned training data of multiple subjects in various poses. In order to obtain a low dimensional model, variations due to intra-person differences and clothing as well as difficult to observe degrees of freedom and differences between certain similar body parts are marginalised over. The resulting model allows uncertainty in measurements to be quantified as well as improving estimation efficiency. Finally, a prior is developed to

encode inter-part constraints and it is shown that due to these constraints smaller configurations contain much of the information of larger configurations.

Although a strong likelihood model is critical in determining the success of human pose estimation many existing models have limitations in terms of discrimination and efficiency when applied to real world images. Therefore, two novel techniques are developed to discriminate people with complex, textured appearance from cluttered backgrounds. A boundary model is developed based upon the divergence between the appearance distribution of the foreground region and its adjacent background. In particular, the distribution of the divergence between the joint colour histograms of these regions is learnt for correct and incorrect configurations. In order to provide a quantitative empirical evaluation the statistics of intensity edges on and off human boundaries are also learnt. It is shown that the new boundary model is more discriminatory and searchable. This is particularly important as early identification of body parts focuses the estimation. Next, a model is proposed that encodes the spatial structure of human appearance. In particular, the statistics of the similarity between regions on the surface of correct and incorrect configurations are learnt. Encoding inter-part similarity is important in discriminating larger incorrect configurations, and due to the combinatorial growth in the number of large configurations is key to efficient estimation in real world images. In addition to these likelihood models a foreground model is developed that encodes the expectation of temporal consistency in appearance for use in human tracking applications. It builds upon previous techniques by matching feature distributions and using clothing structure to improve estimation of the adapting foreground appearance.

Once the model and likelihood have been defined pose estimation can be performed. Two approaches to pose estimation can be identified in the literature. The combinatorial approach identifies body part candidates in the image and then combines the results, for example using dynamic programming, to estimate the overall body pose. Whilst such methods are efficient they rely heavily upon body part

detection which is particularly difficult in the presence of occlusion and clutter. In contrast, the full state space approach searches for whole body configurations and thereby models the complex self occluding appearance. However, due to the high dimensional space such methods use local rather than global sampling and require manual initialisation. By taking advantage of the *partial configuration* formulation and the strong likelihood model a straightforward deterministic search algorithm is able to recover many of the body parts and results of such a search to challenging scenes are presented.

Acknowledgments

I would like to thank my primary supervisor, Stephen McKenna, for all his guidance and friendship. His natural ability at explaining complex issues with great clarity and insight are second only to his amazing memory regarding authors, dates, titles and conferences! Thanks to everyone else at the Department, especially Ian and Gordon, for providing a great environment for discussion and development. To Em, thank you for your constant love and support and listening to my ramblings on our many walks and help with image markup. To Mum and Dave, thank you for your love, support and encouragement. Special thanks to Jan, at West Cheshire College, who cultivated my interest in computing and encouraged me to pursue my education. Finally, I would like to thank all the people that have worked to develop the open tools and free access to information upon which I have relied, especially Citeseer and the Annotated Computer Vision Bibliography.

Declaration

I hereby declare that the work described in this thesis is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted the references listed herein.

Signed

Timothy Roberts

September 2004

Declaration by the supervisors

We certify that Mr. Timothy J. Roberts has satisfied all the terms and conditions of the regulations made under Ordinances 12 and 39 and has completed the required nine terms of research to qualify in submitting this thesis in application for the degree of Doctor of Philosophy.

Signed

Prof. Ian W. Ricketts

Dr. Stephen J. McKenna

Associated Publications

T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Adaptive learning of statistical appearance models for 3D human tracking. In British Machine Vision Conference, pages 333-342, Cardiff, 2002.

T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Online appearance learning for 3D articulated human tracking. In International Conference on Pattern Recognition, 2002.

T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Human Tracking using 3D Surface Colour Distributions. Image and Vision Computing Under Review, 2003.

T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In European Conference on Computer Vision, 2004.

Contents

List of Figures

Chapter 1

Introduction

This thesis describes a system for estimating human body pose from single real world images. In particular, it aims to address what are felt to be two key problems in computer vision based human pose estimation:

1. Discriminating between a subject with a complex, unknown appearance and a cluttered, unknown scene that possibly occludes parts of the subject.
2. Formulating the pose estimation problem such that efficient, accurate global estimation is possible in such conditions.

Indeed, in spite of the recent increase in research into human tracking circa 2002 (Moeslund and Granum [2001]) the limitation of most systems to synthetic scenes and manual (re)initialisation remains.

Before starting the body of the thesis this introductory chapter presents a set of example results and a diagram that illustrates a life-cycle of sample through the pose estimation system.

1.1 Example Results

First, in order assist understand of the objectives of system and its operating conditions the reader is referred to Figures 1.1 and 1.2 which illustrate the results of applying the pose estimation system. Both indoor and outdoor scenes are illustrated. The subjects are unknown and are wearing lose fitting, textured clothing. These images are typical of the those under consideration in this thesis. The output pose configurations with maximum score after searching for 2 minutes are overlaid. Notice that the number of identified body parts varies. In this system the visible parts is determined automatic. This is necessary to account for occlusion by other objects, such as the door in the final image.

1.2 System Outline

In order to assist the reader in understanding the overall structure of the system, Figure 1.3 presents a schematic that describes the life-cycle of pose hypotheses, or samples, through the system. In this system a pose hypothesis is defined as a set of N layered 2D rigid body transformation parameters. These parameters are used to transform a probabilistic shape model into the image plane, which in turn are used to form foreground and adjoining background feature distributions (colour histograms) for each body part. The relationship between these distributions is used to compute likelihood ratios that describe how much more likely this configuration is to be correct than incorrect. A search is performed for the largest likelihood ratio that also obeys a set of hard constraints on inter-part pose. Notice that this diagram, like the thesis concentrates on the likelihood model (as opposed to the search scheme, pose prior or motion model).

Figure 1.1: Typical results from applying the search for partial pose configurations to outdoor scenes. The samples with maximum score after searching for 2 minutes are shown.

Figure 1.2: Typical results from applying the search for partial pose configurations using the divergence based likelihood model to indoor scenes. Notice that in one case a door partially occludes the subject.

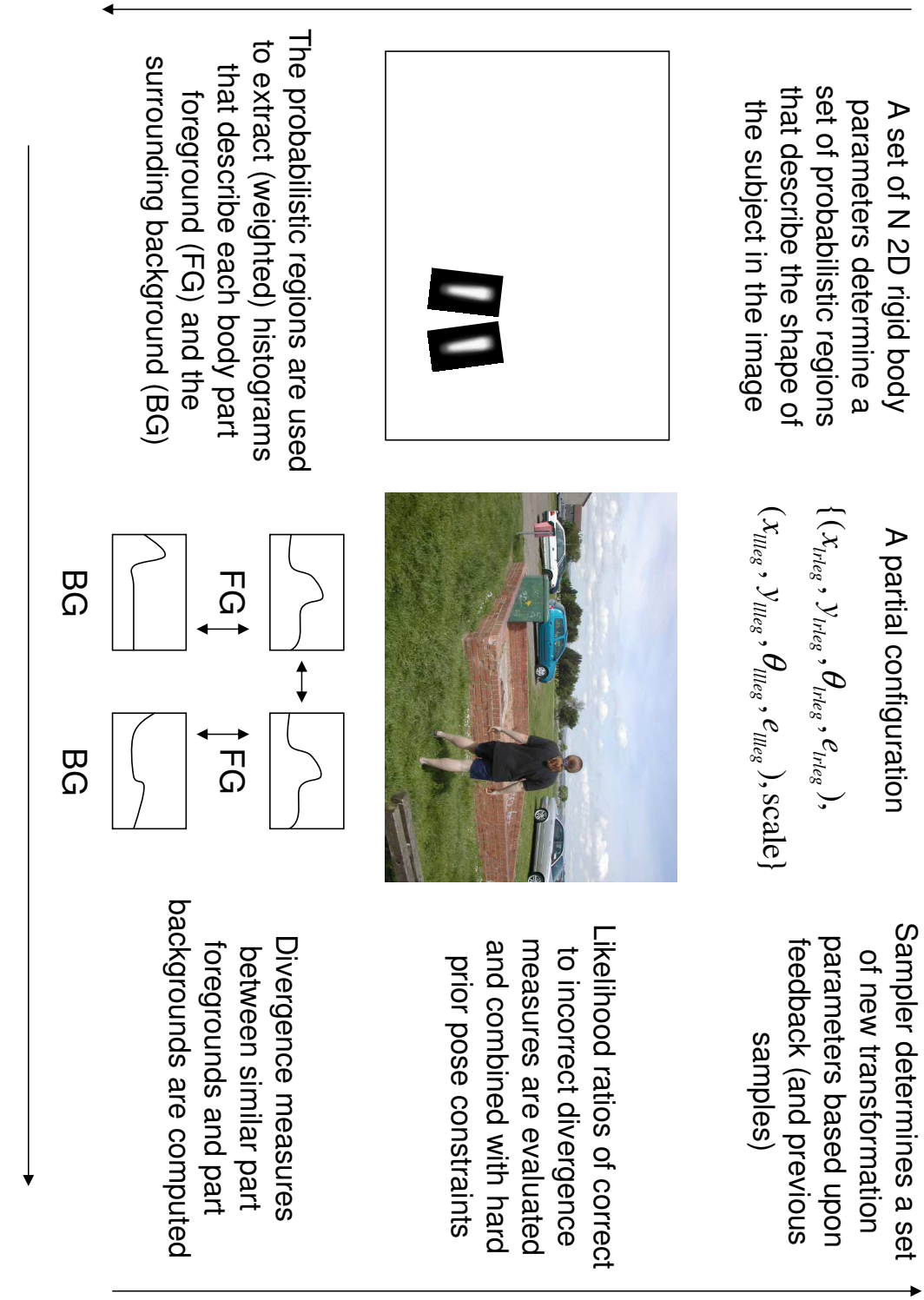

Figure 1.3: A diagram showing the life-cycle of a sample through the pose estimation system. This is also known as the analysis by synthesis loop. This diagram is only meant to guide the reader to understanding how the components of the system fit together.

1.3 Thesis Outline

The body of the thesis is divided into five Chapters. Chapter 2 discusses the problem domain, focusing on applications that require visual pose estimation and discussing what is meant by real world images and pose. The next three Chapters discuss the system in detail. In particular, Chapter 3 introduces the probabilistic, analysis by synthesis approach and explains why part based models are appropriate. Within this Bayesian, part based formulation, *partial configurations* are presented as the partial solution to problem 2. Chapter 4 builds upon this formulation and introduces a novel, highly discriminatory likelihood model based upon both *differences* between feature distributions in the foreground and background and *similarities* between texture on foreground points and thereby addresses problem 1. Chapter 5 builds on the strengths of the new formulation and likelihood and develops a computationally feasible estimation scheme and presents results of pose estimation. The final Chapter attempts to summarise the system as a whole and discuss the important limitations of the work. From this outline it should be clear that each Chapter in this thesis relies on the previous ones for its context and methods and therefore the chapters are best read in order.

Each of the main Chapters follows a similar format. First, in order to provide a clear structure and explicitly identify aim and scope, the chapters begin by posing the key questions that will be addressed in that chapter. Next, a review of research relating to that chapter is presented. Then after identifying the key limitations of previous works the novel methods developed in this work are presented. Finally, to summarise the chapter, the key questions posed at the start are briefly re-answered.

Chapter 2

Problem Analysis

2.1 Introduction

This chapter considers the pose estimation system as a 'black box' and concentrates upon analysing typical input images from various scenes and the requirements of the pose output (Figure 2.1). In analysing the problem domain this chapter attempts to answer the following questions:

- Why use computer vision for pose estimation?
- What are real-world images and why are they difficult to model?
- What is meant by pose and what level of detail is appropriate?

These questions are best answered by considering the applications that require or could benefit from computer vision based human pose estimation.

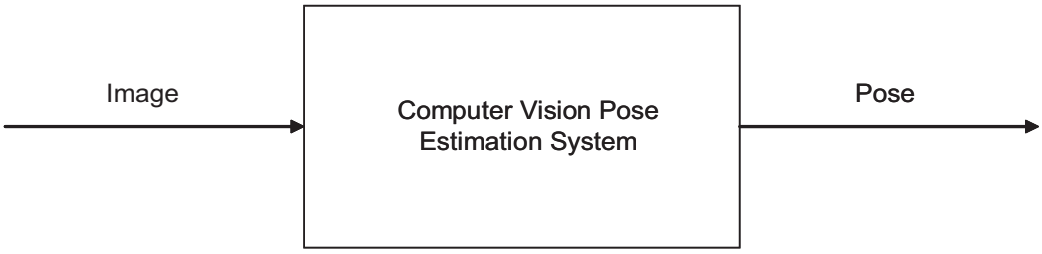

Figure 2.1: A pose estimation system as a black box that transforms (or compresses) images of people, with all their irrelevant details, into a description of pose that an application can easily process. This chapter analyses typical image inputs and discusses the requirements on pose output.

2.2 Applications

There are many applications, both current and envisaged, that depend upon the interpretation of human pose and motion. A perfect pose estimation system would be low cost, non intrusive, accurate and precise. Unfortunately, no such system is available. In reality, these characteristics are often competing and it is therefore unlikely that a one size fits all approach will be successful. For example, to achieve high levels of accuracy and precision usually requires a more expensive and intrusive sensing system. Based upon these differences in emphasis three groups of applications emerge:

Low Detail Examples of low detail applications include existing surveillance systems and smart environments. These applications must operate over long periods of time, in-place and therefore must be robust and operate in real-time. They usually require or emphasise non intrusive sensor modalities and unconstrained environments. These constraints and the nature of the application itself usually result in a low detail pose parametrisation such as presence, or position in a room, although a more detailed pose description could enhance these applications.

Medium Detail Examples of medium level applications include computer games,

virtual reality (e.g. Hilton [1999]) and high bandwidth human computer interfaces. Such applications usually emphasise a low cost solution. Constrained environments and intrusive sensors (such as special gloves for computer games and those considered in Frey et al. [1995]) are sometimes possible but not ideal. Fast or real time processing is often important. These applications usually only require a parameterised description of pose, such as the position and motion of points on the body.

High Detail Examples of high detail applications include medical and professional sports analysis. The most important requirement is the recovery of detailed, accurate pose, usually in the form of 3D structure. In the case of sports analysis this must be performed in the presence of fast motion. Because of the emphasis on accuracy and detail systems are usually expensive and or intrusive. Moreover, these applications are often performed in constrained environments. Off-line processing is often adequate.

2.2.1 Qualities of Computer Vision Systems

A computer vision based solution is clearly necessary when only images or video are available. However, computer vision techniques also provide a compelling alternative to other sensing techniques in applications with certain requirements:

Cost Colour video cameras, such as WebCams, and high performance desktop computers are relatively cheap and widely available technologies for capturing and analysing large amounts of information.

Intrusive In comparison to electromagnetic sensing systems (e.g. Polhemus) cameras are non-intrusive. Some vision systems have relied upon slightly more intrusive markers. There are however privacy issues associated with video

capture. Computer vision systems often assume or require some constraints on the scene.

Accuracy Compared to sensing modalities such as electromagnetic sensors and laser scanners, computer vision is a low accuracy, low precision solution. To achieve higher accuracy and precision often requires carefully calibrated, multiple camera rigs.

Reliability In less constrained environments computer vision systems are usually less reliable than more direct sensing modalities and therefore should not be used in critical situations.

2.2.2 Application Focus

Computer vision techniques have been successfully used for low detail applications such as surveillance. Computer vision systems have also been somewhat successful in highly constrained, high detail applications, although the author feels that other sensing modalities are sometimes more applicable. The focus of this thesis is upon medium detail applications since they are arguably the most promising in terms of future applications. *A solution to this class of application will allow the automated interpretation of pose from existing unconstrained images and video material.*

2.2.3 Analysis of Image Input

Since the majority of existing images are in colour the focus is upon this modality instead of range and infrared modalities. No suitable, standard human pose image databases exist. Therefore images were obtained specifically for the purpose of this work. These images were obtained with a range of still cameras and video cameras.

In particular, frames from the Sony DSW, Sony PC10-E and Canon MV600 video cameras were used, all of which allow direct digital video transfer. The Nikon CoolPix 775 and Konica Minolta A2 were used for still image capture. Some of these cameras use compression (such as JPEG) and this was either turned off or set to the highest quality settings. It is assumed that the intrinsic camera parameters are unknown (or rather uncertain). The resolution of the majority of the input images is 640×480 pixels. This is an appropriate choice since it corresponds to the lowest resolution of modern digital cameras and mini DV camcorders. Examples of typical input images are shown in Figure 2.2.

From the point of view of the pose estimation system, the set of images contain two types of variation: relevant and irrelevant. In this case the relevant variation is due to pose. It can be seen that there is a large and complex irrelevant variation between the images. This irrelevant variation is the primary cause of the difficulties in human pose estimation from images. Indeed, as will become clear, many pose estimation systems are only successful when the irrelevant variation is significantly constrained. Regarding these images the following observations and assumptions are made:

Subject A single subject is assumed to be present in an unknown pose that is to be recovered. The shape and identity of the subject are only known approximately. The appearance (e.g. clothing) of the subject is unknown. It can also be seen that the clothing can be loose fitting and therefore have a complex outline. Furthermore, the clothing can be textured in complex ways and at many scales.

Scene The images are of both indoor and outdoor scenes. It is assumed that the scene has an unknown structure and can contain clutter with a similar scale to the human body. Different types of scenes lead to differences in the shape

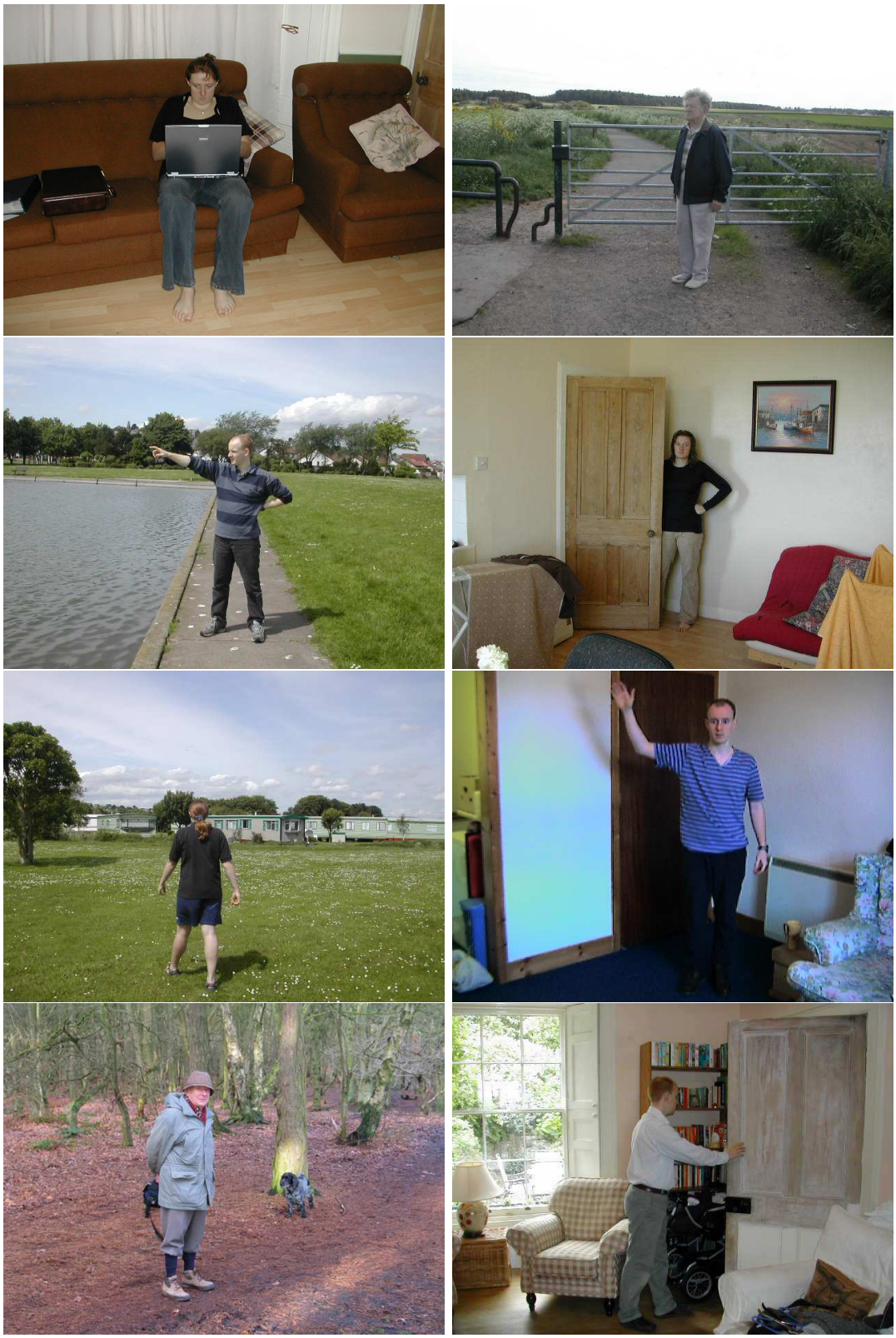

Figure 2.2: Examples of typical images that will be used for human pose estimation

of typical objects that are visible (and presumably differences in pose of the person). Objects in the scene can have a textured appearance.

Viewpoint It is assumed that the scale of the subject is known only approximately. Furthermore, it is assumed that the class of viewpoint (e.g. overhead, profile) is known, although the system should be able to be retrained to work with another viewpoint.

Occlusion It can be seen that in many images a portion of the subject is not visible. This problem of occlusion is a key difficulty in visual human pose estimation. Occlusion can be caused by another portion of the subject and/or another scene object.

Perspective It is assumed that perspective effects are weak. In particular, it is assumed that these effects are small when compared to intra-person variability. Perspective effects can be modelled if the intrinsic camera parameters are known.

Colour All the images are in colour. It has been observed that the colour signal is often very noisy and that some of the cameras have poor colour reproduction.

Illumination The images can have complex, uneven illumination from multiple sources. In some cases the illumination is so poor that certain body parts cannot be distinguished (consider for example the leg in the image in the third row and second column of Figure 2.2). Some images contain strong cast shadows and self shadowing often occurs.

The difficulty in modelling real world images is also apparent in the complexity of graphical models used to synthesise realistic images of clothed humans, which is still under active research (Jojic et al. [1998]).

2.2.4 Requirements on Output

In this thesis, pose (or a pose sample) is taken to be a geometric configuration of a set of body parts (valid or otherwise, correct or otherwise). However, it is recognised that pose need not involve a decomposition into body parts. Medium detail applications typically require the position of key body parts, such as the head and hands, and their motion. This is in contrast to high detail applications that usually require precise information regarding body part shape and position. Consider, for example, an image or video query application (that might be used for surveillance). Such an application might allow the user to search for people (or an individual) in a particular pose or, in the case of video, making particular gestures. Such an application would not require highly precise pose, but rather would need to be able to accurately discriminate between different poses such as standing and sitting. For the purpose of this thesis it is assumed that the system does not require personal metric details such as body sizes to be recovered. Images could be missing information due to occlusion and the system might be required to estimate body pose given the pose of the visible portion.

2.3 Summary

To summarise this chapter, the questions that were posed in the first section are briefly revisited:

Why use computer vision for pose estimation? This chapter identified applications that require the interpretation of human pose and observed that these applications emphasise differing characteristics. Computer vision based systems allow low cost, non-intrusive human pose estimation. Furthermore, a

computer vision solution is necessary for the automated analysis of existing data. Therefore, the analysis of single, real world colour images was identified as a compelling application of computer vision.

What are real-world images and why are they difficult to model? Typical real world images were presented in order to identify the inherent difficulties in estimating pose. These real world images contain an unknown subject with a complex, unknown appearance that can be occluded by both other portions of the body and other objects in the scene. These images also contain an unknown, cluttered background.

What is meant by pose and what level of detail is appropriate? For medium detail applications, pose is described in terms of the position of around 10 key points on the body. This is in contrast to low detail pose (typically 1 to 3 parameters, such as presence and coarse pose) and high detail pose (typically 100s of parameters describing the precise pose and structure). Such applications do not need precise, accurate measurements but rather only enough information to discriminate different poses and gestures and thereby function.

Chapter 3

Formulation

3.1 Introduction

The previous chapter presented an analysis of the problem domain, its typical inputs and the requirements on the output. A promising class of applications was identified and will be the focus for the remainder of this thesis. This chapter considers how a pose estimation system suitable for such applications should be formulated. In particular the following questions will be answered:

- What approach is best for monocular, medium detail applications?
- What are the limitations of current descriptions of pose?
- How can pose estimation be performed in the presence of occlusion?
- How should uncertainty in the subject's shape be represented?
- What constraints exist on the pose?

3.2 Related Research

This section considers previous research relating to the formulation of human pose estimation systems. It begins very broadly by considering the probabilistic framework that is usually used to represent uncertainty, infer quantities and make decisions in the presence of such uncertainty. Then detailed consideration is given to the different descriptions of pose. Finally, a short discussion of motion models is presented.

3.2.1 Mathematical Framework

A camera maps the state of the world to an image, denoted by **I**. Unfortunately, this mapping is non-linear, corrupted with noise and artifacts and loses information. This introduces uncertainty in the state of the world given the image. One can consider computer vision to be the task of inverting this mapping to allow estimation of models of the world and their uncertainty. It is emphasised that this does not imply 3D reconstruction; any model of the world could be used. In our case, the model is one of human pose which is to be developed in this Chapter.

In order to represent uncertainty and make decisions, many human pose estimation systems and vision systems in general, are formulated in terms of probability theory. Probability theory is a well developed, principled logic for inference under uncertainty that follows from the Cox Jaynes axioms (Cox [1946], Jaynes [1986]). Whilst the application of probability theory is important in modelling the inherent uncertainty in the image, it is also important in the construction of an efficient system, since decision making can proceed in the presence of limited data and assumptions. In probability theory, the inverse mapping from the data (image) to model (pose) is represented using a conditional probability distribution $p(model|data)$. This dis-

tribution allows expectations of quantities to be calculated and optimal decisions to be made. Many pose systems follow a Bayesian approach where probability is interpreted as a degree of belief. Bayes rule is a key part of probabilistic modelling since it allows the dependency between the model and data to be reversed:

$$p(model|data) = \frac{p(data|model)p(model)}{p(data)} \tag{3.1}$$

$$posterior = \frac{likelihood \times prior}{evidence} \tag{3.2}$$

This reversal is key to understanding the approach of many pose estimation systems, including the one described herein, since it allows the principled application of the *analysis by synthesis* approach to pose estimation, where pose models are hypothesised and compared to the image. Bayes rule also allows additional, perhaps subjective, prior information to be incorporated, which is important in situations like human pose estimation when the likelihood is uncertain. Probability theory is also related to information theory which is used for example in model selection and density comparison. An excellent introduction to Bayesian methods with examples motivated by computer vision problems is given in Duda et al. [2001]. To complete optimal decision making requires the quantification of risk and it is interesting to note that this emphasises that interpretation cannot be divorced from purpose.

Applying the generative, Bayesian probabilistic modelling framework to spatial pose estimation yields the architecture illustrated in Figure 3.1. Many human tracking systems build upon this spatial framework by assuming a Markov relationship between frames and thereby obtain a temporal prior (e.g. Sidenbladh et al. [2000a]). Taken together these approaches are essentially the probabilistic manifestation of the model based architecture proposed early on by O'Rourke and Badler [1980].

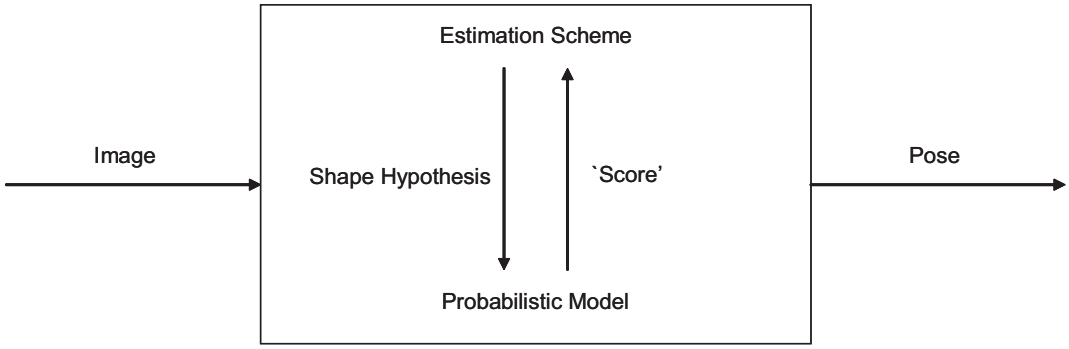

Figure 3.1: The overall structure of the pose estimation system. In order to estimate pose an analysis by synthesis, or hypothesis and test, approach is adopted. For spatial models, the estimation scheme hypotheses shape configurations which are measured using the probabilistic model.

With this architecture in place the following groups of problems must be addressed to complete a pose estimation system:

Formulation A model of pose must be developed that describes the variation of human shape. Prior knowledge on pose should be encoded to constrain the problem.

Likelihood A likelihood model must be constructed that discriminates correct pose configurations (i.e. those that correspond to people) from incorrect ones based upon measurements in the image for the hypothesised shape.

Estimation A computationally feasible estimation scheme must be developed that searches for probable pose hypotheses.

These three groups of problems correspond directly to the structure of the body of this thesis (Chapters 3–5). The success of a pose estimation system relies upon solutions to all these modules (although a particular system may emphasise one over another).

3.2.2 Pose Estimation Reviews

There is a great deal of inter-related research on analysing images of people: human tracking, gesture analysis, activity recognition, face detection and recognition, hand tracking etc. Much of the research regarding whole body pose estimation focuses on tracking a manually initialised model over sequences rather than pose estimation from single images. Furthermore, much of the research is relatively recent; see for example the chart of publications by year presented in Moeslund and Granum [2001]. Due to the large volume of work, difference in emphasis and topical nature, this thesis does not attempt to provide a complete review. The two most relevant review papers concerning analysis of images of people are summarized below.

- Gavrila [1999] began by considering applications which require pose estimation and grouped these into domains: virtual reality, smart surveillance, advanced user interfaces, motion analysis and model based video encoding. Previous pose estimation systems were classified according to the nature of the shape model: 2D without explicit shape, 2D with explicit shape and 3D approaches. This emphasises the importance of the shape model, the choice of which is "largely application dependent". The review also considered hand tracking research since there are parallels in the articulated structure. Human activity recognition was also considered briefly. In the conclusion some important remaining challenges were listed: body model acquisition, significant occlusion, initialisation, modelling loose clothing, pose ambiguity, using 3D data, modelling physical constraints and quantitative evaluation and comparison. Several of these challenges are addressed in this thesis.

- Moeslund and Granum [2001] and the accompanying paper summaries in Moeslund and Bajers [1999] have a larger scope and cover more recent work.

They considered four possible stages in a pose estimation system: initialisation, tracking, estimation and recognition. This structure emphasises the importance of estimation. A useful aspect of this review is the explicit identification of assumptions made by pose estimation systems. The assumptions are grouped into two categories: motion and appearance. Assumptions on motion include single subject, no camera motion, parallel motion, slow and continuous motion, limited occlusion. Assumptions on the appearance include known camera parameters, uniform or static background and constant lighting.

Cedras and Shah [1995] and Aggarwal and Cai [1999] also discussed human tracking but are less relevant since they focus even more upon human motion analysis.

3.2.3 Models of Pose

A model of human pose describes the shape of a person in single images. A model must be able to describe the variation of pose change illustrated in Figure 2.2. Since this variation is large and complex these models can have many parameters. Occams razor suggests that a good model will capture this variation compactly and will encode constraints to reduce the entropy of the state space. It is also important, from the point of view of efficient decision making, that the complexity of the model correspond with the requirements of the application and available input data. Since different applications emphasise different requirements and have different available inputs it is not surprising that various models of pose have been developed. These models can be contrasted in a number of ways. Some models concentrate on modelling the appearance in the image whilst others explicitly model the physical structure of the subject and acquisition system. Some approaches model the body as a whole whilst others break the body into component parts. The

remainder of this section is concerned with discussing and comparing the qualities of the different types of pose models.

Coarse Image Based Models

Coarse image based models describe the subjects using measurements such as position, orientation and scale in the image. For example, Fablet [2002], Oren et al. [1997] and Sidenbladh [2004] recovered the position of image windows around multiple pedestrians in cluttered scenes. Comaniciu et al. [2000] used a constrained ellipse to represent the body for real-time tracking. Such coarse descriptions can be used when the aim is to detect or count people and track coarse interactions between people (low detail applications). The low dimensional representation allows the state space to be sampled globally and makes real-time implementation possible on current hardware which is important for such applications. However, these models usually rely upon a small variation in pose to achieve such a compact representation. Comaniciu et al. [2000] managed to circumvent this problem by assuming the foreground appearance is known and using colour features that change little through occlusion and changes in pose. However, these models still provide only limited information about the pose of the subject. Fablet [2002] reasoned that these coarse models are also useful in initialising more complex models of pose. However, this automatic initialisation is only applicable for constrained poses for the aforementioned reason.

A related approach is to use low level image statistics directly for recognition. Polana and Nelson [1994] advocated this approach for recognising repetitive actions such as walking and running using motion. Regions of independent motion are segmented and normalised spatially and temporally before matching to class templates. The state space is the position, scale and frequency parameters along with the recovered

motion class. Although such low level recognition approaches result in a fast, robust system they require training for specific applications.

Contour Models

A more detailed image based representation is the contour. Contours are used throughout vision to describe shape, see for example the survey in Loncaric [1998]. Active contour models, or snakes, were introduced in the seminal work of Kass et al. [1988] to segment images based upon the minimisation of an energy function composed of image evidence terms, such as edge strength, and regularisation terms such as contour smoothness. Since then many variations have been developed, see for example Blake and Isard [1998]. The contour modelling approach was further developed by active shape models in which contour deformations are learnt from training data (Cootes and Taylor [2001]). In particular, principal component analysis (PCA) is applied to aligned shape instances to produce a low dimensional, linear representation of the shape.

Baumberg and Hogg [1994] used such an approach to learn the shape deformation of the outline of pedestrians from noisy image sequences. Tracking was then performed by local optimisation in translation+PCA space and dynamic filtering using a Kalman filter. Although this resulted in a real time pose estimation system this was achieved by restricting the pose and viewpoint and sampling locally.

A key advantage of contour models is the ability to learn shape variation. This is most useful when limited prior knowledge on shape and structure is available. However, this is not the case for human tracking where the body structure is well understood and anthropometric measurements exist, such as Grosso et al. [1989]. Furthermore, current approaches to learning these models often make limiting as-

sumptions, such as a linear shape space, that can allow physically implausible contours. Non-linear contour models have recently been developed to better model the space of deformations, see for example Bowden [2000].

Contour models are particularly relevant in applications where the background is known or can be fixed. For example, Haritaoglu et al. [1999a,b, 2000] used statistical background models to segment multiple subjects. Higher level techniques can be employed to remove noise and shadows (McKenna et al. [2000], Zhao et al. [2001]).

Human contours have been analysed in a number of interesting ways. For example, Rosales and Sclaroff [2000] and Rosales et al. [2001] described a specialised mappings architecture that is used to infer possible 2D joint locations from the Hu moments of silhouettes. Interestingly, the learning architecture implicitly represents the part based nature of the human body. The technique does not require calibrated cameras and when multiple views are available Expectation Maximisation (EM) is applied to find the most consistent 3D pose and the associated camera views. Haritaoglu et al. [2000] used the silhouettes to recovered body parts in horizontal walking sequences. Haritaoglu et al. [1999a] used contour symmetry and temporal periodicity to detect whether a person was carrying an object and to segment that object. Tabb et al. [2000] applied a neural network to the distances of points around the contour from the centre to detect and analyse human motion. Wilhelms et al. [2000] used an active contour model and interactive correction to recover a 3D body model.

Contours models have also been combined with other models of pose. For example, Bowden et al. [2000] learnt a non-linear point distribution model for combined contour and 2D part measurements to estimate the 3D pose. Ong and Gong [1999] proposed a similar system but learnt a hierarchical PCA model and used multiple, fused viewpoints. Such hybrid models help disambiguate self occluding poses, such as the hands moving over the body, that often occur in the analysis of images of

people, and thereby resolve a key disadvantage of the contour representation.

Many contour based systems assume a manual initialisation is available and then proceed by iterative refinement. In order to tackle the problem of detecting (multiple) contour configurations MacCormick and Blake [1998a] proposed a probabilistic discriminant based upon the ratio of the likelihood of the edge measurements being foreground to the likelihood of them being due to background clutter. This is a similar approach to the one adopted in this thesis. When combined with importance sampling this allows global sampling of an image containing multiple low dimensional targets (in this case the head and shoulders). This partially addresses the problem of automatically initialising the contour model but is not practical for more varied objects.

MacCormick and Blake [1998b] built upon the probabilistic discriminant approach in order to tackle the problem of applying contour models in the presence of other object occlusion. The spatial structure of measurements during occlusion was learnt and used to discriminate occlusion events from weak measurements.

Part Based Models

The structure of the body is well known and a part based description is natural since it corresponds to our rigid bone structure. Part based models use a fixed number, typically 10–14, of body parts that are transformed into the image to model the shape of the subject as a whole.

A common approach to describing the shape of individual parts has been to use geometric primitives. Various levels of sophistication of primitive have been considered. For example, Ju et al. [1996] proposed the cardboard person model where simple rectangular 2D patches were used to model the body. Cham and Rehg [1999] also

used 2D rectangular image regions. Leung and Yang [1995] described a 2D ribbon model where decisions on the ribbon identity were based upon its termination. Wren et al. [1997] described the P-finder system which represents the head and hands using 2D elliptical regions corresponding to the projection of spatial-colour clusters or 'blobs'. Three dimensional body part primitives have also been used. Early work by Hogg [1983] used 3D cylindrical primitives to model body parts. Wachter and Nagel [1999] used cones with elliptical cross sections to model parts. A still more general set of shapes are super-quadrics, as described in Metaxas and Terzopoulos [1993]. Tapered super-quadrics have been used to model the arm (Kakadiaris and Metaxas [1996]) and the whole body (Gavrila and Davis [1996]).

In order to use the part primitives, parameters such as size must be specified. An advantage of physically motivated models is that accurate anthropometric data is available (e.g. Grosso et al. [1989]). However, these data correspond to precise shape descriptions and therefore are infrequently used. Rather the part parameters are usually manually fitted. For example, Kakadiaris and Metaxas [1996] used multiple orthogonal views to automatically estimate the parameters of a super-quadric model. The inter-scene variation due to subject identity is usually ignored. A notable exception is Sminchisescu and Triggs [2001] whose detailed, high dimensional, human model used anthropometric data to specify a prior on the part sizes. Intra-scene variation due to non rigid deformation and clothing motion has not been addressed.

In a part based representation, pose is described in terms of transformations into image space of the part primitives. It is common to consider the body as a tree structure, with the torso as the root, and chain the transforms hierarchically, thereby capturing the kinematic structure of the human body. For example, Cham and Rehg [1999] described the hierarchical scaled prismatic transform which used translation and rotation in the image plane to transform 2D rectangular patches. Although 2D transformations are more compact and easier to visualise than 3D transformations

they do not allow 3D shape variation and perspective effects to be modelled. Various parameterisations of 3D part transformations have been proposed. Wachter and Nagel [1999] used hierarchical 3D transformations with rotation parameterised using Euler angles. However, as the authors pointed out, this representation has singularity problems. Singularity problems arise when changes in state become unobservable. Morris and Rehg [1998] demonstrated that the 2D scaled prismatic parameterisation does not suffer from such problems. Bregler and Malik [1998] described the 3D twists and exponential map formulation, also used in robotics, where rotations are described by a rotation and a unit vector around which the rotation is performed. This formulation linearises the kinematics and removes the singularity problems. Deutscher et al. [1999] point out that singularity problems can also be overcome by a random sample estimation scheme such as Condensation (Isard and Blake [1996]) that does not rely upon local derivatives. In order to model complex joints like the shoulder (which is actually composed of four bones, the thorax, clavicle, scapula and humerus) it is necessary to consider a more flexible representation than relative orientation. Baerlocher and Boulic [2001] investigated the parameterisations of ball and socket joints. Many systems account for joints like the shoulder by allowing relative translation of parts. Although the relative translation is constrained by a spring force the size of the state space is still increased. Moeslund and Granum [2000] considered reducing the size of the state space for the arm by using a phase space. Part based transformations also allow inverse kinematics techniques, such as Zhao and Badler [1994], to be applied.

Three dimensional part models account for self occlusion by representing depth and using standard hidden surface removal. In order to account for self occlusion with 2D part models, Rehg and Kanade [1995] proposed representing a depth ordering. Although depth ordering is a less principled approach it is a more compact representation. It is not, however, necessary to predict self occlusions at all. For example,

Cham and Rehg [1999] were able to track through un-modelled occlusion events by propagating multiple hypotheses through the occlusion events. Furthermore, even with a detailed 3D model it can be difficult to describe the appearance of partially occluded parts. To address this Kakadiaris and Metaxas [1996] proposed actively choosing viewpoints from which to compute the likelihood based upon part visibility.

One significant advantage of physically motivated 3D models is that hypothesises can easily be related between multiple views. For example, Gavrila and Davis [1996] used four "nearly" orthogonal views to estimate pose. However, the focus of this thesis, and of much other human pose estimation and tracking research, is on monocular estimation. Barron and Kakadiaris [2001], Taylor [2000] and Mori and Malik [2002] studied the inherent problems of monocular estimation by estimating 3D pose from known 2D joint points in images taken with cameras with uncertain parameters.

Part based models also allow constraints on the body, such as joint angle limits to be encoded. Beyond simple joint angle limits priors over whole pose can be defined. For example, Karaulova et al. [2000] learnt a hierarchical PCA model to represent the 3D state space. Furthermore, physically motivated 3D models allow constraints based upon part inter-penetration to be expressed, e.g. Sminchisescu and Triggs [2001].

In summary, the part based approach is natural given the prior knowledge of the structure of the human body. In comparison to image based models, such as contours, using this prior knowledge removes much of the burden of learning highly varied shape models and implicitly accounts for non-linear changes resulting from self occlusion, for example. Furthermore, the part based parameter space is easier to interpret and allows constraints to be encoded more easily than contour descriptions. In comparison to physical models, view oriented models are more compact since (absolute) depth is not parameterised. The 3D models allow multiple views

to be related and self occlusion and perspective effects to be modelled explicitly but these advantages come at the expense of increased dimensionality and problems with ambiguities.

3D Surface Models

In certain applications, such as some in medical analysis, rigid part based models are too crude and knowledge of the 3D body shape is required. Due to the large number of degrees of freedom such models usually require more than a single image to estimate.

Early work by O'Rourke and Badler [1980] used a collection of intersecting 3D spheres to model the body. Other early work by Pentland and Horowitz [1991b] used a finite element model of the human body. More recently, Plankers [2001] and Plankers and Fua [2001, 2003] described a body modelling scheme where rigid body parts are replaced by articulated soft objects. Each of these objects defines a field which specifies the body surface. This model can then be fit to the subject's body using optimisation techniques and used for subsequent tracking. Mikic et al. [2001] considered the problem of fitting body parts to voxel data.

Commercial packages also exist to model humans; see for example the Poser from Curious Labs (http://www.curiouslabs.com/) and the VRML H-Anim specification (http://h-anim.org/). However, such models are designed for realistic synthesis and are less suited for visual analysis.

3.2.4 Models of Motion

The human body exhibits complex patterns of motion that are challenging to model and the research literature reports significantly better performance, due to constraints and better sampling, when appropriate motion models are employed. Motion models are also of independent interest for gesture recognition and motion synthesis. However, since the focus of this thesis is on developing spatial models for estimation from single images, motion models are considered only briefly here. Motion models usually make smoothness assumptions and it not clear how they would perform in the case of large other object occlusions.

Sidenbladh et al. [2000a] and Ormoneit et al. [2000] considered automated learning of models of cyclic motion. Galata et al. [1999] proposed using variable length hidden Markov models at two temporal scales to capture both the prototypical pose configurations and high level activity patterns. Pavlovic et al. [1999a,b] and Pavlovic and Rehg [2000] described the switched linear dynamic network. In this approach a Bayesian network is used to model the data, with a latent variable switching between different linear models. Different approximate inference techniques are derived and the work demonstrates the efficiency of the approach over hidden Markov models for fronto-parallel walking and jogging sequences. Tissainayagam and Suter [2000] also considered switched models. Zhao et al. [2002] presented an unsupervised approach based upon minimum description length vector quantisation techniques. They recover both motion primitives and transition probabilities.

3.3 Partial Configurations

In the remaining sections of this Chapter a novel formulation is developed. As we have seen, computer vision based pose estimation systems typically adopt a Bayesian, analysis by synthesis approach. This thesis also adopts this framework. Within this framework a representation of pose must be developed. It is not surprising that a good representation of pose is key to accurate measurement and efficient estimation. *Indeed, the author believes that some of the limitations of current pose estimation systems when applied to real world images are symptomatic of a poor representation and that these limitations cannot be resolved, either efficiently or at all, by simply improving the likelihood model and estimation scheme.*

3.3.1 Limitations of Current Part Based Formulations

A part based approach is adopted because, as discussed above, it naturally encodes the structure of the human body, corresponds well to the level of detail required by the applications, has an easy to interpret parameter space, models self occlusion and allows constraints on the body to be encoded. However, current part based formulations have some significant limitations.

The approach of *partial configurations* is a key contribution of this thesis as it removes many of the limitations of previous part based approaches. Recall that previous part based models use a fixed number of parts, determined manually prior to estimation. Although this seems a sensible physical model, given that the majority of people have a known fixed number of body parts, considering the number of body parts as variable leads to a better *visual model* for the following reasons.

Efficiency and Accuracy A key problem in visual pose estimation is efficiently

searching the high dimensional pose space for correct configurations. Indeed, most human tracking systems, which search the fixed, full dimensional pose space, rely upon manual initialisation. One approach to solving automatic (re)initialisation is to use background models or motion to initialise the model, but this is not possible with single image pose estimation. Instead current pose estimation systems use part detectors and grouping to efficiently estimate pose. However, part detection in the presence of clutter and self-occlusion is very difficult and the grouping of these detectors (often based on dynamic programming) leads to poor results. These approaches are discussed in more detail in Chapter 5. Describing the pose using *partial configurations* allows hypotheses with varying numbers of parts to be compared without making assumptions on the relationships between parts. Since configurations of different sizes can be compared and body parts can be related between configurations it allows the identification of small configurations to focus the search. For example, a system could begin by searching for individual un-occluded parts and then these could be combined and used to predict the position of other parts whilst modelling inter-part relationships such as occlusion. *In summary, partial configurations make pose estimation and automatic (re)initialisation possible without relying on ad-hoc models or assuming limited inter-part relations. Although the state space is larger when the number of parts is allowed to vary, this formulation allows more efficient and/or accurate estimation.*

Other Object Occlusion and Robustness In many formulations there is no model of, and little robustness to, other object occlusion. Since a 3D model of the scene is usually unavailable, and requiring one would limit the applicability of the system, other object occlusion cannot be predicted like self occlusion and it is therefore necessary to consider a different model to account for other object occlusion. Although current approaches allow the number of parts to

be identified prior to tracking, other object occlusion can occur sporadically in tracking sequences. Therefore, an automatic method for dealing with other object occlusion is necessary. Here it is suggested that occluded parts are modelled by simply not including them in the hypothesis. This is not the only possible approach to parameterising other object occlusion. For example, one might add a visibility variable to each part and make measurements based on the part not being present at that position. However, the approach of partial configurations has the advantage of allowing initial sampling in lower dimensional space to focus the search. Furthermore, if an individual part happens to be poorly modelled, for example due to a complex texture, poor illumination or structural differences (e.g. the subject is wearing a kilt!) the system should be able to proceed by estimating the pose of the remaining parts. Current systems are not robust to such conditions. In the proposed approach parts which are difficult to detect can be missing from the pose hypothesis. This adds a degree of robustness to such situations.

3.3.2 Approach

Partial configurations is a formulation for comparing pose hypotheses with variable numbers of parts. It is emphasised that possible partial configurations include single part hypotheses, fully parameterised hypotheses and everything in between and therefore subsume previous part based approaches. It is also emphasised that *partial configurations are not the same as independent parts.* From hereon, a partial configuration hypothesis, $\mathbf{C}$, is denoted by a set of parameterised part hypotheses, $\vec{c}$:

$$\mathbf{C} = \{\vec{c}_i\} \tag{3.3}$$

Clearly, for this approach to be useful it must be possible to compare partial configurations of different sizes. Moreover, larger correct hypotheses should be preferred to smaller correct hypotheses. First, consider how hypotheses in a fixed size state space are usually compared. The most popular approach is to find the maxima of the posterior $p(\mathbf{C}|\mathbf{I})$, i.e. MAP (maximum *a posteriori*) estimation. If, as is usually the case, the probability of the pose is not required, just the pose with maximum probability, it suffices to only compute the likelihood and prior and ignore the evidence. This assumes however that the image contains (at least) one subject, since if such a target did not exist a maximum would still be found and the system would have no idea if this was correct. This approach is not applicable in the case of partial configurations since essentially multiple models exist (some of which may not have a corresponding instance). One approach would be to compute the normalising factor, the evidence, for each combination of parts and thereby compute probabilities that can be compared. However, this is not computationally feasible.

Instead, the problem is treated as one of discriminating between subjects and the background *at each point in the state space*. The state space is therefore augmented by a class label v that labels the hypothesis as either for a person or for a background process. The optimum decision between the subject and background classes for a particular pose is found by choosing the class with the highest probability (assuming uniform risk) (Duda et al. [2001]). This is equivalent to forming the posterior ratio, $PR(\mathbf{C}|\mathbf{I})$, as given in (Equation 3.4) and classifying hypotheses as people when the ratio is greater than 1 and as background otherwise. The posterior ratio is related to the Bayes Factor methodology of model selection (Kass and Raftery [1995]).

$$PR(\mathbf{C}|\mathbf{I}) = \frac{p(\mathbf{C}, v = person|\mathbf{I})}{p(\mathbf{C}, v = background|\mathbf{I})} \tag{3.4}$$

Posterior ratios have been applied before in pose estimation. As observed in Siden-

bladh and Black [2001] this formulation emphasises detection and estimation based upon knowledge of both people and backgrounds. This approach has also been applied to images for detection of multiple targets (MacCormick and Blake [1998a]). Perhaps the most closely related work is that of Ioffe and Forsyth [2001a] where a mixture of trees was used to approximate the posterior ratio and achieve efficient estimation. However, the mixture of trees algorithm places restrictions on the form of the fitness function that make it less appropriate for analysis of real world images where initial limb localisation is difficult due to clutter, camouflage and occlusion. For example, the pairwise appearance constraint introduced in the next chapter cannot be incorporated into such a scheme.

The application of posterior ratios to comparing fully inter connected, variable part human models is novel. A fully inter-connected model (i.e. where the posterior ratio is a function of all parameterised parts) is necessary to accurately represent self and other object occlusion, and to encode the inter-part relations (such as the inter-part similarity model described in Chapter 4) that are necessary for reliable estimation in complex scenes.

The posterior ratio allows hypotheses from *multiple models* to be compared based upon how different each hypothesis is to a statistical process describing the background class. From the point of view of efficiency, *the key point is that whilst varying the numbers of parts creates multiple models the parts themselves can be related between models.*

Due to the structured appearance of people, hypotheses with larger numbers of parts are easier to discriminate from the background than ones with smaller numbers of parts. Therefore larger correct configurations have higher posterior ratios than smaller correct configurations.

3.4 Modelling Part Pose

The approach of partial configurations, as described above, can be applied to both 2D and 3D part transform parameterisations. However, a depth ordered 2D model is preferred for the class of applications under consideration since they emphasise:

Monocular Estimation Monocular estimation makes accurate estimation of depth difficult. Furthermore, genuine ambiguities exist with 3D monocular estimation that complicate estimation.

Limited Perspective Effects The class of images under consideration does not contain significant perspective effects.

Automatic Initialisation A more compact state space makes more efficient sampling easier.

Uncertainty in Shape The uncertainty in shape from perspective effects and 3D variation introduced because of the 2D model is comparable to the uncertainty due to clothing and intra-personal variability.

A disadvantage of the 2D model for the class of applications under consideration is the inability to encode extra physical constraints.

To describe the transform of the pose of a body part i into the image, a 2D parameterisation similar to the scaled prismatic parameterisation proposed in Cham and Rehg [1999] and Morris and Rehg [1998] is employed. Since partial configurations do not have a single root body part and parts in a chain are often missing, it is not possible to directly compare different hierarchical representations. Therefore, a non-hierarchical representation is employed. Although this removes the automatic

kinematic behaviour, it is not clear whether this behaviour eases estimation anyway. Since the non-hierarchical representation assigns independent transformation and orientation parameters to each part this makes the state space larger than using just relative orientation. However, as was observed above, it is often necessary for complex joints like the shoulder to use an inter-part distance. Since one can convert between the hierarchical and non-hierarchical representations the difference is one of convenience. In order to reduce the size of the state space a common scale parameter is used for all parts.

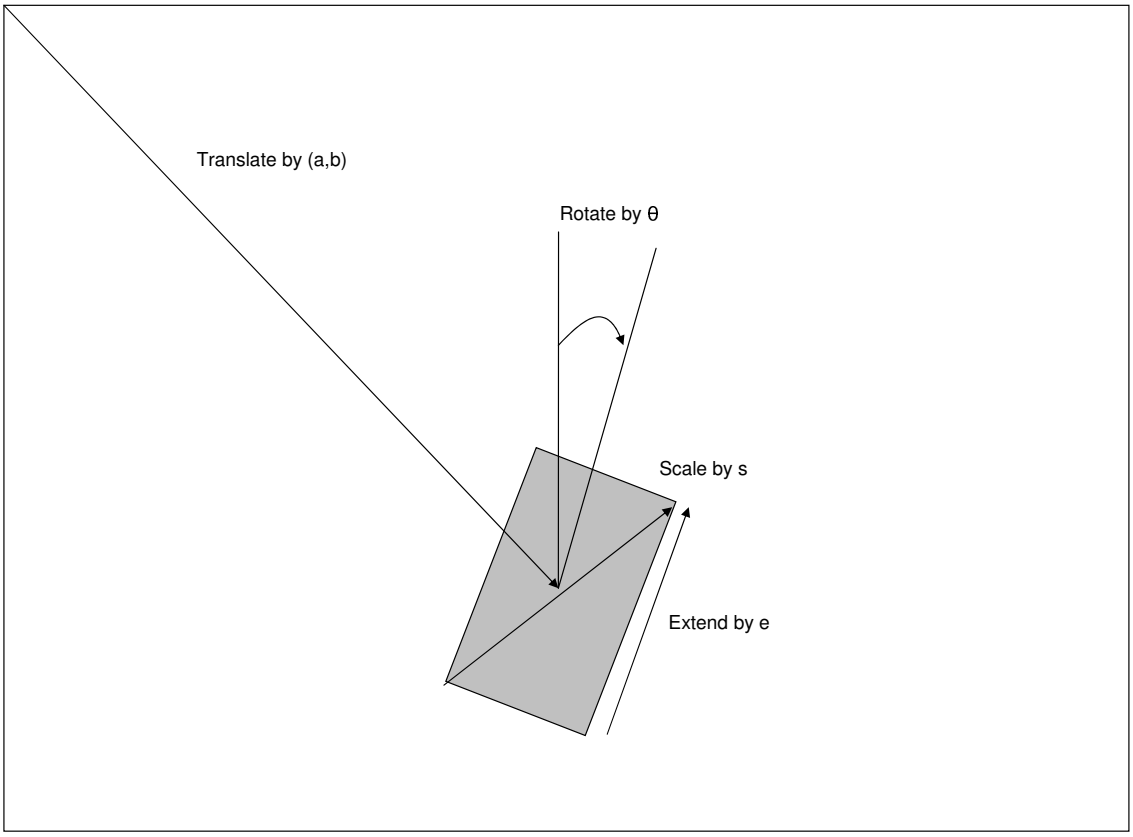

Figure 3.2: A part, shown in grey, is transformed into the image using the part pose parameters.

The transformation from the part space to image space, shown in vector format in Equation (3.6) has a straightforward interpretation. First, the probabilistic region is translated so that the centre is at position (a_i, b_i). Then, the probabilistic region is rotated by θ_i in the image plane and an extension, denoted by e_i, is applied to

model out of plane rotation. Finally, the part is scaled by a common scale factor s. Figure 3.2 illustrates this transform.

$$\vec{c_i} = \{a_i, b_i, \theta_i, e_i, s\} \tag{3.5}$$

$$T_i(x_{region}, y_{region}) = \begin{pmatrix} x_{image} \\ y_{image} \end{pmatrix} = s \begin{bmatrix} \cos\theta_i & e_i \sin\theta_i \\ -\sin\theta_i & e_i \cos\theta_i \end{bmatrix} \begin{pmatrix} x_{region} \\ y_{region} \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix} \tag{3.6}$$

3.5 Probabilistic Regions

Current part based systems use *ad hoc* geometric primitives, either 2D (e.g. rectangles, ellipses) or 3D (e.g. cylinders, tapered super-quadrics), to model body parts. Furthermore, the parameters of the primitives are usually tuned to specific individuals prior to estimation and fixed. The problems of uncertainty due to intra-person variability, clothing and non-rigid deformation are ignored. However, this limits the generality of the system when a strong prior is not available and inter-subject variability is significant (although the sensitivity to such assumptions depends upon the likelihood model employed). In contrast, the approach of *probabilistic regions* presented here encodes the shape uncertainty explicitly. In this approach no hard distinction is made between a hypothesised foreground and background.

Consider the cropped images of body parts illustrated in Figure 3.3. It can be seen that the variation in shape is, in part, due to clothing and inter-person variation. Although it would be possible to model these variations explicitly this is unnecessary

since the class of applications under consideration does not require such information. Uncertainty is also introduced by assumptions in the model of pose. For example, for 2D models, un-modelled perspective effects and 3D shape variation increase the shape uncertainty. Again, it depends on the input data (and thereby the application) as to whether parameters such as 3D position are important. Therefore, *shape model uncertainty is due to un-parameterised variation.* This is in contrast to pose uncertainty which is inherent in the problem.

Figure 3.3: Cropped images of body parts that were aligned using manually specified parameters for the 2D transform. The rows correspond to torsos, heads, upper arms, lower arms, upper legs and lower legs. The torso images are shown at a different scale to the other parts.

3.5.1 Learning the Probabilistic Region Templates

To represent shape uncertainty, the non-transformed shape of a body part, i, is represented using a non-parametric probabilistic distribution denoted by $\mathbf{Mask}_i$ and termed a probabilistic region template. Each point in this region represents the probability of that point being on the part. To estimate these probability masks, part segmentations were manually aligned using the 2D transformation presented in the last section. To do this a small application was constructed to mark up the images and thereby specify the alignment transformation. Training data was gathered from photos of people at a fixed distance from a camera with fixed lens parameters. To simplify training limbs were only extracted from parts with maximal extension (i.e. with the major axis of the part approximately parallel to the image plane) thereby reducing the number of transform parameters to 4. These were specified using two predefined, opposing boundary points, for example the elbow and shoulder for the upper arm. This transformation was also used to extract a sub image containing the untransformed part. Then the foreground was manually segmented by masking the region in an image editor. In some situations the boundary of a part is subjective (consider for example the head and neck). This issue is partly circumvented by choosing a consistent segmentation and partly by modelling the inter-part appearance similarity as described in the next chapter. Figures 3.3 and 3.4 illustrate typical cases along with example segmentations. Twenty segmentations were performed for each limb part (making a total of 160), 20 for the torso and 40 for the head. Notice that the segmentations were deliberately not re-scaled to take account of changes in the physical size of the subject in order to account for this variability.

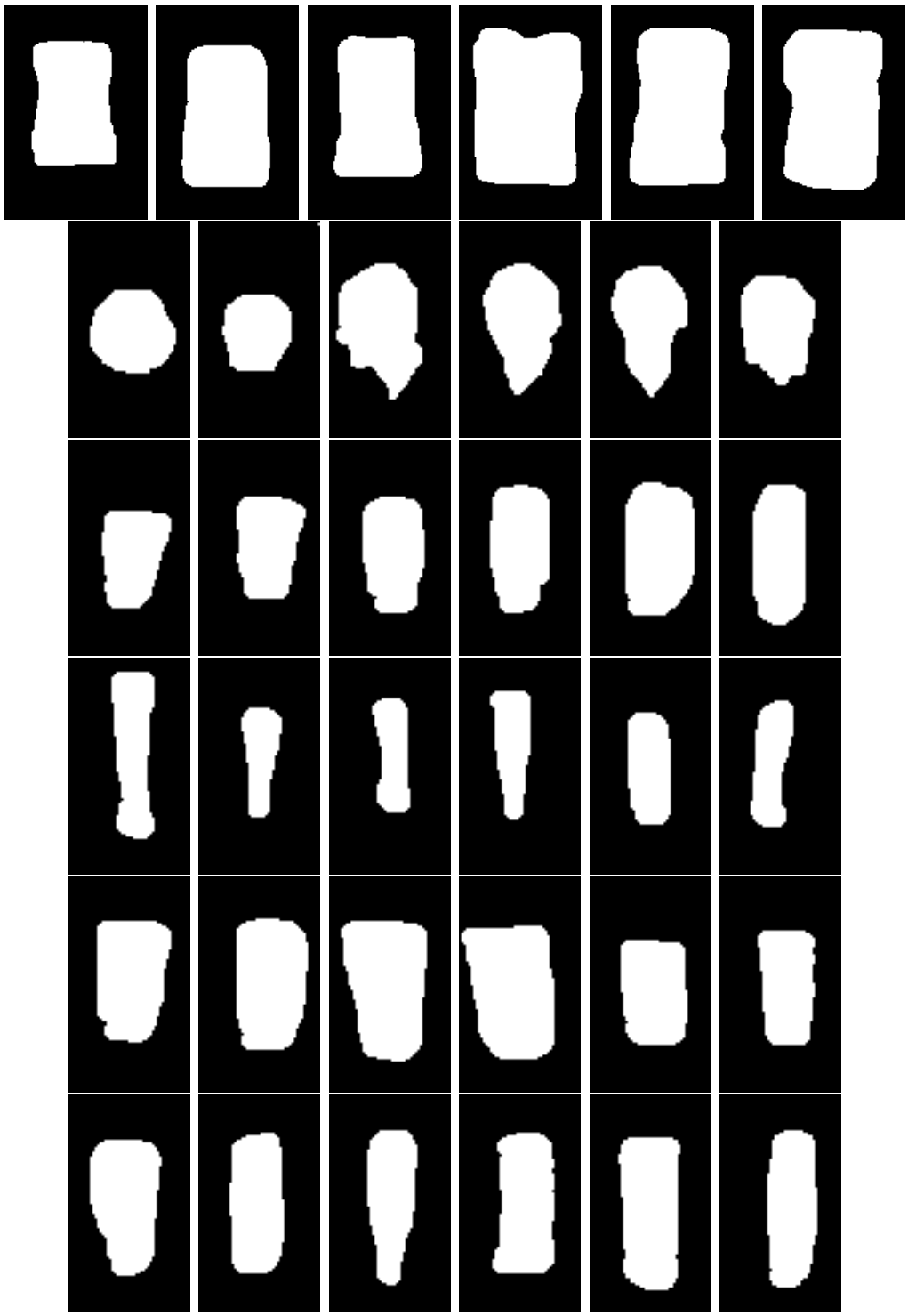

Figure 3.4: Examples of manually segmented part foreground. The rows correspond to torsos, heads, upper arms, lower arms, upper legs and lower legs.

3.5.2 Reducing the Size of the Search Space

Probabilistic regions are a principled approach to representing uncertainty in the shape model based upon the marginalisation over un-parameterised shape variation. The number of parameters with which to describe the object's shape can be varied and the effects on its shape uncertainty investigated. While it would be possible to augment the rigid transformation parameters with a set of basis regions, the mean was found to be sufficient (especially when considered in the context of the likelihood model described in Chapter 4). In order to describe other objects, especially rigid objects that have significant 3D variation, such as cars, it would be necessary to learn a basis of weights or use an explicit 3D approach.

In summary, the probabilistic region template is estimated using the frequency at each point across all segmentations and not parameterising a degree of freedom (and marginalising over it) reduces the size of the state space. This reduction in the size of the state space is key to efficient automatic initialisation. With this in mind, the rotation about each limb's major axis is not parameterised since these rotations change the shape and appearance very little. Furthermore, the limbs are constrained to be symmetric about the vertical axis thereby reducing the size of the space of rotations by two (i.e. the segmentations were flipped vertically and used for learning). The resulting probabilistic region templates are shown in Figure 3.5. Note that whilst the equations assume these probabilistic region templates have infinite extent they are illustrated, and implemented, as masks with a finite size that contains all the non-zero probabilities.

It can be seen that the resulting limb and torso regions are similar to tapered cylinders. Indeed, these masks could be approximated by such a shape with, for example, an exponential function describing the drop in foreground probability. However, in order to represent more general shapes and their uncertainty in a simple

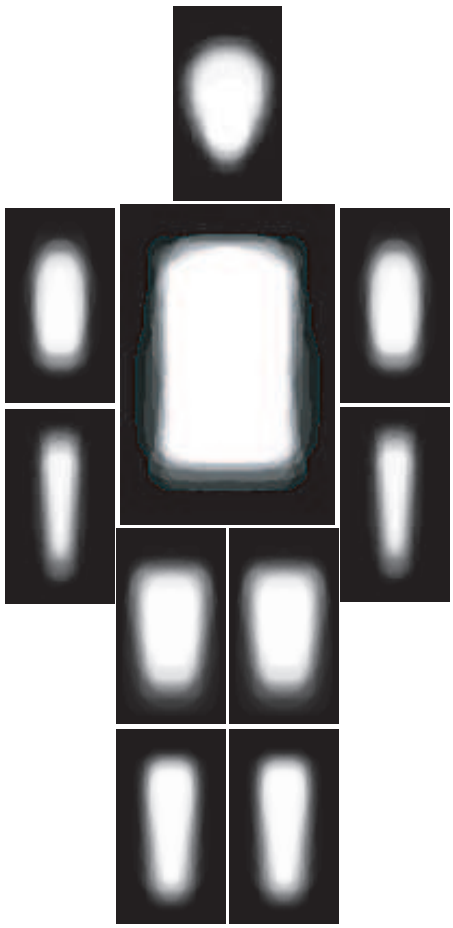

Figure 3.5: The probabilistic region templates, all at the same scale, that result from marginalising over the foreground segmentations and enforcing horizontal symmetry.

manner, the non-parametric approach described here is preferred.

3.5.3 Probabilistic Self Occlusion

In this thesis depth ordering is used to account for self occlusion (although a 3D model was used for earlier work that is presented in the Appendix). Since the depth ordering is not known prior to estimation in most cases, the ordering must be included as part of the pose hypothesis. Depth order has been applied to model based tracking previously (Rehg and Kanade [1995]).

To represent this ordering a pose hypothesis is taken to be a sorted set of part

hypotheses, with the nearest part first. Since the model of shape is probabilistic, multiple parts could be 'visible' (in the probabilistic sense) at a particular point in the image. More specifically, the probability that a part, i, is visible at image point $\mathbf{I}(x, y)$, or its *foreground* probability at that point, is determined by the inverse part transform, as shown in Equation (3.7) where j labels closer, instantiated parts.

$$p(visible|(x, y), i, \mathbf{C}) = \mathbf{Mask}_i(T_i^{-1}(x, y)) \times \prod_j (1 - \mathbf{Mask}_j(T_j^{-1}(x, y)) \tag{3.7}$$

The probability that the background is visible at a given point is given by Equation (3.8), where j indexes all instantiated parts.

$$p(\overline{visible})|(x, y), \mathbf{C}) = 1 - \sum_j p(visible|(x, y), j, \mathbf{C}) \tag{3.8}$$

These are *key equations* that will be used to form the appearance model that is used for likelihood measurements. In particular, the probability of a part being visible, or the foreground probability, at a point in the image determines a weight that is used to form a distribution that describes the appearance of the body part. The likelihood model also expands the notion of a part's background probability $(1 - p(visible))$ to the notion of a contrasting background probability that depends on which other body parts are visible at that point and how visible they are likely to be at that point. This notion of contrast is necessary in order to account for the appearance similarity of different body parts (i.e. the upper arms are often similar to the torso).

3.6 Pose Prior

In order to constrain the pose to valid configurations a prior is proposed. In this thesis this is a simple hard constraint prior based upon learning the upper and lower bounds on the relative pose of anchor points defined on pairs of body parts. This prior is learnt from approximately 150 instances of standing, walking, pointing, waving and sitting poses from various viewpoints (i.e. not always face on). In all these poses the person is upright in the image. The prior embodies scale and translational invariance. Whilst a more specific model, defined in terms of global pose, would improve discrimination, such models require more data to estimate and are not the focus of this work. The prior used does not constrain the orientation, scale or extension of the body parts.

Consider two parts, i and j with configuration $\vec{c}_i$ and $\vec{c}_j$. For each part, a set of anchor points is defined that corresponds to the position of idealised joints in the body. These anchor points are specified manually. The limb has an anchor point at each end, the head has a single anchor point at the neck and the torso has anchor points at the neck and limb joint points. Let the vector (specified in Cartesian co-ordinates) that connects the anchor points between parts i and j be $\mathbf{v}_{i,j}$. The prior probability that the pair is correct is considered to be a top hat function over the relative position of these anchor points. The parameters of the distribution, the maximum and minimum relative horizontal and vertical translations are specified from the training data. The prior over background poses is also considered to be uniform, but over the entire image. The prior probabilities of being a person, $p(person)$, or background, $p(background)$, are unimportant because only a single maximum is sought. They would become important for detection and scenes containing multiple people. The prior on the pose as a whole is formed by assuming part independence and is given in Equation (3.9). Body parts are also constrained

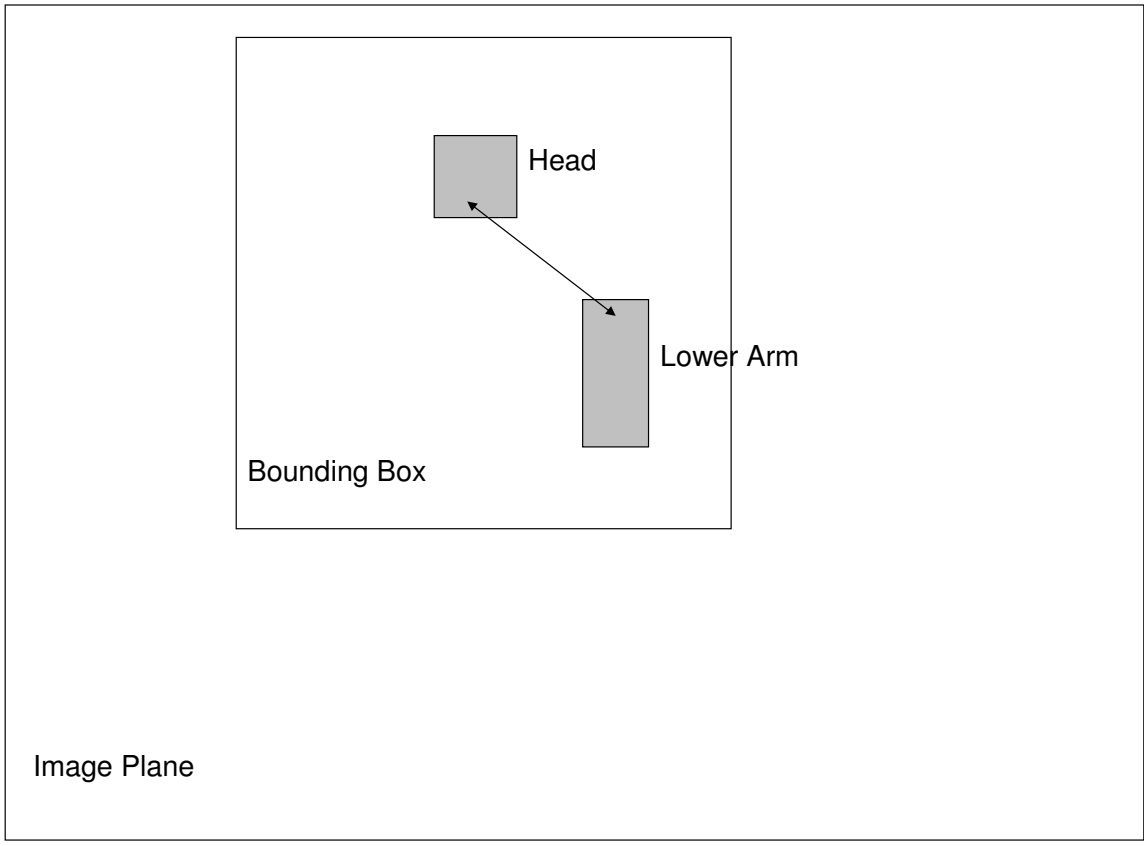

Figure 3.6: The prior specifies hard constraints on the relative position of anchor points on the projection of body parts. These constraints can be visualised as bounding boxes in the image plane. Note that all parts are connected in this manner in the non-hierarchical model described here.

to lie within the image.

$$p(\mathbf{C}, v = person) \propto \prod_{i,j:j>i} p(\mathbf{v}_{i,j}|person) \tag{3.9}$$

3.7 Summary

In order to summarise this chapter the questions that were posed in the first section are briefly revisited:

What approach is best for monocular, medium detail applications? A Bayesian,

analysis by synthesis approach, whereby pose models are hypothesised and compared to the image and prior constraints, is a principled approach that is often adopted. Within this framework a part based representation has many advantages. Moreover, for this class of application a lower dimensional 2D part model is preferred.

What are the limitations of current formulations? A significant disadvantage of existing formulations is that they either do not make use of 'bottom-up' body part identification, and are therefore highly inefficient and require manual (re)initialisation, or that they make restrictive assumptions on the relationship between parts (e.g. self occlusion) and therefore give poor results. Another significant disadvantage of most current part based approaches is that they assume no sporadic other object occlusion, loss of visibility due to poor illumination or camouflage. The partial configuration formulation was proposed as a solution to both these problems.

How can pose estimation be performed in the presence of occlusion? For 2D part models, depth ordering can be used to represent and predict self occlusion. Partial configurations can be used to cope with occlusion by other objects in the scene.

How should uncertainty in the subject's shape be represented? Probabilistic regions were proposed in order to represent the uncertainty due to shape. Uncertainty in shape, which is distinct from uncertainty in pose, is represented to increase efficiency.

What constraints exist on the pose? In a part based formulation a range of constraints exist on the solution. In the current implementation, a simple prior was proposed based upon the relative position of pairs of body parts.

Chapter 4

Likelihood

4.1 Introduction

In the previous chapter a model of human pose was developed. This chapter discusses how models of pose can be used to make measurements in the image and thereby discriminate correct from incorrect hypotheses. It is not surprising that the likelihood is critical to accurate visual pose estimation. However, due to the complexity and variation of human appearance, as discussed in Chapter 2, building a general but discriminatory likelihood model is difficult and still a topic of active research. In discussing, reviewing and developing likelihood models this chapter aims to address the following questions:

- What visual information can be used to compute a likelihood model?
- How can single body parts be better discriminated?
- How can the correct pose be discriminated from the huge number of incorrect multi-part configurations?

4.2 Related Research

Whilst learning the joint PDF of measurements conditioned upon model parameters, $p(\mathbf{I}|\mathbf{C}, v)$ is the optimal framework for likelihood models, it is impossible to reliably estimate this distribution due to the large number of parameters. Therefore, in the interests of generalisation a model must be established that encodes conditional independencies, representing, for example, invariance to position and changes to foreground appearance. Various likelihood models have been proposed and this section reviews and critiques the various approaches. Furthermore, these individual models are often fused to improve performance and interesting or important combinations are discussed. The aim is not to present an exhaustive enumeration of the implementations of different types of likelihood models. Rather the aim is to identify the main approaches and their fundamental assumptions and limitations.

The literature relating to likelihood models for detailed pose estimation is limited and can be divided into boundary and foreground models. These models are discussed in the Sections 4.2.2 and 4.2.3. The literature for human tracking likelihoods is more rich and includes, for example, optical flow and background models. Since a human tracking system was also developed as part of this thesis (and is discussed in the Appendix), tracking specific likelihoods are also briefly discussed. As there is a limited amount of research relating to likelihood models for human pose estimation and since similar problems occur in describing the appearance of other objects, references are sometimes made outside the human pose estimation literature.

4.2.1 Properties of Likelihood Models

To begin the discussion of likelihood models consideration is first given to identifying the characteristics of a good likelihood model. Although the likelihood model should

be tested as a component within the context of the entire pose estimation system, two factors can be identified as important in a likelihood model:

Discrimination Strong discrimination is necessary for good posterior classification since the number of incorrect instances is much larger than the number of correct instances. From the point of view of discrimination the ideal likelihood model would be a delta function on the correct model configuration. The discrimination of a model should be measured over a large representative set of test images and thereby implicitly account for the generalisation of the model.

Efficiency The likelihood surface is usually multi-modal and complex and therefore requires iterative sampling in order to determine the maxima. Since the search space is large and human pose estimation emphasises efficient estimation the likelihood model should allow rapid sampling and/or efficient sampling techniques (e.g. local gradient ascent).

The discrimination and efficiency of a likelihood model can be competing goals. Furthermore, there is a tradeoff between the generality of the likelihood model and it's discrimination. For example, in situations when a foreground appearance estimate is available better discrimination is possible. As discussed in Chapter 2, a key problem in human pose estimation is that limited information is available regarding the foreground and background appearance.

4.2.2 Boundary Models

Likelihood models based upon the difference in appearance of the foreground and background around the boundary of the model are popular since they are largely

invariant to changes in foreground and background appearance. Much of the literature on human boundary models relates to tracking but is applicable here. *However, since human trackers use temporal constraints, and the tracking sequences are usually short and manually initialised, strong discrimination is less important than in global pose estimation.*

Intensity Edge

Matching model configurations to the intensity edge field has a long history in computer vision, for example, in industrial vision applications. The focus on analysis of intensity is due in part to the availability of grey scale cameras in early research. Edge responses are largely invariant to illumination changes and are easy to extract and match. It is natural therefore that edge based matching be applied to human pose estimation.

Early work by Hogg [1983] used a threshold on the magnitude of the Sobel filter response to detect edges. Projected model boundary segments were then inspected to find edges within a specified distance and relative orientation. These measurements were combined using an average over the boundary segment and then each part was weighted depending on its size and visibility. In order to constrain the search pose candidates, frame differencing was also employed.

Gavrila and Davis [1996] used the results of edge detectors from multiple viewpoints to determine pose. However, instead of counting edge features within some threshold, a robust variant of the chamfer distance transform was computed from detected edges for matching. The distance transform has the advantage of providing a gradual change between the contours and a long range effect. However, edges that are far from the hypothesis, not moving nor background were removed in order to improve

the result.

Instead of matching to the results of edge detection, Wachter and Nagel [1999] convolved a model edge directly with the filter response. The edge was modelled using a 1D Gaussian with manually specified variance. Strong candidate model edges were actively selected based upon the overlap with similar parts.

The energy term of active contour models is often formulated using intensity edge measurements. A standard approach involves casting contour normals and making measurements along these 'measurement lines' (Blake and Isard [1998]). To form the likelihood either edge detection is performed along these lines (Isard and Blake [1996]) or the intensity profile is matched to a learnt profile (Cootes and Taylor [2001]). Making non-localised measurements along lines allows errors in the shape model and the detection process. However, as the length of the measurement lines increases the constraints on shape decrease since consistency between lines is not enforced. The approach of casting measurement lines can also be applied to rigid part models.

MacCormick and Blake [1998a] developed a probabilistic formulation for contour localisation using measurement lines. The resulting contour discriminant was similar to the likelihood ratio approach adopted here. However, the focus was on modelling the distribution of features on the foreground and background in contrast to the emphasis of this thesis, which is on learning these distributions. Building upon this contour discriminant formulation, MacCormick and Blake [1998b] considered removing the assumption of measurement line independence in order to deal with occlusion. The idea was that occlusion generates poor measurements that are structured. The dependence between measurement lines was encoded as a prior on the structure of the measurements around the contour and the prior was represented using a Markov random field learnt from previous occlusion instances.

A state of the art 'bottom-up' statistical edge model, that is particulary relevant to this thesis, is described in Konishi et al. [2003]. This work emphasised the importance of modelling filter responses from the non-boundary edges (e.g. texture) in order to characterise the strength of a set of edge measurements. Ground truth segmentations were used to learn the PDF of filter responses on and off object boundaries. Many filters were considered and applied at multiple scales. By estimating the joint distribution, represented non-parametrically using an adaptive histogram, the work attempted optimal fusion of the filter responses. The ratio between the two learnt PDFs, the likelihood ratio, is a non linear mapping from edge features to a measure of the edge strength. In comparison with standard model based techniques excellent results were reported and this represents the state of the art in 'bottom up' intensity edge detection.

Sidenbladh and Black [2001] took a similar approach but learnt object specific feature distributions. The PDF of intensity edge features was learnt for points around human boundaries and for points on the background. Likelihood ratios were then combined by assuming independence. In contrast to many other systems, it was also reasoned that the important edge information is contained in the orientation and scale of the edges rather than in the magnitude and therefore the image should be contrast normalised. It is interesting to note that the empirical distribution of edges does not correspond to the oft assumed Gaussian form. A conclusion of the work was that intensity edges provide a sparse cue and that a statistical model of colour and texture would improve results.

Ronfard et al. [2002] took a different approach to using the variation in intensity by learning support vector classifiers for whole parts (and one for the whole body) based upon orientation and scale specific Gaussian derivative filters (a 2016 dimensional feature vector per image location). Although the system is part based it does not explicitly parameterise occlusion, i.e. if present, the effects of occlusion must be

learnt. Surprisingly, the foreground appearance, i.e Gaussian smoothing, was also learnt. The classifiers were trained upon 100 manual part identifications. The false part detection rate (in contrast to person detection which makes use of grouping) was reported to be approximately 80% (although this does not include confusion between parts).

Due to the limitations of using edges cues in real world images, edges are often fused with other cues. For example, Wachter and Nagel [1999] used a foreground template to stabilise tracking.

Colour and Texture Boundary

Intensity edges provide a suboptimal representation of model boundaries since they do not make use of colour information. Furthermore, they do not account for texture. For the typical images of people presented in Chapter 1, texture is often apparent at scales of 2 to 20 pixels. Using intensity edges in textured scenes leads to poor boundary discrimination (large numbers of false positives and false negatives). There is, however, a limited amount of research pertaining to 'top down' colour and texture boundary models, especially for human pose estimation.

The compass operator described in Ruzon and Tomasi [1999] used the divergence between colour distributions either side of a circle's oriented bisector (hence the name compass) to find edges in colour images. Martin et al. [2004] used a similar approach but used intensity edge, colour and texture features to detect boundaries. However, in contrast to the compass operator, the parameters of the system were learnt from ground truth segmentations in a similar manner to Konishi et al. [2003] (described in the previous section). Colour was represented in the CIE-Lab space and the colour gradient was formulated using the χ^2 measure between colour histograms either side

of the boundary. Texture was represented using cluster labels, or textons (Malik et al. [2001]), based on Gaussian derivative responses and the texture gradient was formulated using the χ^2 measure between the texton distributions. This approach provided good performance for a 'bottom up' method.

Shahrokni et al. [2004] used zeroth and first order Markov processes along a measurement line to model texture and determine the most likely position of a texture boundary (assuming the line crossed the boundary). Their results on tracking rigid textured objects in cluttered scenes emphasised the importance of modelling texture and the limitations of intensity edge cues. The Markov measurement line formulation allowed fast local tracking.

4.2.3 Foreground Models

Absolute Foreground

The absolute foreground appearance provides a strong cue for pose estimation. However, in the case of human pose estimation a description of the absolute appearance is usually unavailable, primarily due to changes in clothing appearance. In tracking scenarios, foreground appearance can be assumed to be known from manual initialisation. However, the absolute appearance is susceptible to adaptation due to lighting change and clothing motion. Appearance estimation in the presence of large pose uncertainty and appearance adaptation is a key unanswered question in foreground appearance modelling. For this reason foreground cues are often fused with a 'stablising' cue.

A case when absolute appearance is known with some certainty is skin. In relation to human pose estimation, Forsyth and Fleck [1997] learnt the texture distributions

off-line in order to detect naked people in images and Ioffe and Forsyth [2001a] used template matching to find limb segments and torsos of naked people. Park et al. [2000] used a segmentation scheme and then applied a colour classifier to detect skin coloured body parts. Although skin colour provides a strong cue, and has been used in many successful realtime systems, it is not used since it is sensitive to illumination and subject identity, places restrictions on pose (i.e. that the skin is visible) and is often only applicable to the head and hands.

Template matching is an established technique for localisation. Cham and Rehg [1999] used a probabilistic formulation of template matching on intensity to localise body parts for tracking. In particular, each pixel on the model surface was assumed to be normally distributed with mean found from the manual initialisation in the first frame and a global, manually specified variance. The total likelihood was formed by assuming independence between pixels. No method was presented for adapting the appearance information. In contrast Wachter and Nagel [1999] used the difference between the points on the hypothesis and the points on the previous maximum (but also relied upon manual initialisation). Measurements were combined using a sum of squared differences. Neither of these approaches is robust to outliers and the assumed Gaussian noise model may not be appropriate in the case of significant relative clothing motion and foreground texture.

In contrast to the template matching approach, where spatial information is used, histogram matching is based upon marginalisation over spatial extent. This approach is fast, eases foreground appearance estimation and adds a degree of robustness to relative clothing motion. An example of such a technique is Comaniciu et al. [2000] in which the mean shift, a local search technique, was employed to maximise the Bhattacharyya divergence between the hypothesised and estimated colour histograms. The technique demonstrates good performance in the presence of background clutter and significant un-modelled partial occlusions and operates

in realtime. However, these advantages come at the cost of reduced discrimination and localisation. In situations where absolute appearance information is available and this appearance is non-uniform it is likely that a matching method that makes use of spatial information whilst being robust to movement of the textured surface would discriminate and localise best.

In order to account for such complex variations in appearance Sidenbladh et al. [2000b] proposed learning a linear subspace representation of the surface texture for a particular subject from a set of views (linear subspace models have been widely applied in face modelling, see for example Turk and Pentland [1991]). Ground truth for a 3D cylindrical shape model was provided by a motion capture system which was in turn used to project the image onto the model surface. Since each view provides only a portion of the 3D limb model's surface the principal components algorithm was modified so that regions on the surface are weighted by visibility. The likelihood is then formed based upon the distance between the learnt model and the hypothesised appearance in this linear subspace. Notice that this model allows rotation about the limb's major axis to be recovered if the limb's surface has distinguishing non-symmetric features such as an emblem. A disadvantage of this approach is that the whole appearance must be learnt off-line and the result cannot be adapted between different targets and may be complex to adapt online for a specific target.

Ramanan and Forsyth [2003] used motion, appearance consistency and kinematic constraints to find a colour representation of the foreground appearance of individual limbs automatically before tracking. The subject was then tracked efficiently using the mean shift algorithm.

Absolute foreground appearance models are also of interest for maintaining subject identity across sequences. For example, McKenna et al. [2000] and Haritaoglu

et al. [1999b] used colour and texture information to track specific people through occlusions and group interactions.

Foreground Structure

Foreground structure models are those models that use the relationship between foreground features. In relation to human pose estimation and tracking, only a limited amount of research has made use of such relations.

A clear example of the discriminatory power of foreground structure is provided by the similarity templates as discussed in Stauffer and Grimson [2001]. A similarity template is the concatenation of all the relationships between pairs of pixels in an image window. These similarity templates have been used for pedestrian detection and provided good discrimination. However, it relied upon limited variation in pose and was slow to evaluate.

The Pfinder system (Wren et al. [1997]) used clusters in colour-position space to find the head and hands. This approach allows a real time implementation which reported good generalisation. However, foreground regions were first segmented using a background model and the recovered pose was coarse.

Sidenbladh and Black [2001] learnt the distribution of ridge features at the scale of the body part for correct and incorrect configurations. A ridge is a point that has strong change in one direction and little change perpendicular to that direction. Ridges occur because of the curvature of the body parts. Ridge features were formulated in terms of second derivative filters of intensity at the scale of the body part in a similar manner to Lindeberg [1998]. It is not clear how dependent ridge features are on illumination.

4.2.4 Other Models

In this section, likelihoods models that are applicable to (some) human tracking scenarios, but not human pose estimation, are discussed briefly.

Optical Flow

Low level optical flow fields can be used to estimate the parameterised motion of the human body and provide a strong likelihood cue. Using an optical flow field has many of the benefits of the foreground appearance modelling techniques, but since it uses frames that are close temporally, it is less sensitive to illumination changes. A general survey of motion based recognition systems was performed by Cedras and Shah [1995].

Both dense and sparse optical flow fields have been used for human tracking and motion analysis. An early example of the use of dense optical flow for human tracking is provided by Ju et al. [1996] where the motion of a 2D 'cardboard person' model was recovered. However, the system was applied in a constrained environment to track parallel lower leg motions and did not provide a technique for dealing with self occlusion. Pentland and Horowitz [1991a] used optical flow to recover a 3D model. Sidenbladh et al. [2000a] used optical flow in a Bayesian framework with explicit occlusion handling to track a 3D model. Sminchisescu and Triggs [2001] combined a correlation based optical flow field with edge features by weighting their proximity to the motion boundary. Song et al. [2000] used the position and flow of (sparse) Lucas-Kanade feature points to detect and label human motion in the presence of simple background movement. Fablet [2002] proposed using the eigenvectors of the full body motion field to detect human motion and initialise a more detailed model.

Many optical flow systems implicitly assume that only the target is moving and it is

unclear how successful they would be in more complex environments. Furthermore, optical flow based likelihoods can suffer from accumulation of errors resulting in the model drifting from the subject. Finally, due to the complexity of the clothing surface the brightness constancy assumption may be inappropriate.

Background Models

In scenes where the background can be estimated it is possible to segment the subject. These models usually require that the background be static or adapting slowly. Therefore applying such techniques to moving camera sequences is difficult. Much research has been done regarding background models, some of the research relating to human tracking includes:

For the realtime W4 surveillance system (Haritaoglu et al. [2000]) a background model was employed to quickly segment multiple subjects. The background model consists of per pixel minimum and maximum intensity and maximum inter-frame change. Pixels are first determined that differ from either the minimum or maximum by more that the maximum inter-frame change. Then thresholding, morphology and connected component operations are applied to remove noise and find foreground regions of significant size. High level grouping is used to determine if pixels should be adapted. Further processing was performed to recover body parts. This system was extended in later work to take into account multiple people (Haritaoglu et al. [1999b]) and people carrying objects (Haritaoglu et al. [1999a]). McKenna et al. [2000] described a more principled probabilistic background model. In particular, the distribution of colour and colour gradient at each pixel was modelled using a Gaussian distribution with the colour channels assumed to be uncorrelated. A pixel was classified as foreground if it differed from the mean by more than three standard derivations (which by Chebyshev's theorem, includes at least 88.9% of the

data regardless of the distribution). Shadows, which are typically a problem with background models, were handled by making decisions based upon chromaticity and texture information. Zhao et al. [2001] extended this idea by predicting shadows from scene knowledge. Mikic et al. [2001] used such background models from multiple calibrated viewpoints to construct a voxel (3D) representation of the subject.

Since the segmented silhouette resulting from background models gives limited pose information it is often combined with other cues. For example, Deutscher et al. [2000] and Sminchisescu [2002] used the background model to weight edge measurements.

Depth

In situations where depth information is available it can be used in the likelihood model. For example, Darrell et al. [2000] proposed using depth to segment subjects in cluttered environments and then used colour and texture models to identify the head and hands. Okada et al. [2000] integrated optical flow and depth information to estimate the 3D pose of the subject.

4.3 Spatial Likelihood

Although several likelihood models have been proposed, it is the author's opinion that more emphasis needs be placed on likelihood models in order to achieve accurate, efficient human pose estimation from real world images. In particular, the high level shape information has not been used to best effect in formulating boundary and foreground structure models:

Boundary Models Part boundary likelihood models in most current pose estima-

tion and human tracking systems are based upon bottom up boundary models that do not make use of the high level shape information (e.g. Deutscher et al. [2000], Sidenbladh and Black [2001]). The performance of these localised bottom up boundary models on finding boundaries in whole real world images was significantly lower than human performance (Martin et al. [2004]). This is not surprising given that texture is an inherently non-localised property that can occur over large scales and that the high human performance depends upon having large regions either side of the boundary. Bottom up approaches are unable to account for differences in large scale textures that are common in images of people and therefore lead to poorer discrimination and lower efficiency. Moreover, using texture on measurement lines (such as Shahrokni et al. [2004]) assumes that the texture can be described by this line, which can be violated, e.g. when the surface undergoes rigid deformation.

Foreground Models Whilst it has been shown that the relationships between pairs of foreground features can be used to reliably detect pedestrians (Stauffer and Grimson [2001]), this information has not been used for detailed, part based pose estimation. In particular, the relationship between body parts which usually have a similar appearance, has not been used to enhance discrimination. In this system the similarity between the appearance of opposing body parts is used to improve discrimination of larger configurations and thereby constrain the estimation.

In this section a model is developed that provides a common approach to boundary modelling and foreground structure modelling based upon the divergence between regions in the image formed by the high level shape information.

4.3.1 Part Boundary Model

Approach

The approach taken here to discriminating single body parts is to use the divergence between the appearance of the foreground, as induced by the high level shape model, and its adjacent background. These appearances will be dissimilar as long as a part is not completely camouflaged. This is illustrated in Figure 4.1.

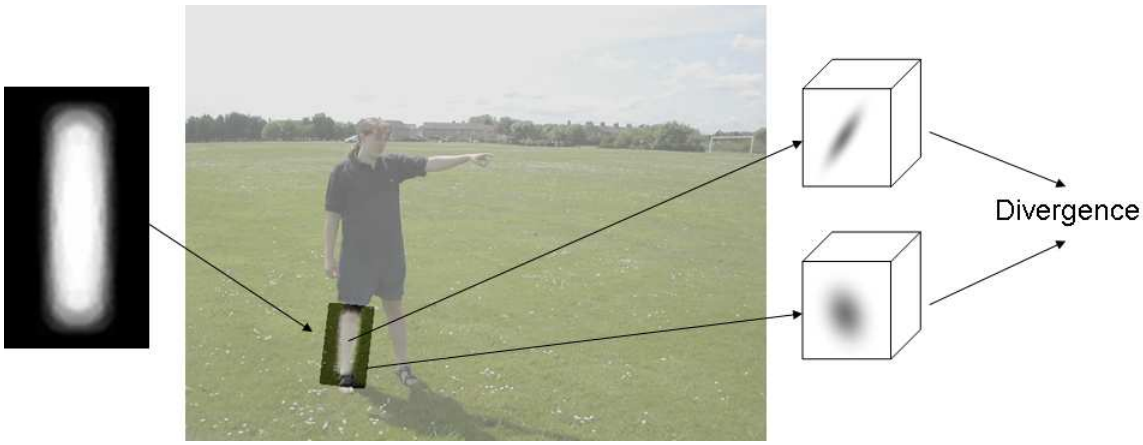

Figure 4.1: The probabilistic region template for the lower leg is transformed into the image. The probabilistic region is used to estimate the foreground and adjacent background appearance distributions. The likelihood model is formed based upon the divergence between these distributions.

Divergence between feature distributions in a region has been used for bottom up detection (with small compass regions as discussed above) and texture patch discrimination (e.g. Puzicha et al. [1999]). The model proposed here is the natural application of this approach to high level shape models.

It may aid the reader to interpret this as a model based segmentation scheme and contrast this approach to both the boundary detection approach and the region homogeneity approach. The difficulties in applying boundary detection to real world images were discussed above. Segmentation by region homogeneity is also not generally applicable since homogenous regions do not correspond to the body part

structure (parts can be non-homogenous and homogenous regions can cross part boundaries). The approach here is based on the assumption that the appearance of the region as a whole is different from the appearance of the adjoining background.

In order to demonstrate the improvements in discrimination and efficiency that this approach affords it will be compared with an intensity edge model.

Foreground Appearance

A generic foreground appearance model is adopted here based upon marginalising features over the foreground region, weighted by their visibility, $p(visible|(x, y), i, \mathbf{C})$, as given by Equation (3.7). This appearance model, makes the assumption of limited foreground structure and is more applicable to clothed parts, that are characterised by large regions of unknown, uniform texture than other objects, such as faces, that can be described by local feature vectors (Fergus et al. [2003], Schiele and Crowley [2000]). In human pose estimation the image regions that can be related occur at a larger scale (i.e. at the size of body parts, instead of localised features) and the relative pose of these regions is highly varied (in contrast to facial features). Furthermore, this approach is most useful for discriminating highly deformable, textured objects and would be less useful in localising rigid man made objects such as cars.

Improving Efficiency By Combining Part Models

Some individual limb segments have a similar shape and therefore give similar responses to boundary models. This is especially true of divergence based likelihood models that are less sensitive to the boundary position. In order to improve the efficiency when initially identifying limb candidates (which, as described in the next

Chapter, comprises a significant amount of the search time) the small differences between individual limb segment shapes can be marginalised over without much loss in boundary discrimination. Therefore, in this system the lower arms and lower legs are represented using a single limb segment probabilistic mask template.

Background Appearance

In order to identify parts based upon a difference in appearance between a part's foreground and adjacent background it is necessary to form a background appearance with which to compare. *A key difficulty in modelling the boundary of body parts is that different parts often have similar appearance and are either neighbouring or overlapping.* For example, points on the upper arm are often similar to points on the torso and these parts are adjoining and often neighbouring and/or overlapping. It is for this reason (in addition to more frequent occlusion) that it is more difficult to discriminate torsos 'bottom up' than outer limb parts, such as the lower arms. In order to discriminate overlapping and adjoining parts a model for the notion of a contrasting background is proposed that is different to the model of non-part membership, $1 - p(visible|(x, y), i, \mathbf{C})$, as defined by Equation (3.7). Let the probability that a point is contrasting to the foreground be $p(contrast|(x, y), i, \mathbf{C})$. This point-wise model of contrast can then be used to form the contrasting background appearance distribution and compared to the foreground appearance to form the single part boundary likelihood term.

In the absence of other parts the contrasting background appearance distribution is extracted from a region of approximately equal area (in the probabilistic sense) to the foreground region and with equal probability (weight). This choice is supported by the fact that the discrimination (as determined by the induced likelihood ratio described below) is weaker when larger and smaller regions are used (although better

weighting schemes could exist). One might expect there to be a tradeoff between shape specificity and obtaining a good estimate of the contrasting region distribution. In particular, this region is formed for each part by finding the Euclidean distance d such that all points with $p(visible) = 0$ that are within a distance d of a point with $p(visible) > 0$, in addition to those with $p(visible) < 1$, give an equal (probabilistic) area to the foreground. For example, in the case of the lower arm this is all those points in the mask within 3 pixels of a foreground point. Example contrasting regions are shown in Figure 4.2.

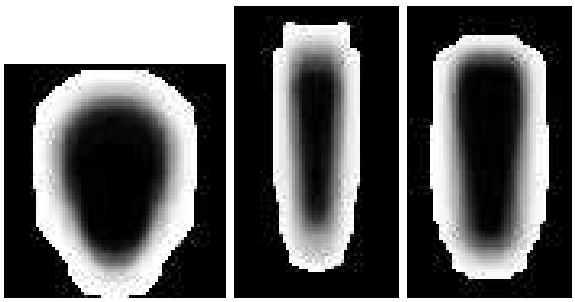

Figure 4.2: The contrasting background probabilistic region templates for the head, lower arm and lower leg.

In the presence of other parts it is necessary to consider a non-uniform weighting of points within the adjoining background region. However, the nature of contrasting part boundaries is complex and subtle. Firstly, because of the 3D shape of body parts, points close to the boundary are often shadowed and therefore an appearance difference exists even when the part is overlapping a similar part. Secondly, if orientation specific texture features are used, similar parts with different orientations, and that have oriented texture, can be discriminated. Currently a straightforward model of contrast is employed based upon the expected contrast between whole parts, $p(cp|i, j)$. For example, the head and torso are assumed to be highly contrasting whereas the upper arm and torso are usually not highly contrasting. This is similar to the approach proposed by Wachter and Nagel [1999] for contrasting edges. It is important to understand that using more discriminatory features, such as texture, will result in higher expected contrast between parts. The parameters

for this model are specified manually. In particular, the probability of contrast between adjoining limb segments is set to 0.1 and between other combinations is set to 0.5. The probability that a point is contrasting to part i is then modelled as Equation (4.1), where j labels all instantiated parts and $p(\overline{visible}|(x,y),\mathbf{C})$, as discussed in Section 3.5.3, is the visibility of the background. Learning a more principled model of inter-part contrast is deferred to future work.

$$p(contrast|(x,y),i,\mathbf{C}) = p(\overline{visible}|(x,y),\mathbf{C}) + \sum_j (p(cp|i,j) \times p(visible|(x,y),j,\mathbf{C})) \tag{4.1}$$

Finally, to complete the model of a contrasting region, consideration must be given to the evaluation of partial configurations (i.e. poses that do not describe all parts). In this case the *expected* position of the missing body parts can be used to obtain a contrasting region. For example, when detecting single body parts, the performance can be improved by distinguishing positions where the background appearance is most likely to differ from the foreground appearance. In the current system this non-adjoining region is specified manually during training by identifying regions that are most likely to be adjoining with a weight equal to that defined above. For example, a region at the top of the lower arm where it usually joins the upper arm is identified and the weight is set to 0.1 (the weight for adjoining parts). A more principled approach, where the adjoining regions are estimated using the expected position of missing parts given the current parts might be better. However, calculating the expected position of missing parts online would be computationally expensive and it is important that the detection of small configurations be efficient (low complexity). It would also require a better representation of the pose prior than has been developed in this work. It is important to note that this is only important, and thus used for, better bottom-up identification of body parts. When the

adjoining part is specified using a multiple part configuration, the standard model of contrast described in the paragraph above is employed.

Learning Region Divergence

The appearance distributions are represented using joint intensity-chromaticity histograms (3D distributions). A histogram representation is used since texture can often result in multi-modal distributions and histograms are fast to compute. These histograms have $8 \times 8 \times 8$ bins. For scenes in which the body parts appear small, semi-parametric density estimation methods such as Gaussian mixture models would be more appropriate. In general, local filter responses could also be used to represent the appearance e.g. Schiele and Crowley [2000].

Let the foreground appearance histogram for part i be denoted by F_i and the background appearance histogram by B_i. There are many approaches that could be used to compare the two appearance histograms including: χ^2 (which estimates the likelihood of one distribution being drawn from the other), the Kullback Leibler (KL) divergence (which relates to the mutual information of the distributions), the Jeffrey distance (which is a symmetric version of the KL divergence), the Bhattacharyya measure and the Minkowski metric. Alternatives have been proposed specifically for comparing histograms including histogram intersection, the quadratic form and the earth movers distance (the latter two being global measures of histogram similarity). Puzicha et al. [1999] provided a comparison of several distribution similarity measures in the context of texture region comparison. Based on its success in colour based tracking (Comaniciu et al. [2000]) this system uses the Bhattacharyya measure, given by Equation (4.2), to compare appearance distributions. The Bhattacharyya measure is related to the Bayes (Duda et al. [2001]).

$$DIV(F_i, B_i) = \sum_{\mathbf{f}} \sqrt{F_i(\mathbf{f}) \times B_i(\mathbf{f})} \tag{4.2}$$

To discriminate parts from the background the Bhattacharyya divergence between the foreground and background appearances is learnt for correct ($v = person$) and incorrect ($v = background$) configurations in a supervised fashion. In particular, a $v = person$ distribution was estimated from data obtained by manually specifying the transformation parameters to align the probabilistic region template to be on parts that are neither occluded nor overlapping. The $v = background$ distribution, which encodes the likelihood of observing a part shaped object in the class of scenes under consideration, was estimated by generating random alignments elsewhere in 100 images of outdoor and indoor scenes and smoothed. Equation (4.3) defines the border divergence ratio, BDR_i, as the ratio of these two distributions for a part, i.

$$BDR_i = \frac{p(DIV(F_i, B_i)|v = person)}{p(DIV(F_i, B_i)|v = background)} \tag{4.3}$$

In order to obtain a smooth log likelihood function and interpolate/extrapolate the learnt data a parametric function was fit to the data. In particular, rather than fitting a parametric function directly to both PDFs the ratio was first formed and a parametric function fit to this single result. It was found that the Boltzmann sigmoid function, with a functional form given in Equation 4.4 was a good fit (with r-value, or correlation coefficient, as determined by XLFit, of 0.96). This is the function used to 'score' a single body part configuration and is plotted in Figure 4.3. This learnt sigmoid function acts as a soft classifier for body parts based upon the divergence measure.

$$S(x) = a + \frac{b - a}{1 + e^{\frac{c-x}{d}}} \tag{4.4}$$

The log likelihood ratio is central to identifying body part candidates and subsequent pose estimation. Any hypothesis that results in a histogram divergence with log likelihood above zero is more *likely* (i.e. not taking into account the prior) to be a body part than not a body part.

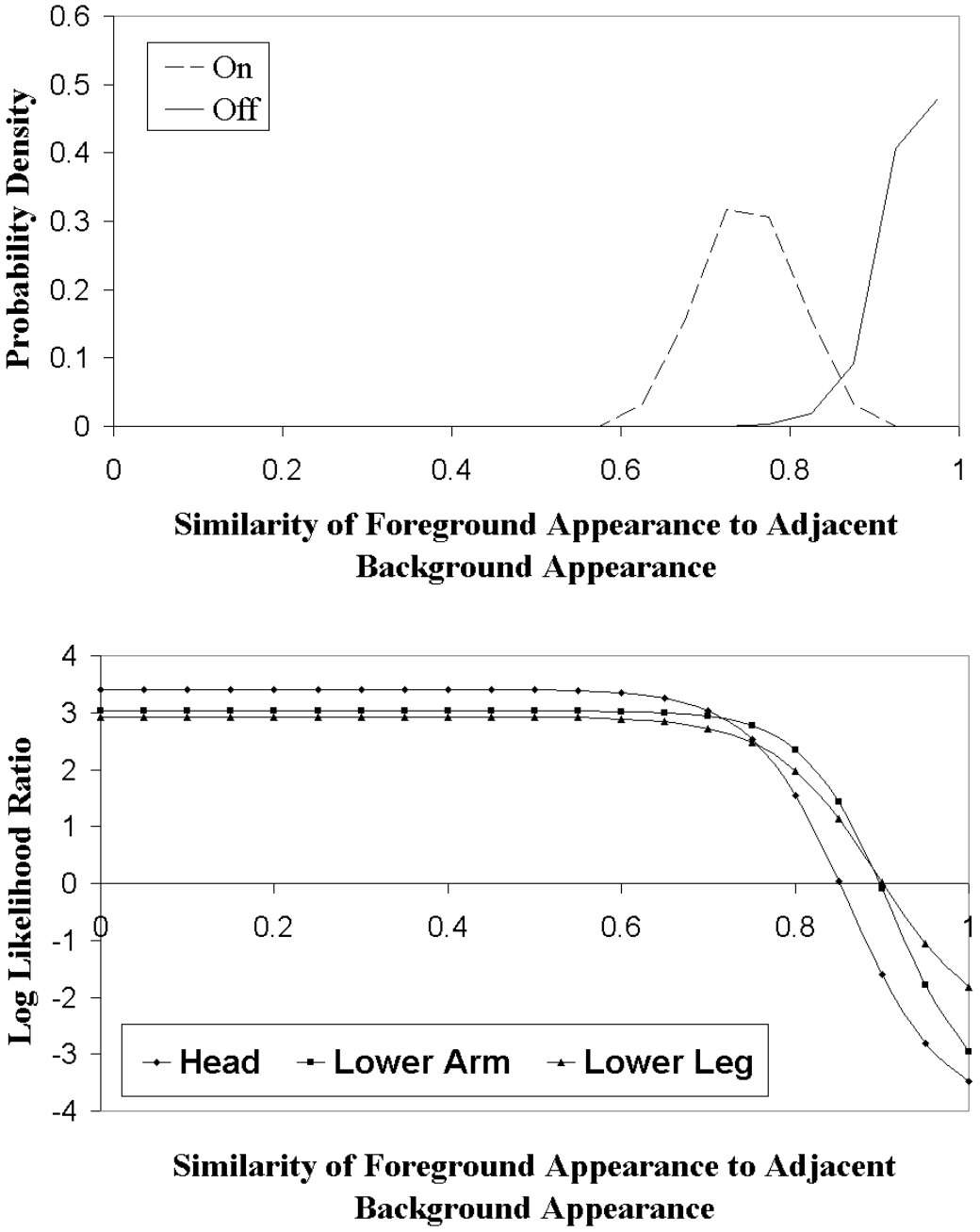

Figure 4.3: Top: A plot of the learnt PDFs of foreground to background appearance similarity for the $v = person$ and $v = background$ part configurations of a head template. Bottom: A plot of a Boltzmann sigmoid function fit to the log of the likelihood ratio data for head, lower arm and lower leg parts. It can be seen that the distributions are well separated.

Intensity Edge Model

In order to judge the effectiveness of the divergence based boundary model an intensity edge model was implemented (for single part hypotheses only). Intensity edge magnitude and orientation is computed using 3×3 Sobel filters. In order to form the single part edge likelihood for a part i the expected position and orientation of edges must be determined. Since the derivative is a linear operator the mean spatial gradient can be found by taking the spatial gradient of the mean (i.e. the probabilistic region). Using the derivative of the probabilistic region in this way provides a more principled approach than edge detection and making assumptions about the form of the model boundary, e.g. Gaussian with fixed variance as in Wachter and Nagel [1999]. The edge response can now be formed by convolving the derivative of the probabilistic mask with the image.

In a similar way to contrasting boundary responses, the magnitude of edge responses is spatially structured. Without a model of this structure, measurements in regions with weak expected contrast, for example around the joints, would be treated in the same way as regions with high expected contrast, thus reducing discrimination. For simplicity, and since this investigation is for single parts only, regions of large expected contrast are manually identified in a similar manner to current systems (e.g. Sidenbladh and Black [2001]).

Using the body part training data, object specific responses can now be learned in a supervised fashion (i.e. manually align the parts to determine foreground responses and randomly sample to determine background responses). It is important to note that correct responses are learnt even when the response is weak since to do otherwise would greatly penalise correct poses that have a weak response. Rather than learning both magnitude and orientation, the component of the edge in the direction of the model edge is learnt. Notice that this approach does not use contrast normalisation

or a multi-scale approach as expounded in Sidenbladh and Black [2001] and this may partly explain its poor performance. The response for the part as a whole is computed by assuming independence of the individual measurements.

Investigation

Figures 4.4 to 4.6 show the projection of the log likelihood ratio computed using Equation (4.3) onto typical images containing significant clutter. The first image shows the response for a head while the other two images show the response to a vertically-oriented limb filter. It can be seen that, in comparison to the intensity edge model, the new method is highly discriminatory and produces relatively few false maxima.

Figure 4.7 illustrates the typical spatial variations of both the body part likelihood response proposed here and the edge-based likelihood. The edge response, whilst indicative of the correct position, has significant false, positive likelihood ratios.

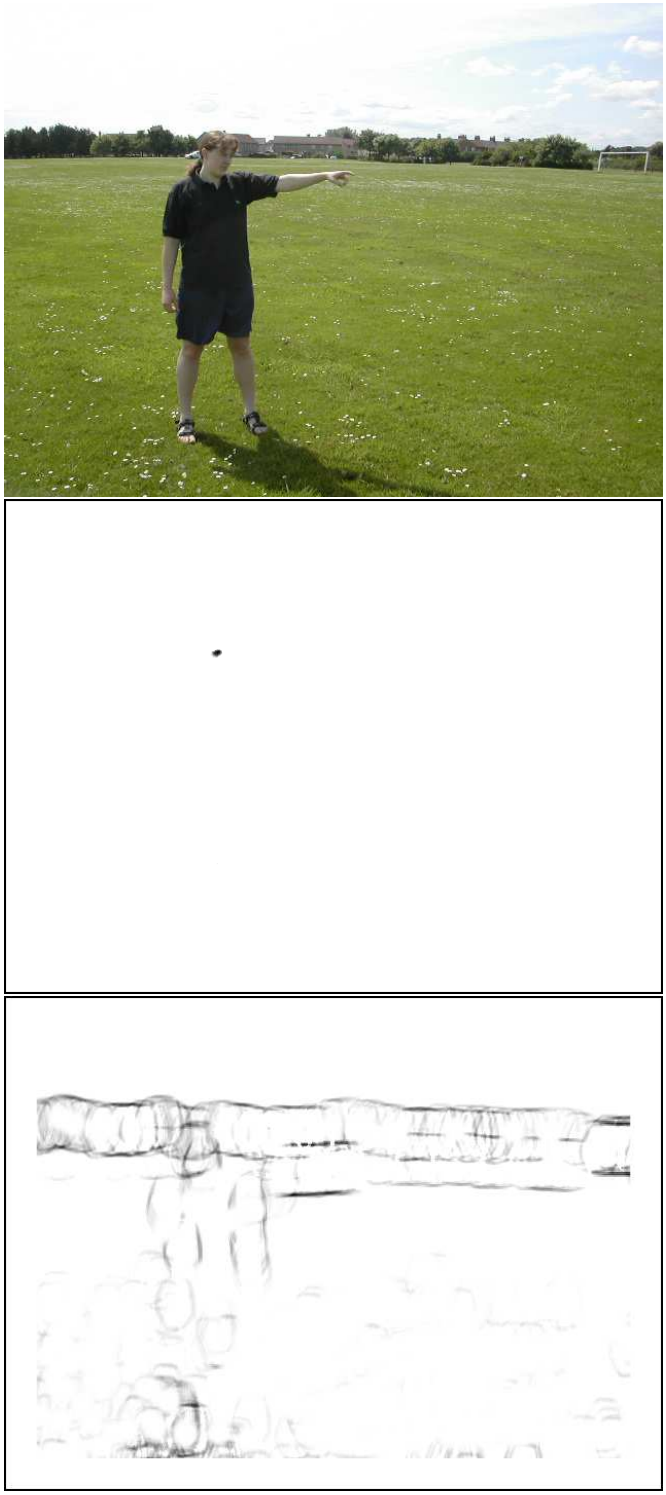

Figure 4.4: An image of an outdoor scene along with the projections of the log likelihood (positive only, re-scaled) for a head part filter: first for the colour divergence model developed here and then for the intensity edge model.

Figure 4.5: An image from a challenging outdoor scene along with projections of the log likelihood for a vertically oriented limb. Notice the large response of the edge based model to the sail masts. This is typical for an intensity edge based model. Also notice the false response in between the legs for the model presented here, the space between the legs is itself shaped like a leg.

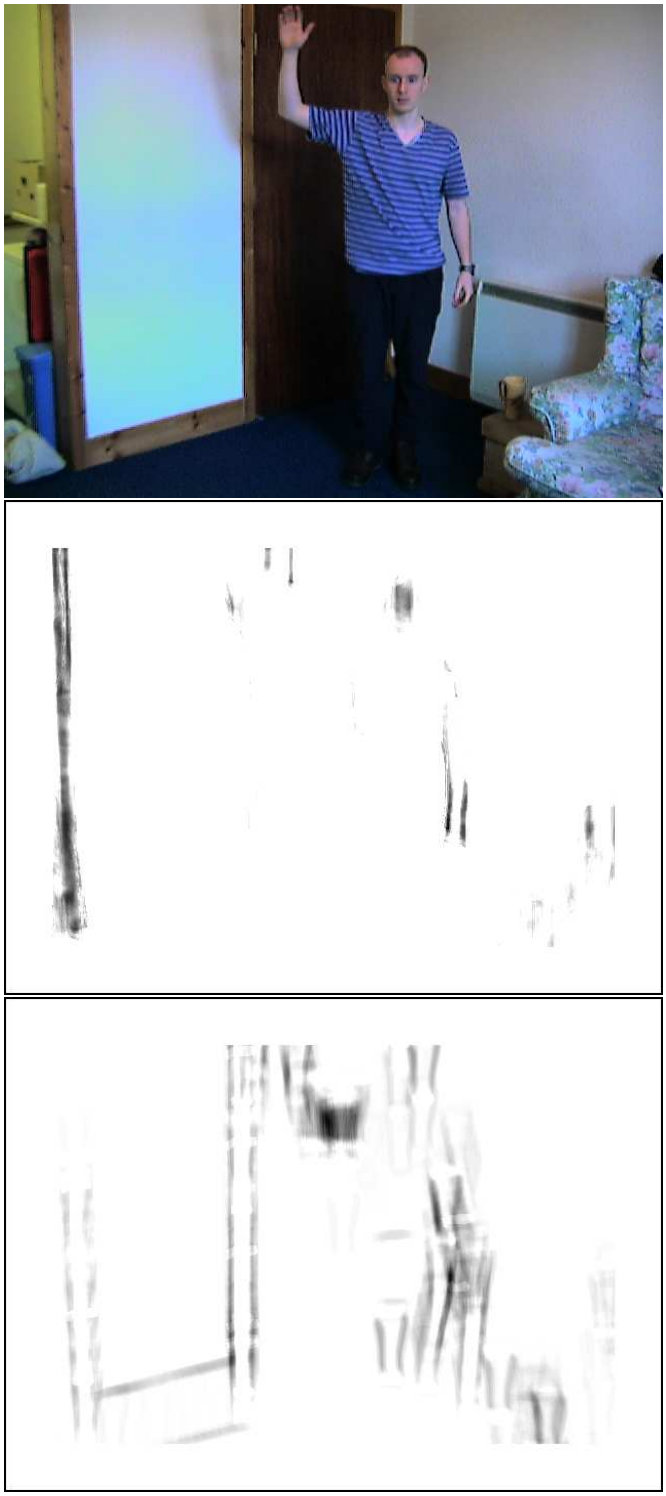

Figure 4.6: An image from a cluttered indoor scene along with projections of the log likelihood for a vertically oriented limb. Notice the strong likelihood response from the door frame. Also notice, in contrast to the edge model, that the head gives a strong response (in relation to the correct arm) for the model proposed here.

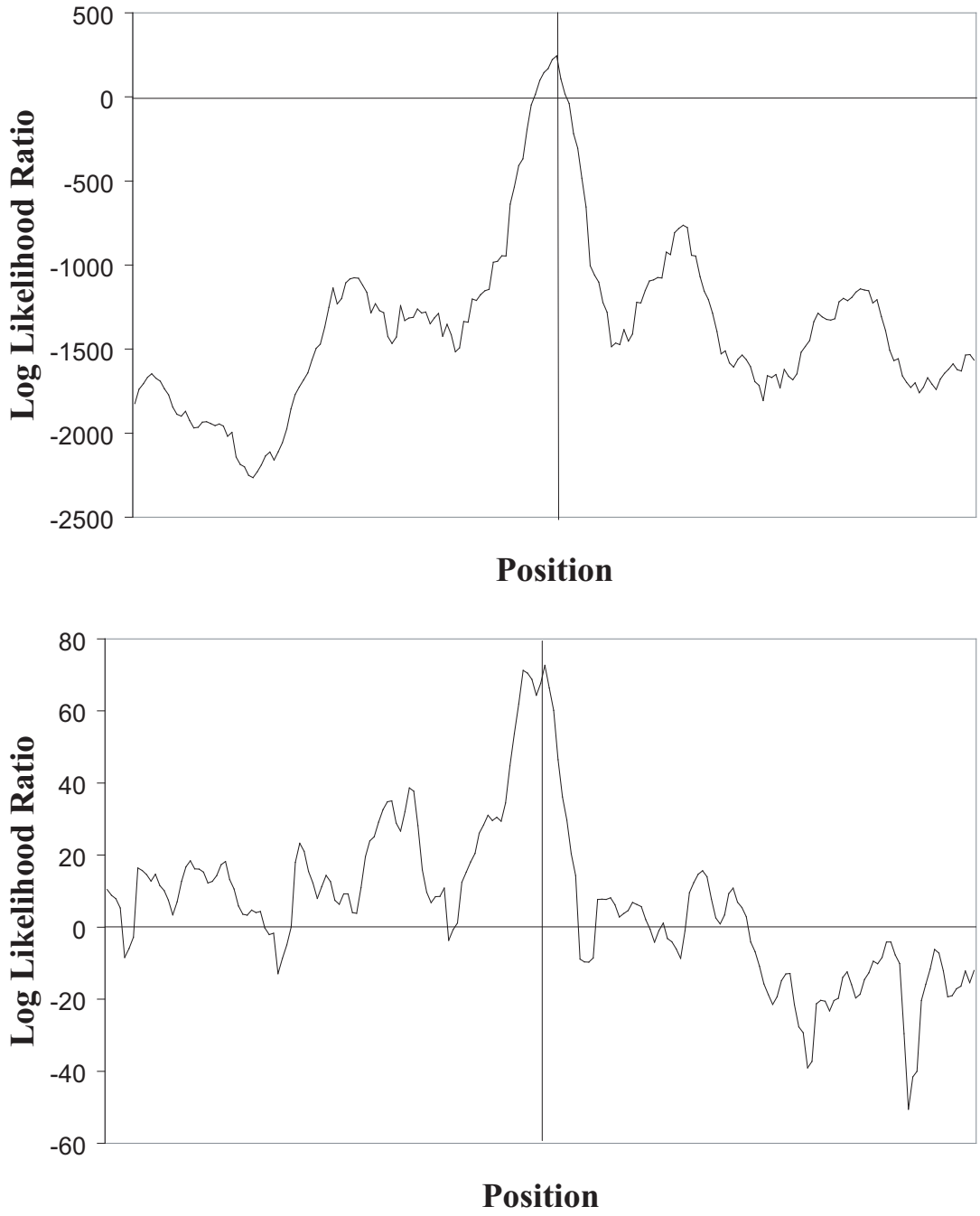

Figure 4.7: Comparison of the spatial variation (plotted for a horizontal change of 200 pixels) of the learnt log likelihood ratios for the model proposed here (top) and the edge-based model (bottom) of the head in Figure 4.4. The correct position is centered and indicated by the vertical bar. Anything above the horizontal bar, corresponding to a likelihood ratio of 1, is more likely to be a head than not.

Although the proposed part likelihood is more expensive to compute than the edge-based filter (approximately an order of magnitude slower in the current implementation) it is far more discriminatory and as a result, fewer samples are needed when performing pose search, leading to an overall performance benefit. Furthermore, it is necessary to estimate the foreground appearance in order to compute inter-part similarity as discussed in the next section.

4.3.2 Inter-Part Model

Since any single body part likelihood will usually result in many false positives it is important to encode higher order relationships *between* body parts to improve discrimination. In addition to the spatial constraints between parts it can be seen, from the typical images of people presented in Chapter 1, that certain pairs of body parts have a similar foreground appearance. For example, a person's upper left arm will nearly always have a similar colour and texture to the upper right arm. Long range structure provides a mechanism for discriminating *large* correct configurations from *large* incorrect configurations and thereby 'pruning' incorrect hypotheses.

Approach

In this section a model of inter-part similarity is developed that encodes the long range similarity using the divergence between pairs of parts. In particular, the divergence between opposing pairs of limbs is learnt (e.g. upper left arm and upper right arm). Since rotation about a limb's major axis is not parameterised (since it cannot usually be accurately recovered) and clothing can move relative to the part's surface, the texture at a point on two limbs can be very different. Matching texture features is further complicated by the rotation of texture features on the surface of

the limb relative to the image plane. Therefore, in the same way as the boundary model presented above, colour histograms are compared.

Learning the Divergence

To learn the similarity of the appearance of opposing part pairs for correct configurations the model was manually aligned on part pairs (supervised learning). In particular, 20 pairs of upper and lower arms and legs were used. It is assumed that the distribution of similarity is the same for all pairs. To learn the similarity for incorrect configurations random unaligned hypotheses of part pairs were hypothesised with the inter-part separation drawn from the prior described in Section 3.6. It is therefore assumed that this distribution is unchanged if one of the parts in the pair is correctly aligned (and therefore more likely to be homogenous).

To encode this knowledge, a PDF of the divergence measure (computed using Equation (4.2)) between the foreground appearance histograms of paired parts and non-paired parts is learnt. Equation (4.5) is the inter-part log likelihood ratio, IDR, that results from these two distributions. For parts that are not paired this ratio is set to 1.

$$IDR_{i,j} = \frac{p(DIV(F_i, F_j)|v = person)}{p(DIV(F_i, F_j)|v = background)} \tag{4.5}$$

Figure 4.8 shows plots of two Gaussian distributions fitted to the recovered correct and incorrect responses along with the resulting likelihood ratio. It can be seen this model strongly penalises opposing part pairs that are not similar.

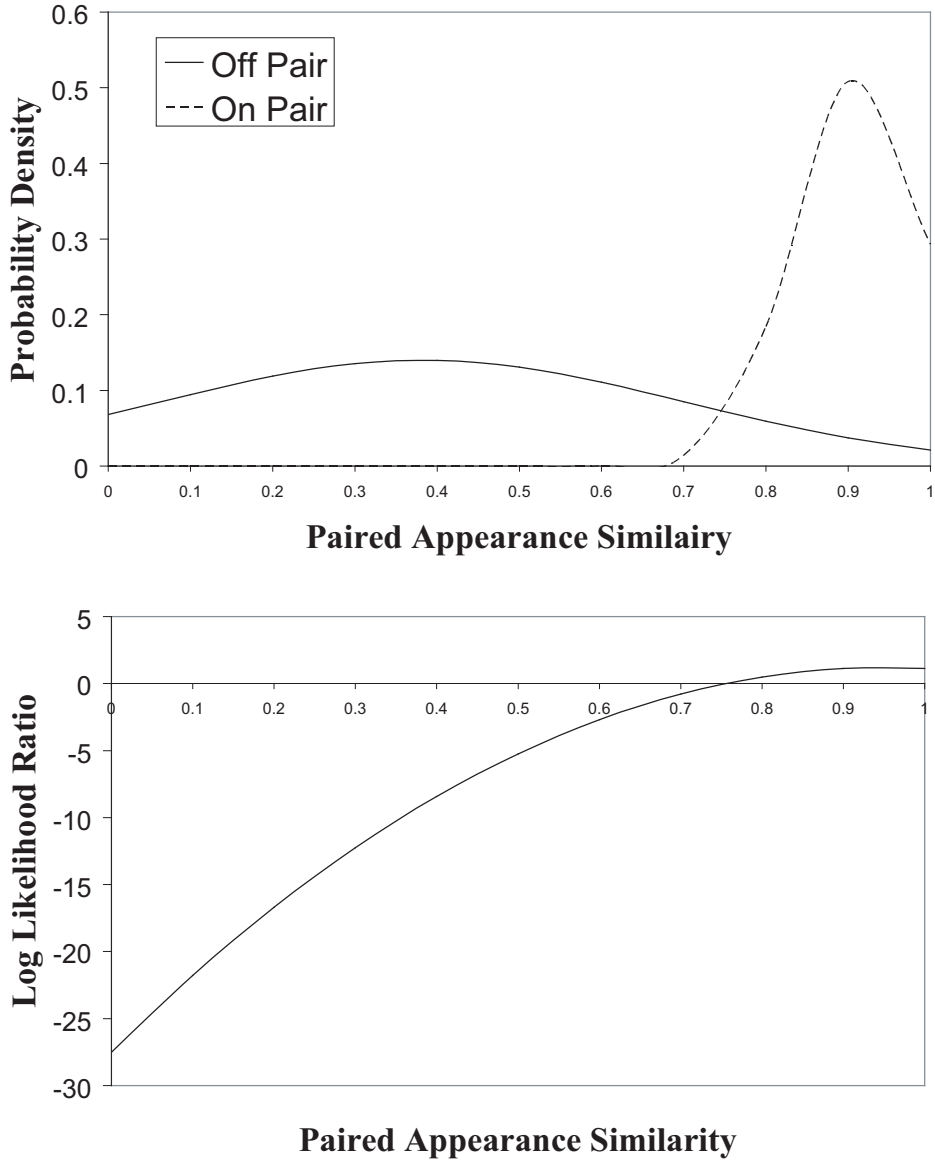

Figure 4.8: Top: A plot of the learnt PDFs of foreground appearance similarity for paired and non-paired configurations. Bottom: The log of the resulting likelihood ratio. It can be seen, as would be expected, that more similar regions are more likely to be a pair.

Investigation

Figure 4.9 shows the projection of this likelihood ratio onto a typical image and shows the technique to be highly discriminatory. This cue limits the possible configurations if one limb can be found reliably and helps reduce the likelihood of incorrect large assemblies. Furthermore, by allowing larger incorrect assemblies to be pruned, a deterministic combinatorial search is more feasible.

Figure 4.9: Investigation of a paired part response. Top: an image for which significant limb candidates are found in the background. Bottom: the projection of the log likelihood ratio for the paired response to the person's lower right leg in the image.

4.3.3 Combining the Models

The individual likelihood ratios are combined by assuming independence. The overall likelihood ratio, (LR), is then given by Equation (4.6) where the first term is a product over all parameterised parts, V, in the partial configuration and the second term is a product over all pairs of parts (recall the inter-part ratio, IDR, is set to 1 for non-opposing parts).

$$LR(\mathbf{I}|\mathbf{C}) = \prod_{i \in V} BDR_i \times \prod_{i,j \in V} IDR_{i,j} \tag{4.6}$$

4.4 Temporal Likelihood

The subject of the early investigations performed for this thesis was full body 3D human tracking. Indeed, it was partly the limitations of this work and trackers in general, in particular the requirement of manual (re)initialisation, that led to the focus on single image pose estimation. As part of this earlier work a novel likelihood model was developed based upon foreground appearance estimation and is included in Appendix A. The work is also briefly summarised here with particular emphasis on how it relates to the model described above.

The likelihood model proposed in Appendix A uses an estimate of the foreground appearance in order to discriminate the person from the background. The central problem faced by such models is reliable tracking in the face of appearance adaptation and foreground uncertainty. Therefore, the proposed model uses the regular structure of clothing to group points on the body to aid estimation. However, even with grouping and appearance adaptation, the rapid and complex changes in appearance, that occur in many sequences, can cause the tracker to diverge. Moreover,

it was difficult to solve this problem efficiently by propagating multiple appearance models (in a manner similar to Condensation or multiple hypothesis tracking).

More recently the author has adopted a different view of how to incorporate an estimated foreground appearance. In particular, in situations such as human tracking where the foreground appearance estimate is highly uncertain due to adaptation, this appearance should be used instead to improve the efficiency of a more reliable method, for example by importance sampling.

4.5 Summary

In order to summarise this chapter the questions posed at the start are briefly re-answered:

What visual information can be used to compute the likelihood? Likelihoods for single image pose estimation systems have been proposed based upon both boundary differences and foreground structure. A more diverse set of likelihoods is available for human tracking including background models, low level motion and absolute foreground appearance.

How can single body parts be better discriminated? Using the divergence between the appearance distribution of the foreground region and its adjacent background allows better discrimination of single parts (provided a model of background contrast is employed). Since the divergence approach is insensitive to the boundary position the small differences between limb segments can be marginalised over to reduce the number of samples needed to find individual limb candidates.

How can the correct pose be discriminated (from the huge number of incorrect multi-part configurations?) The inter-part appearance relationship is a highly discriminatory cue that in addition to prior constraints allows larger correct configurations to be discriminated from incorrect ones. This is key to efficient pose estimation in real world images.

Chapter 5

Estimation

5.1 Introduction

In the previous two Chapters a state space was proposed to describe human pose in images and a likelihood model was developed to efficiently discriminate correct from incorrect pose configurations. With this in place, pose estimation now consists of finding those pose configurations that are more likely to correspond to people than background. Due to the complex, multi-modal nature of the likelihood surface and the high dimensionality of the space an efficient, iterative search is needed. This chapter aims to answer the following questions:

- What approaches to estimation have been proposed?
- What are the limitations of these approaches?
- How can these limitations be overcome?

5.2 Related Research

Two broad approaches can be identified in the literature for pose estimation and human tracking: the combinatorial approach and the state search approach.

An important characteristic that differentiates estimation schemes is the representation of the solution(s). Most pose estimation systems (e.g. Ioffe and Forsyth [2001a]) and some human tracking systems (e.g. Wachter and Nagel [1999]) represent the solution using a single pose and an estimate of the local uncertainty/error. In contrast, most human tracking systems represent the solution using multiple hypotheses in order to make tracking more reliable in the presence of large uncertainty. For example, Cham and Rehg [1999] take a semi-parametric approach by recovering the local maxima and describing the density using piecewise Gaussians. Many other systems (e.g. Deutscher et al. [2001], Sminchisescu and Triggs [2001]) use a non-parametric representation of the pose distribution, also known as a particle set (Isard and Blake [1996]). The question of how to use/visualise the particle set is often neglected.

5.2.1 Combinatorial Approach

The combinatorial approach consists of finding candidate body parts and then grouping these to construct the best global fit as determined by inter-part constraints. This is the approach taken by most single image pose estimations systems.

A good general example of the combinatorial approach is pictorial structures (Felzenszwalb and Huttenlocher [2000]). In this approach pose estimation is formulated as a global energy minimisation problem with energy comprised of a per part term and an interaction term. By assuming that the interactions between parts can be

modelled as a tree, the problem can be solved in time linear with the number of assemblies using dynamic programming. An advantage of their particular approach is its probabilistic MAP interpretation. In addition to its application to general articulated object recognition and estimation the approach was applied to human pose estimation in indoor scenes.

In relation to human pose estimation, the combinatorial approach has been most actively developed by Forsyth *et al.* In early work that aimed to build better content-based image retrieval systems, Forsyth and Fleck [1997, 1999] described a highly view invariant model for detecting naked humans and animals based upon hierarchical grouping schemes called body plans. A body plan is a sequential classification approach to articulated pose estimation. It is based upon cylindrical algebraic decomposition in which a decision surface in high dimensions (e.g. corresponding to a whole human) can be projected to multiple lower dimensional spaces (e.g. corresponding to single parts and pairs) to improve efficiency. For example, first individual parts are detected, resulting in many false positives. The results of these part detections are fed into a pairwise classifier, which is in turn fed into a three part classifier, with each stage further pruning the candidates. It is important to note that the topology of the classification network is fixed prior to learning the individual classifiers (which for example were learnt from 38 horses).

A sequential classification approach was also described in Ioffe and Forsyth [2001b] but was criticised for its reliance on binary classification of limbs and the inability to measure the likelihood of the configuration (i.e. a pose was hard classified as either a person or not). A new approach was then proposed based upon drawing assemblies proportional to a likelihood of the full (fixed size) assembly. This likelihood was formulated in terms of a set of independent features. In order to efficiently draw samples from this likelihood, a set of marginal likelihoods was proposed. The marginal likelihoods are assumed to be independent of other parts and as the au-

thors point out this was an important limitation. Assemblies were then built in a fixed order (torso, upper limbs, lower limbs) by re-sampling the marginal likelihoods. Due to inter-part constraints, such as the requirement of being distinct, this model no longer has a tree structure and cannot be inferred using dynamic programming. Counting the modes of the likelihood was also used to count the number of people present.

Ioffe and Forsyth [2001a] points out that a single tree structure is not suitable for representing humans due to significant and frequent self occlusion. In order to address the problem of efficient inference under such conditions, a mixture of trees approach was developed to describe the posterior ratio. This approach allowed detection, localisation and tracking in time proportional to $O(M^2)$, where M is the size of the candidate sample sizes. The resulting system was used to automatically initialise and track subjects in the Muybridge sequences (Muybridge [1989]).

Most recently, these ideas were used to construct an automatically initialising human tracking system (Ramanan and Forsyth [2003]). The system first used motion, appearance consistency and kinematic constraints to find the appearance of limb candidates. This appearance was then used to efficiently track the subject using the mean shift algorithm. The system gave impressive results and tracked through other object occlusions.

5.2.2 State Search Approach

The state search approach makes samples in the full high dimensional space that fully models inter-part relations such as self-occlusion. This is the most popular approach for human tracking where a temporal prior is available. However, the approach usually requires manual initialisation and re-initialisation upon error to

be computationally feasible.

Many state space search techniques have been proposed. The majority of which are for localised sampling when a strong prior is available and are therefore only considered briefly here.

Gavrila and Davis [1996] described a hierarchical best first search approach to find a single best pose estimation. Whilst this greatly reduces the size of the search it gives problems when occlusion is strong or the temporal prior is poor.

Wachter and Nagel [1999] employed an iterative extended Kalman filter to incorporate the results of local Newton-Raphson optimization. The Kalman filter (Kalman [1960]) is a straightforward, principled method for temporal filtering and the assumptions on which it relies, namely process linearity and a Gaussian posterior and noise, and its derivation are presented in Maybank [1979]. Segawa et al. [2000] used the constraints of the articulated model to improve the Kalman filter technique. The Kalman filtering approach is problematic however when the distribution becomes multi-modal (e.g. due to clutter) and when singularities occur (Deutscher et al. [1999], Morris and Rehg [1998]).

In such situations, which often occur in human tracking, a particle set representation is more applicable. The most popular particle filtering technique is the well known Condensation algorithm (Isard and Blake [1996]), in which samples are drawn from the temporal prior, diffused and resampled. The problem with this technique, when applied to human tracking, is that it requires a large number of samples and is therefore inefficient. Many techniques have been proposed to increase efficiency. For example, Cham and Rehg [1999] recovered the modes of the distribution using local Gauss-Newton search and then approximated the distribution using piecewise Gaussians.

Deutscher et al. [2000] used ideas from simulated annealing to modify the Condensation algorithm to more reliably estimate the global structure of the posterior distribution. The standard single re-sampling step was replaced by an annealing run where particles were 'drawn' through a series of smoothed weighting functions and diffused by different amounts at each stage. The approach was demonstrated to work well but is still computationally expensive. Deutscher et al. [2001] extended the work further by 'automatically soft partitioning' the search space. In essence this involved adding less jitter to parameters with smaller variance. A genetic algorithm style particle cross-over operation was also developed which had the effect of searching the state space in parallel. Together these techniques took advantage of the (semi) hierarchical nature of the problem and produced a significant reduction in the required sample size.

In highly related work, Choo and Fleet [2001] expounded the advantages of using the hybrid Markov Chain Monte Carlo (MCMC) filter to recover high dimensional density functions. In this approach, the local shape of the posterior distribution defines a potential energy field which can be used to assign particles a momentum without biasing the sampling behaviour. In addition to the inclusion of this 'local optimisation' scheme the work used a multiple Markov chain approach that made exploration of multiple modes more efficient.

The work of Sminchisescu *et al* has focussed on developing novel sampling schemes that explicitly model the characteristics of the distributions found in monocular human tracking. In the covariance scaled sampling approach for example, a set of modes is used to compute the directions of largest uncertainty and samples are made from a Gaussian with a covariance matrix inflated in these directions. The authors reasoned that for monocular tracking it is in these directions, along the valley of the distribution, that alternative minima are more likely to be found. Another technique, called kinematic jumping (Sminchisescu and Triggs [2003]), was

proposed that 'flips' the orientation of parts in order to escape the local maxima that can occur in monocular estimation. These local maxima are related to those poses that cannot be reliably discriminated even when joint positions are known (Taylor [2000]).

5.3 Approach

The key approach of this work is that by improving the formulation and likelihood model the estimation problem can be eased. In particular, by better discriminating individual parts and using the relation between parts to prune larger incorrect configurations only a simple estimation scheme (in comparison to the other approaches discussed above) is necessary to perform pose estimation. The purpose of this Chapter is not to develop novel search schemes but rather to demonstrate the effectiveness of the new likelihood formulation.

The partial configuration formulation presented in Chapter 2 is particularly important from the point of view of estimation. In particular, it allows bottom up sampling to focus the search (making global sampling and thereby automatic (re)initialisation possible) whilst still allowing self occlusion and inter-part similarity to be modelled for strong discrimination. Since the structure of the model does not allow exact inference (due to inter-part relations) techniques such as dynamic programming cannot be applied. Instead an iterative combinatorial search with local optimisation is employed. This approach, although less efficient than methods such as pictorial structures (Ronfard et al. [2002]) and mixtures of trees (Ioffe and Forsyth [2001a]), is more flexible and is made feasible because of the strong likelihood model developed.

5.3.1 Assumptions

Recall that it is assumed that only a single subject is present in the image (and that one is always present). The aim of the sampler can then be treated as one of global maximisation rather than density estimation. It was also assumed, as is common with many other single image pose estimation systems, that the scale parameter of the subject is known. In practice the scale parameter is manually specified prior to estimation using the head model as a guide. Since the system parameterises scale it is easy to incorporate this information and unnecessary to account for changes in scale at the image level for example by forming an image pyramid. It is estimated that the current system would function over a range of scales of 0.8 to 5 (In order for the likelihood to be reliable the histograms must be well formed which puts a lower bound on the scale). The elongation parameter of all limbs was constrained to be above 0.7.

It is emphasised that the most important body parts, in terms of information content and for human computer interface control, for example, are the outer limbs and the head. Furthermore, the torso and upper limbs are usually harder to identify due to a lack of contrast with neighbouring regions. Therefore, the current search scheme aims to identify only the head and outer limbs (i.e. lower arms and lower legs). This assumption also allows simplification of the search scheme since the labelling of parts becomes simpler.

5.3.2 Samplers

Coarse Sampling

The first stage of the search scheme is coarse sampling, i.e. uniform tessellation, of the parameter space of single part configurations. In particular, the translation parameters were sampled at intervals of 3 pixels (in a 640×480 image). The part rotation parameter was sampled at 4 orientations, (Recall that only orientations between 0° and 180° are meaningful since the part is horizontally symmetric). All those part configurations with log likelihood ratio greater than a threshold T are stored for later grouping. Unfortunately, large amounts of clutter can cause excessively long runtimes (due to combinatorial explosion in the number of larger configurations) which necessitates the need for a higher log likelihood threshold. The log likelihood threshold is specified manually and set as low as possible, with a minimum of 0 (i.e. all the parts that are more likely to be correct than not), so that pose estimation is completed within 2 hours. The scale and log likelihood threshold are the only parameters which were specified between runs.

When coarse sampling, a 'guess' is made at the part label (e.g. lower leg) based upon the division of the image into quadrants. Getting the labelling correct at this stage is not crucial since re-labelling is included as part of the later optimisation. At this stage the head and outer limbs are often identified (if they are un-occluded and not camouflaged) along with false positives due to background clutter.

Local Optimisation

The results of the coarse sampling are then locally optimised. Even though the divergence based likelihood model was found to provide a smooth profile it has been

found that a local search which does not rely upon the evaluation of local derivatives performed better than one which does. This could be due to escaping local minima or the evaluation being more efficient in high dimensions.

Local search was performed by perturbing the pose for a particular part by a uniform random amount and accepting the proposal if the posterior ratio was higher. This could be improved upon by using a Metropolis Hastings (MCMC) proposal scheme. The translation was perturbed by up to 2 pixels in each direction, the orientation by up to approximately 12 degrees and the elongation by a factor of up to 10%. To account for self occlusion, depth ordering is also searched by moving the part up or down 1 layer (Recall that self occlusion is modelled using depth ordering). It is not possible to combinatorially search over depth ordering. However, in this case this is not crucial since the contrasting, outer parts do not usually overlap.

Combinatorial Search

The results of coarse scanning and local search are then combined to find the best overall pose. In this system a search is performed by sequentially building up larger and larger configurations. The search begins by evaluating all possible pairs of parts. Then from all the parts present in the set of valid pairs (with likelihood ratio $> T$) the set of triples is formed. This proceeds to a maximum of 5 parts.

Each level of the search proceeds as follows. First, the parts in the configuration are labelled. Since a maximum of 5 parts are currently considered labelling is greatly eased (in comparison to configurations where upper and lower limb segments are possible). To perform the labelling a part is chosen as an anchor (either the head or if this is not present the upper left part) and the other parts are labelled in terms of this part. For example, the left most part with a vertical distance from the head

of at most $s * 120$ pixels is labelled the left lower arm. Next, the configuration is checked against the hard prior constraints described in Section 3.6 to determine if it is valid (i.e. whether the anchor points on each part lie within the bounds). If the configuration is valid the log likelihood ratio is then evaluated (i.e. the boundary and inter-part appearance models are evaluated).

At each stage of grouping, inter-part relations on configuration (the prior) and appearance (inter-part similarity) reduce the number of possible parts to consider. The search is elitist in that the best configuration is kept irrespective of whether it passes later grouping stages. The best configuration is locally optimised for 200 iterations prior to output, including the global scale factor which is perturbed by up to 0.05 (but constrained to be within 0.1 of the original estimate).

Clearly, this combinatorial feed forward approach is limited and should be extended in future work. However, the five part combinatorial search will produce configurations which contain much of the pose information and from which a non-feedforward search could begin. This is not considered in this system but is discussed in the final chapter when considering future work.

5.4 Results and Discussion

The system was implemented using an efficient, in-house C++ framework. The histograms are built efficiently by projecting scan-line segments and iteratively computing the mask co-ordinates inside these segments. In addition, the colours in the image are preprocessed into histogram bins. The system samples single part configurations at the scale shown in Figure 4.1 at approximately $3KHz$ from an image with resolution 640×480 on a 2GHz PC. The runtime for the complete pose estimation system ranges from 2 minutes, when limited part candidates are identified,

to 2 hours, when many part candidates are identified.

Figures 5.1 to 5.20 show input images (top left) with results of the search (top right) as well as single part likelihood projections for the head (bottom left) and lower limbs (bottom right). The log likelihood projections are provided in order to give insight into the nature of the clutter. The projections show positive log likelihood ratios re-scaled to the range $(0, 255)$. The limb projections show the *maximum* over all orientations (sampled to 0.6 radians and locally optimised). The analysis of individual results is provided in the Figure captions.

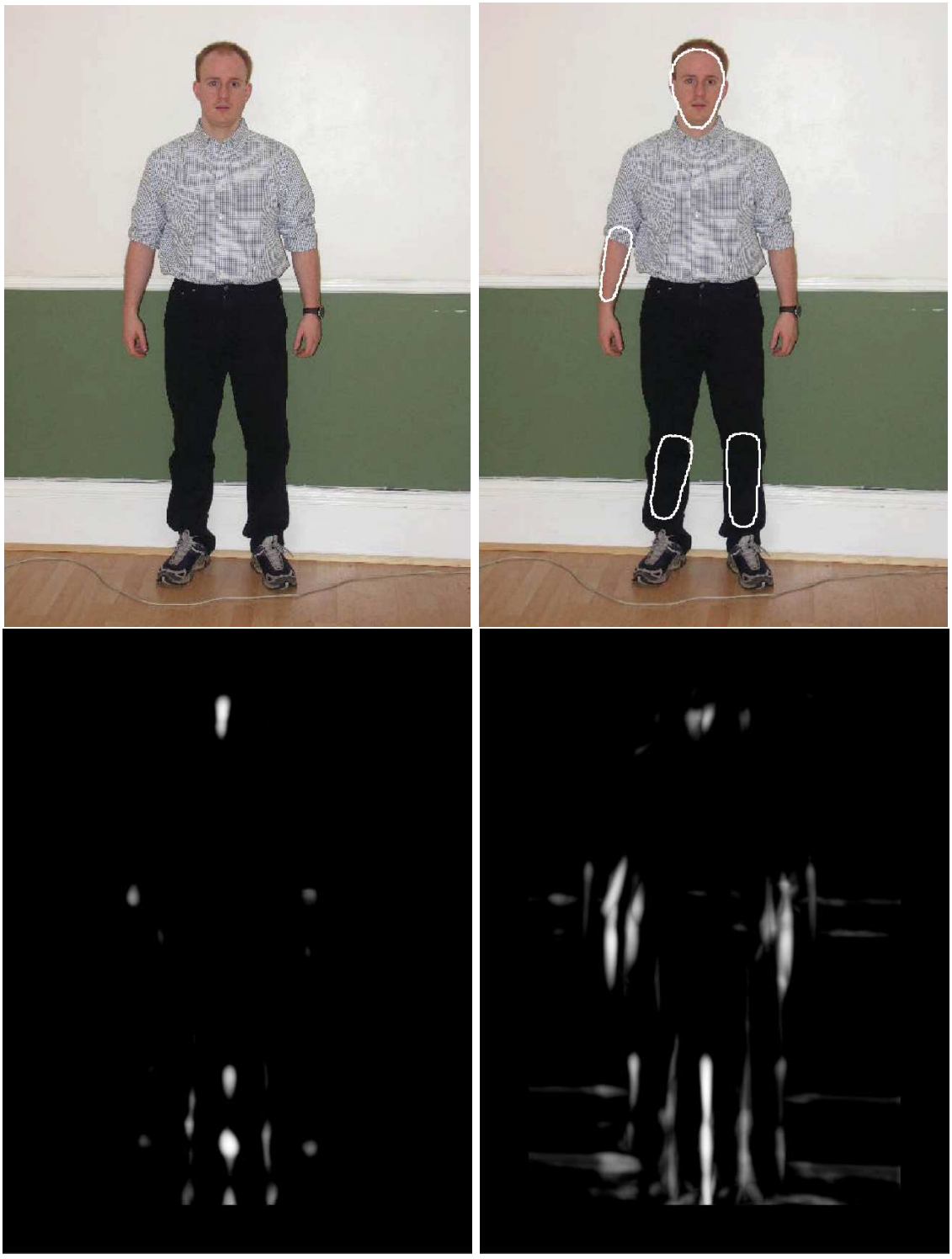

Figure 5.1: A simple indoor test image (with a textured foreground). Three of the limbs are correctly identified. The lower right arm is not identified. This is due to either scale errors, the presence of the watch or the approximate nature of the search. At the first stage 802 parts candidates were identified. At stage 2 this number had been pruned to 102 candidates (196 likelihoods ratios were evaluated). At stage 3 this reduced to 89 parts (220 likelihood evaluations). At stage 4 this reduced to 18 parts (126 likelihood evaluations). The last stage found 12 parts (89 likelihood evaluations). As can be seen the maximum occurred at level 4.

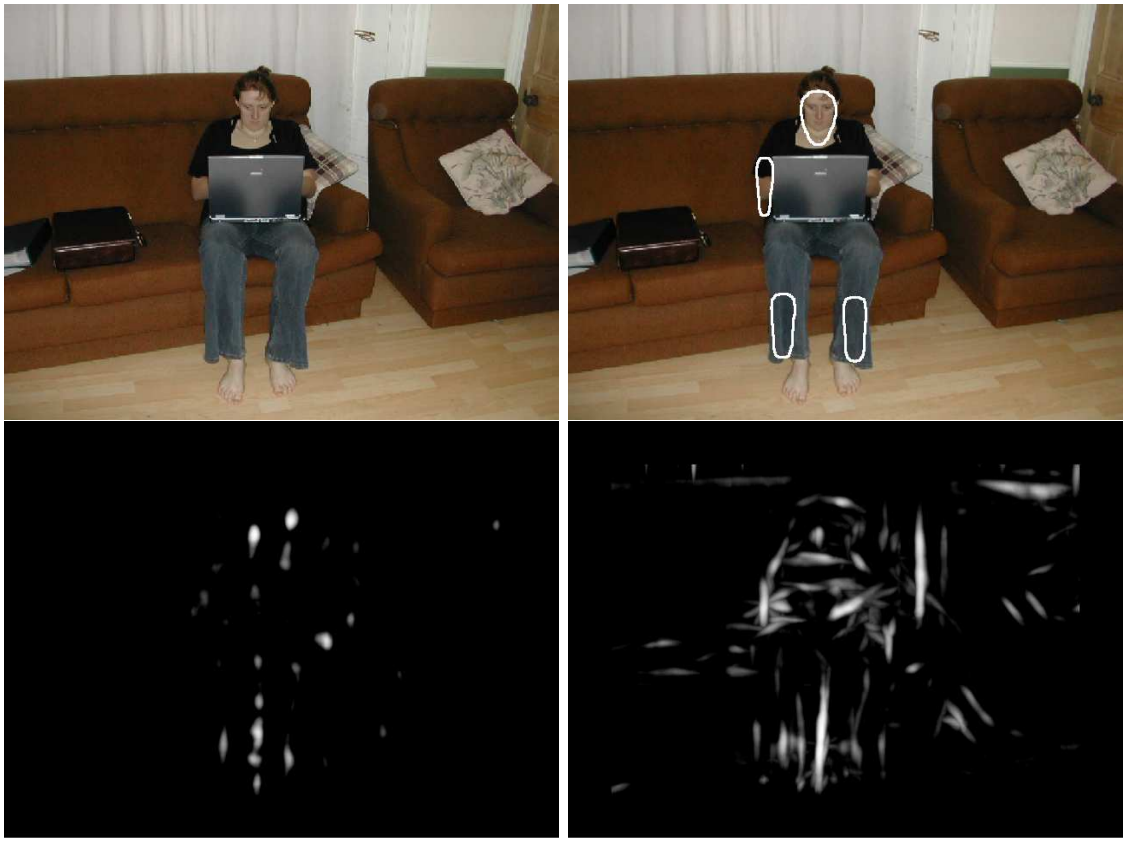

Figure 5.2: An indoor scene with occlusion from a laptop. The system is able to determine the correct position of the head, a lower arm and the lower legs.

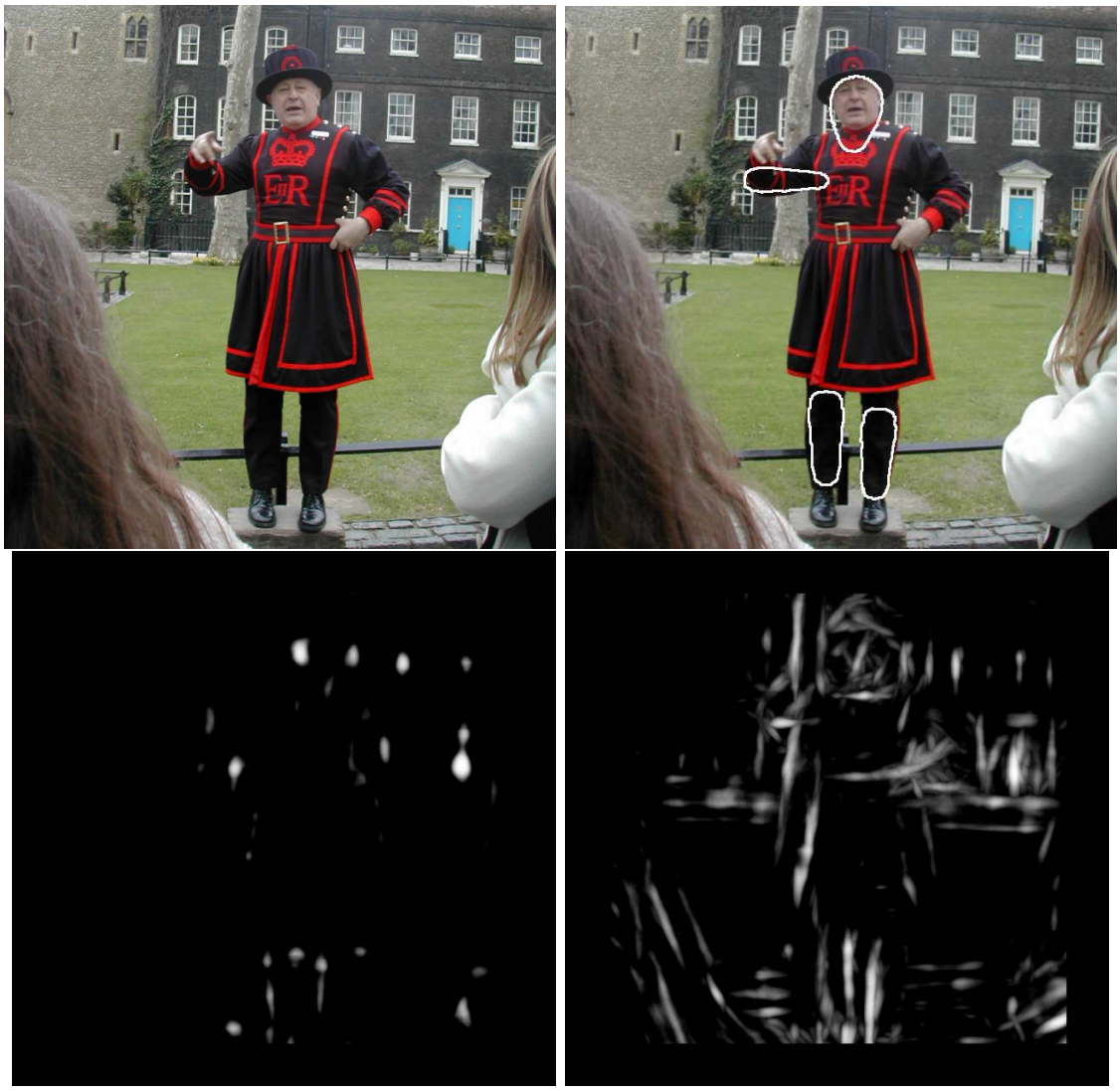

Figure 5.3: An outdoor scene containing a subject with an unusual clothing style. The head and lower legs are correctly identified. The lower arm is identified but is misaligned. The windows in the background cause many false responses to both the head and limb models.

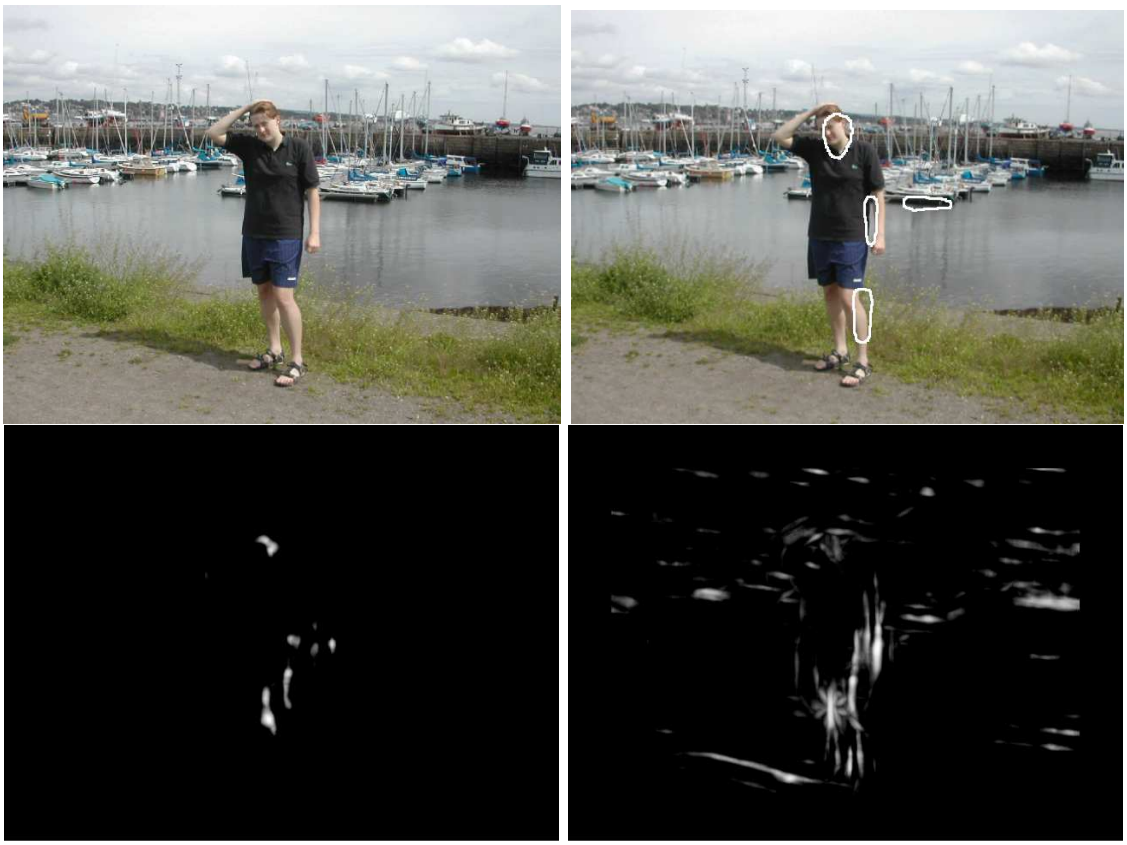

Figure 5.4: An outdoor scene with a subject wearing short sleeves and shorts. The large vertical edge responses from the sail masts make this a challenging image for edge based likelihood models. It can be seen that the new likelihood is able to discriminate such clutter. However, the space between the arm and torso is a good match and is able to pair with clutter on the background. The head and a lower leg are identified correctly. The other leg cannot be paired since it is heavily shadowed.

Figure 5.5: An outdoor scene with the subject walking away from the camera with an arm occluded. The system correctly identifies the head and lower leg but does not find the paired lower leg (due to shadowing) or the un-occluded lower arm (due to scale difference)

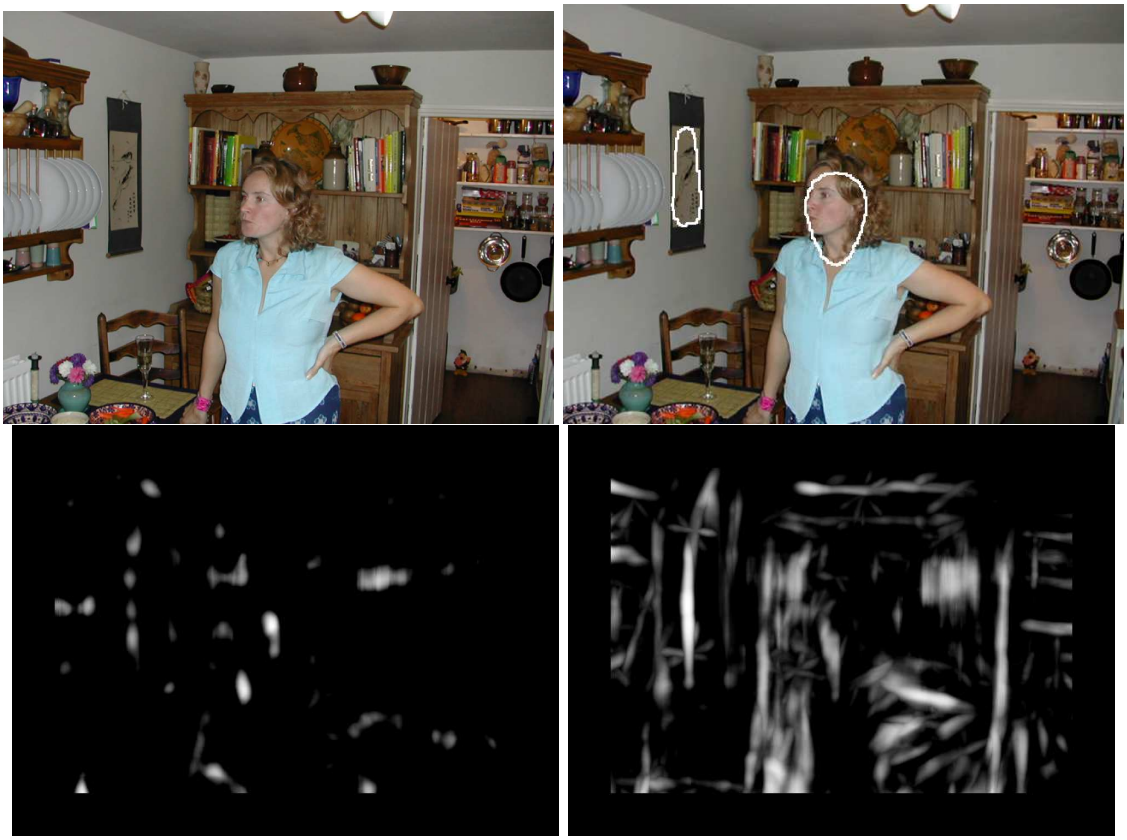

Figure 5.6: A *scale* = 1.6 indoor scene. The head is correctly identified. A lower arm is incorrectly identified on a picture in the background. It is likely that a larger configuration (with upper limbs) would be able to discriminate this configuration. The projection shows that there are large amounts of background clutter which respond to the limb model.

Figure 5.7: In this scene the head and a single lower arm are identified. The search did not find the paired lower arm. The lower legs are camouflaged into the background due to poor illumination.

Figure 5.8: An outdoor scene with a subject dressed in a challenging manner. The hat and socks make a single optimum scale parameter difficult to identify. However, the system correctly identifies the head and a lower leg. The sunlight at the bottom of the tree gives a strong false response to the head model.

Figure 5.9: An outdoor walking scene. Interestingly, the system is able to localise an arm that neighbours the body (and this is not accounted for by the adjoining region model). However, due to differences in scale the lower legs are not identified. Small changes in scale gave quite different answers: a smaller scale identified the rucksack strap as an arm, a larger scale identified the lower legs.

Figure 5.10: An outdoor sports scene. The system is able to find the head and lower legs correctly. It is not clear why the arms are not located since these are more contrasting than the lower legs. The basket ball and post give strong responses to the head and limb models respectively. At slightly different scales, different configurations are identified with large likelihood ratios and therefore this result should not be considered robust.

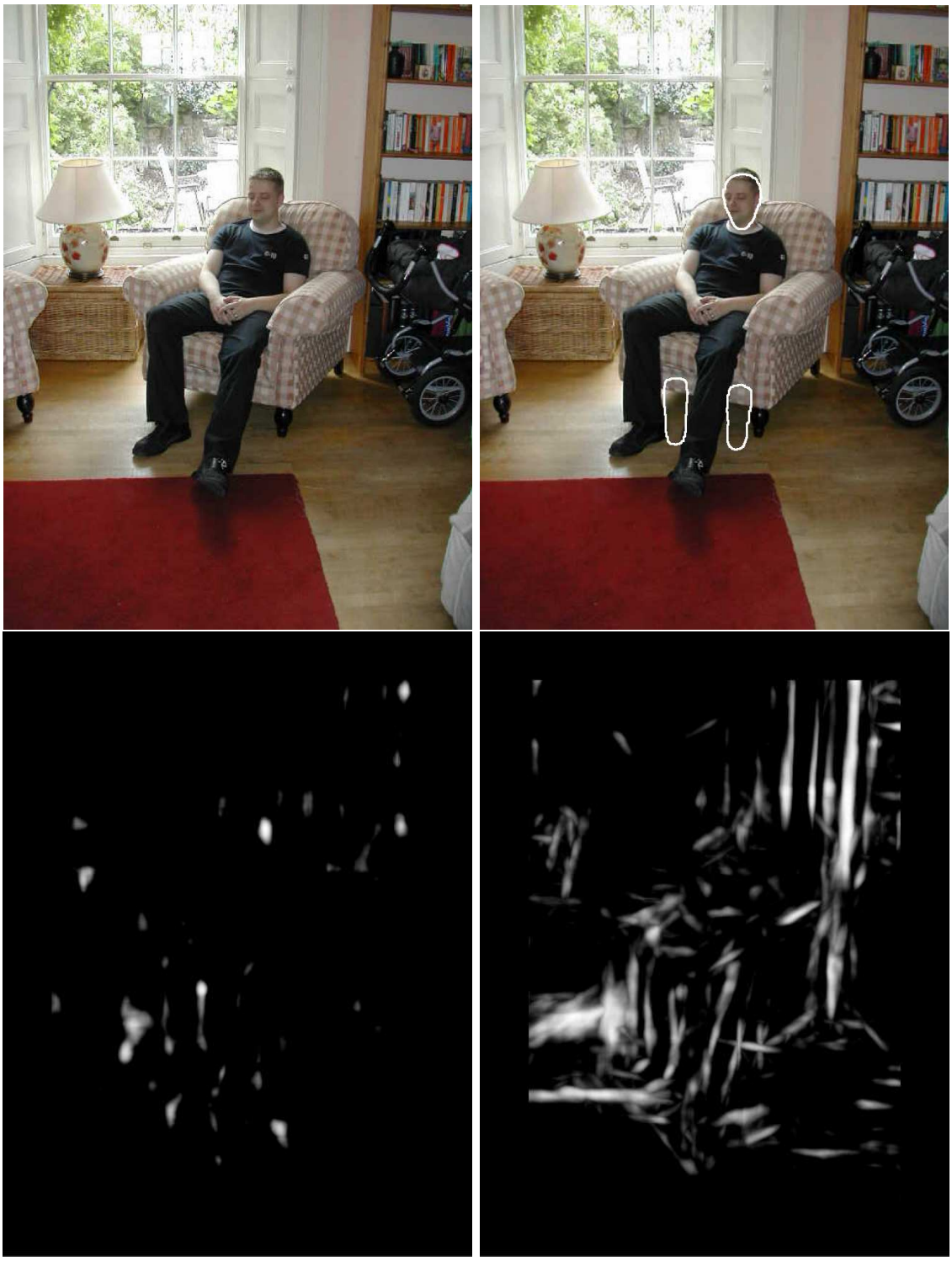

Figure 5.11: An indoor scene with the subject sitting down. The head is correctly identified. The system identified the lower legs incorrectly. This is likely due to perspective effects causing a difference in scale between the head and legs.

Figure 5.12: An outdoor test for a subject at a scale of $s = 4.5$. The system is slow to run since the number of points in the foreground is very large and the coarse sampling has not been tweaked. The system correctly identified the head and a lower arm. This result is not robust to small changes (0.25) in scale.

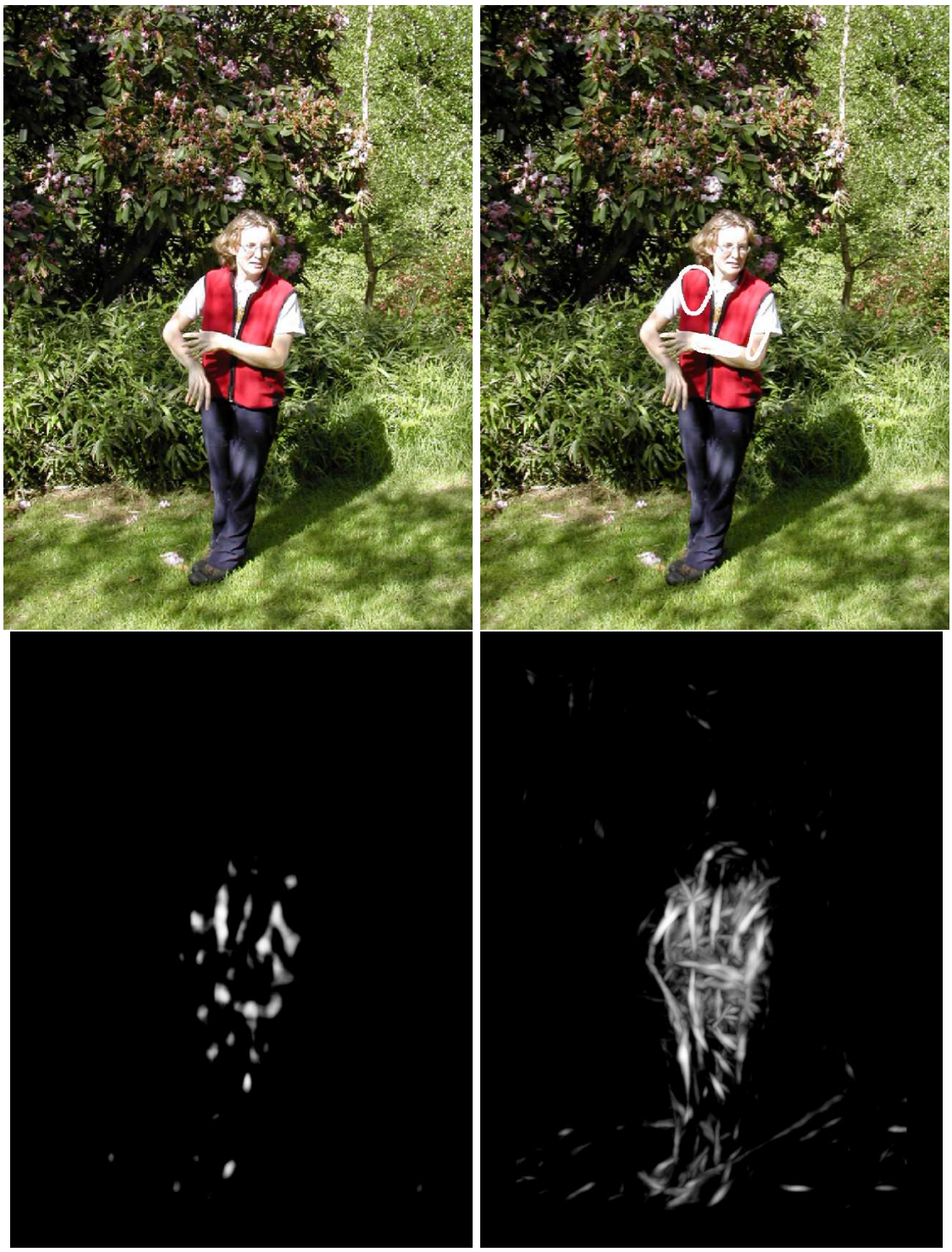

Figure 5.13: An outdoor scene with a subject wearing bright, contrasting clothing. It can be seen from the projections that there are many false positives on the body and that the head does not give a strong response. The head is incorrectly identified in the final estimate. A lower arm is correct, however it is paired with the upper arm. A stronger pose prior would improve performance in this situation.

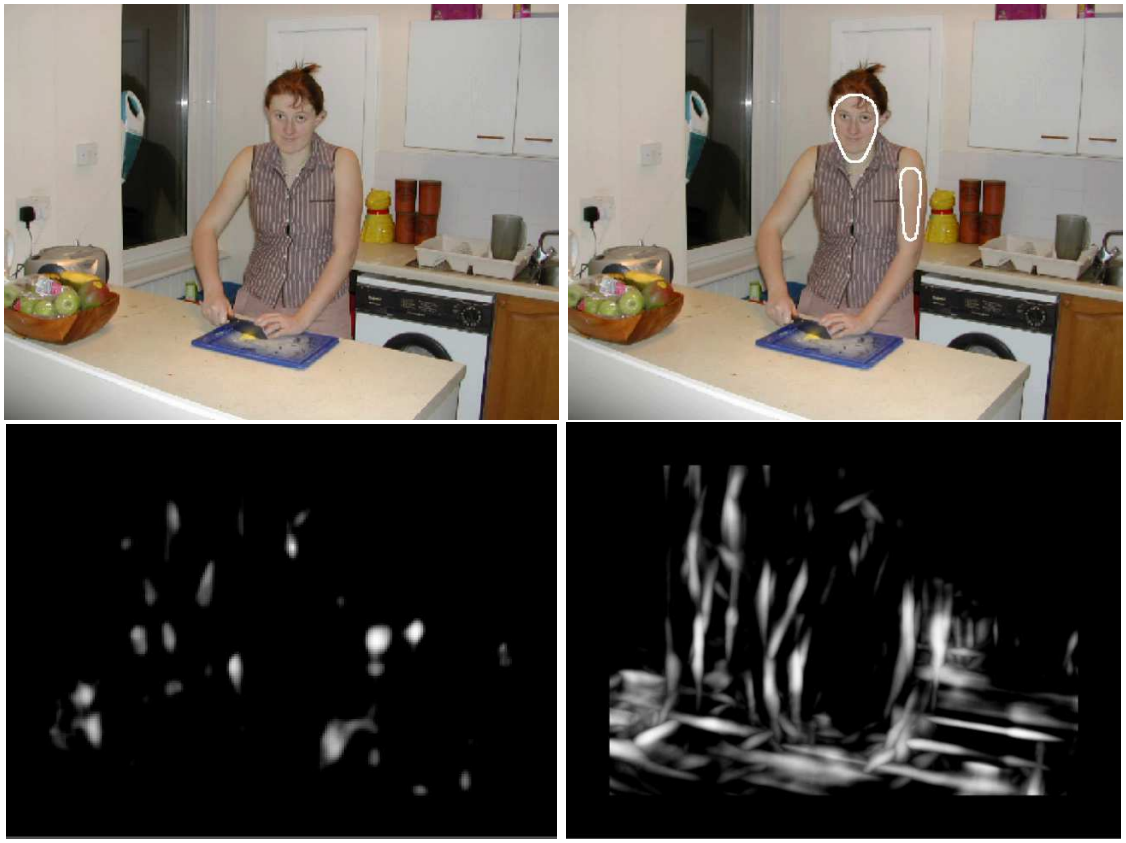

Figure 5.14: An indoor scene with other object occlusion. The head and lower arm are correctly identified.

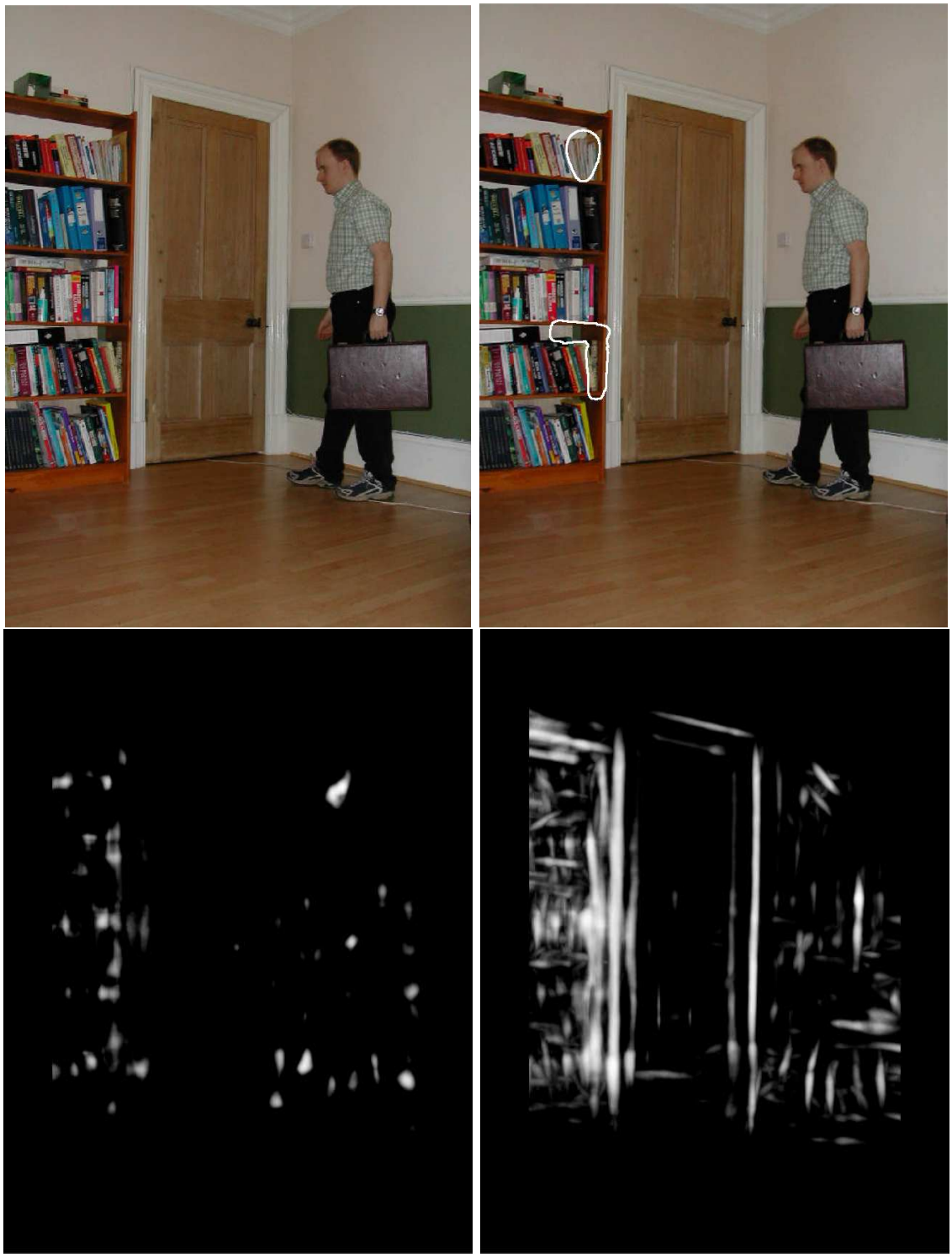

Figure 5.15: The clutter on the book case causes a completely incorrect result. A correct two part configuration did however have a high likelihood ratio.

Figure 5.16: A cluttered outdoor scene. The system correctly locates the head and a lower arm. The (neighbouring) lower legs are not identified. One leg is significantly shadowed.

Figure 5.17: An indoor scene with a subject walking and performing an action. The system is able to localise the head and legs correctly. The lower arms are not located. This could be due to scale differences between the identified parts and the unclothed lower arms.

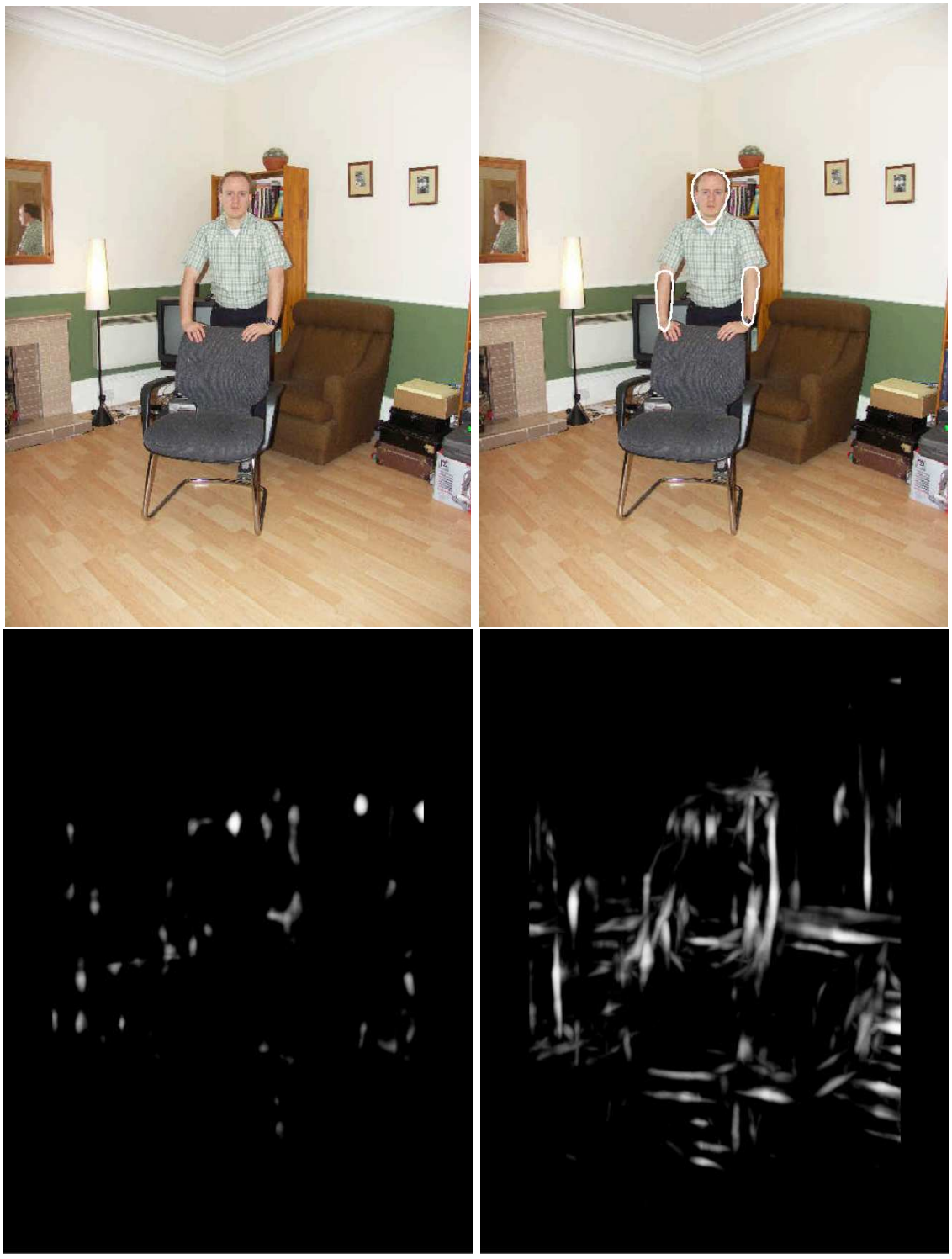

Figure 5.18: A cluttered indoor scene with a subject occluded by a chair. The system correctly localises the head and lower arms. The discrimination of the lower arms depends critically upon the inter-part appearance constraints.

Figure 5.19: An indoor scene with clutter from a bookcase and lounge furniture. Pixel saturation due to a reflection from a garden chair causes a strong head hypothesis which results in a completely incorrect configuration. The system is not able to discriminate enough of the limbs on the correct configuration.

Figure 5.20: An indoor scene with significant shadowing and other object occlusion. The system correctly identified the head. The upper arms were identified instead of the lower arms as the system was able to pair them. A stronger prior may help improve this result. A lower leg was incorrectly identified on the sofa.

It should be emphasised that although inter-part links are not visualised here, these results represent estimates of pose configurations with *inter-part relationships* (in terms of appearance and pose) as opposed to independently detected parts. Furthermore, it is re-emphasised that the camera parameters and subject identity are unknown (apart from scale). Close inspection of the input images also shows significant JPEG artifacts.

These results support the initial hypothesis that it is possible to efficiently find highly informative partial solutions in real-world images using a strong single part and inter-part likelihood model. Furthermore, the final result confirms that the system is able to recover pose in the presence of other object occlusion and over large changes in scale (i.e. 600% variation) with manual initialisation of the scale parameter.

Two important issues can be identified from these results. First, it can be seen that baggy clothing and perspective effects cause changes in scale between body parts. Indeed it is difficult to identify a single 'correct' scale for some images. Second, it can also be seen from the limb likelihood projections that pose estimation in indoor scenes is more difficult due to clutter from man made objects (such as door frames). Further research into border likelihood models would therefore likely improve performance for these scenes.

To improve these results, better search techniques are needed that are able to hypothesise larger configurations that account for self occlusion and lack of contrast due to neighbouring parts. For example, in order to find an upper leg, given the position of the lower leg, it is necessary to hypothesise the position of both upper legs to take into account the lack of contrast and pairwise similarity. A stronger pose prior would also improve the efficiency and quality of the results.

Since much of the information about pose is contained in the smaller sub-configurations, especially in the outer limbs, finding small numbers of parts is not as significant a drawback as one might immediately assume. Moreover, these results compare favorably with state of the art pose estimation systems that require more restricted scenes and assume more is known about the appearance (e.g. Ioffe and Forsyth [2001a,b]). In particular, these other systems also often only find three to five body parts.

5.5 Summary

To summarise this chapter the questions that were posed in the first section are briefly re-answered:

What approaches to estimation have been proposed? Two broad approaches can be identified in the literature for human pose estimation: the combinatorial approach and the state search approach.

What are the limitations of these approaches The combinatorial approach relies upon bottom up part detection and restrictions on inter-part relationships in order to efficiently estimate pose. The state search approaches make no such assumptions but relies upon manual initialisation and a strong temporal prior in order to be successful.

How can these limitation be overcome? The limitations of previous approaches can be overcome by using a strong likelihood model, that allows good discrimination of single parts and encodes inter-part appearance constraints, coupled with a iterative search scheme that is able to search different partial configurations. This Chapter developed the basis of such a search scheme and gave encouraging but limited results on a set of challenging images.

Chapter 6

Conclusion

The previous Chapters have described a vision system that performs human pose estimation from real world images. The system was described in terms of three layers: formulation, likelihood and estimation. However, as with any discussion by parts, the relationship between the parts, which is often important, can be neglected. Furthermore, for complex systems it can be difficult to discern the relative importance of different points. Therefore, in this, the concluding chapter, a more holistic discussion is presented that relates the different components and identifies the most important contributions and limitations.

6.1 Summary of Contributions

It was proposed that the two fundamental problems of visual pose estimation that should be focussed upon are:

1. Discriminating between a subject with complex, unknown appearance and a cluttered, unknown scene that possibly occludes parts of the subject.

2. Formulating the pose estimation problem such that efficient, accurate global estimation is possible in such conditions.

This is in contrast to the majority of human tracking and human pose estimation systems that focus upon the estimation problem. Indeed, the approach taken in this work was to ease the estimation problem by improving the formulation and likelihood.

With this in mind, the first important contribution made in this thesis was the *partial configuration* formulation that allows pose configurations with variable numbers of parts to be compared in a principled manner. Adopting such an approach has two key advantages. Firstly, it allows *other object occlusion* to be modelled when the structure of the scene is unavailable (which is the case in the great majority of applications). Other object occlusion is common in real world images but has been largely ignored in previous research. Secondly, encoding pose using partial configurations allows more *efficient and flexible* search schemes to be implemented. In particular, it allows bottom up part hypotheses to be used to focus the search for larger configurations and thereby achieve efficient localisation. However, in contrast to previous formulations, it does not make restrictive assumptions on (self and other object) occlusion and inter-part relations.

The second important contribution of this thesis is an efficient, highly discriminatory spatial likelihood model. This model is composed of two complementary components. The first component is used to efficiently discriminate single parts based upon the difference between the texture in regions induced by the high level shape model. Using the high level shape model in this way allows better discrimination of shapes that have a complex, textured appearance than models based upon bottom up boundary measurements. The second component is used to efficiently discriminate larger configurations based upon inter-part appearance similarity.

In addition to these two primary contributions other ideas were proposed to further improve visual pose estimation. A probabilistic model of shape was proposed that, when transformed into the image using the pose parameters, encodes the uncertainty in visibility of parts at points in the image and thereby forms the basis of the measurement process. This probabilistic approach is important for efficient pose estimation where there is significant un-parameterised variation due to factors such as clothing and inter-person variability. Moreover, additional gains in efficiency can be made by combining similar part models and removing degrees of freedom that have little effect on the appearance. This probabilistic region formulation is further developed by probabilistic notions of self occlusion and contrast.

Lastly, in addition to the research on human pose estimation a novel human tracking likelihood based upon absolute foreground appearance was presented as an Appendix. The model attempted to address the problem of appearance estimation in the presence of rapid appearance adaptation by taking advantage of the structure of clothing.

6.2 Future Work

The estimation scheme is the most significant limitation of the current work. Although the current method is able to localise many of the key outer body parts, which contain most of the pose information and would be used for human computer interface applications, for example, the system is not capable of localising the remaining parts. This is due to the simplistic algorithm that was employed that does not attempt to predict the position of missing parts or hypothesise multiple parts at the same time and thereby account for the significant lack of contrast some parts have with their neighbours (e.g. the upper legs). Therefore, future work will de-

velop sampling schemes that, given the position of a set of body parts, form larger configurations that account for the lack of contrast and self occlusion. The efficiency of such a scheme will depend upon the sophistication of the prior.

Indeed, the approach of partial configurations opens the door for novel future estimation approaches to greatly improve estimation efficiency. It is envisaged that, due to the flexibility of the parameterisation, a set of optimization methods such as genetic style combination, part based importance sampling (possibly using skin colour), prior based prediction, local search, depth re-ordering and part re-labelling can be combined using a scheduling algorithm and a shared sample population to achieve rapid, robust, global, high dimensional pose estimation.

Another key limitation is the assumption of a single subject. The posterior ratio formulation in principle allows detection and pose estimation for multiple people but the system was trained and evaluated on images containing only single subjects. It is likely that the inter-part appearance likelihood model will be useful in discriminating between body parts of different people.

Future work might use a texture in addition to colour to enhance discrimination of body parts, especially those that are overlapping parts with similar appearance and have oriented texture. In particular, future work would address issues regarding the description of local texture (e.g. Gabor filter banks or joint image statistics (Varma and Zisserman [2003])), the conditional independencies between measurements and their fusion with other cues. Indeed, early investigations into texture models were started but gave inconclusive results.

The system also lacked a comprehensive pose prior and, although it worked in spite of this omission, it is likely that performance will be improved with its inclusion. Furthermore, the assumption that body parts are treated as equally likely to be

occluded needs to be addressed.

Future work might also address the problem of efficiently calculating the optimum 'marginalised' likelihood model for smaller partial configurations based upon the expected position of missing parts.

An interesting future direction might also be to combine the pose estimation ideas with the temporal likelihood model proposed in Appendix A. The main problem with the temporal tracker (as with most trackers) was the need for manual (re)initialisation and the pose estimation system partly addresses this limitation. However, it is the author's opinion that the estimated foreground appearance would be better used to focus the search (i.e. through importance sampling) than as part of the likelihood, since the foreground appearance estimate is subject to large uncertainty and adaptation.

Finally, as with most human pose estimation and tracking systems, there is a need for large scale testing and evaluation. In order to draw quantitative comparisons between different systems a public database with ground truth data, similar to the CMU motion database (Gross and Shi [2001]), would be a useful development.

6.3 Closing Remarks

This thesis is a small but hopefully significant step toward human pose estimation from real world images. Some of the methods presented in this thesis have been patented and have already made a valuable contribution to a commercially viable pose estimation system.

Appendix A

Temporal Likelihood

A.1 Introduction

In this Appendix a likelihood model for detailed human tracking in real world scenes is presented. In contrast to the rest of this thesis, the formulation used here is a more standard, fixed number of parts 3D articulated model. In this formulation, the appearance, modelled using feature distributions defined over regions on the surface of an articulated 3D model, is estimated and propagated as part of the state. The benefit of such a formulation over currently used techniques is that it provides a dense, highly discriminatory object-based cue that applies in real world scenes. Multi-dimensional histograms are used to represent the feature distributions and an on-line clustering algorithm, driven by prior knowledge of clothing structure, is derived that enhances appearance estimation and computational efficiency. An investigation of the likelihood model shows its profile to be smooth and broad while region grouping is shown to improve localisation and discrimination. These properties of the likelihood model ease pose estimation by allowing coarse, hierarchical sampling and local optimisation.

Human tracking research often uses a generative, high-level model. Firstly, a parameter space is established, in this case describing the pose of the human. The aim of the tracking system is then to estimate the parameters of the model $X_t = \{x_0, ..., x_T\}$, given the observations $Y_t = \{y_0, ..., y_T\}$ and a body of prior knowledge $p(X_t)$ (where $x \in \Re^n$, y_t represents an image frame and $t \in [0, T]$). However, due to system noise, model inaccuracies and loss of information, there are genuine ambiguities and our knowledge is represented using a conditional probability distribution $p(X_t|Y_t)$ instead. In general, distributions at new times can be found using (A.2).

$$p(x_t|Y_t) = \int ... \int p(X_t|Y_t)dx_0...dx_{t-1} \quad (A.1)$$

$$\propto \int ... \int p(Y_t|X_t)p(X_t)dx_0...dx_{t-1} \quad (A.2)$$

The first term on the RHS of (A.2) represents the likelihood of a path through state space given the image sequence and the second term is the prior over that path. In can be seen that the dimensionality of this integral grows with time as more information is considered and direct evaluation becomes prohibitively expensive. Therefore, it is usual to consider the state evolution to be a Markov process and the distributions can then be found recursively:

$$p(x_t|Y_t) \propto p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|Y_{t-1})dx_{t-1} \quad (A.3)$$

The first term on the RHS of Equation (A.3) represents a single image likelihood model and the second represents the probability of transitioning from the previous posterior distribution. It is important to realise that the posterior probability distribution is induced by the chosen likelihood model and motion model. Accu-

rate modelling of these terms allows for easier and more accurate estimation. The topic of this paper is the derivation and evaluation of a likelihood model that allows accurate, efficient tracking in real world environments in a Markov-Bayes, analysis-by-synthesis framework. Modelling of human motion is not discussed here.

A difficulty in the case of visual human tracking, is that, due to the large variation in a subject's appearance, the single frame likelihood model is not known *a priori*. Such models may be constructed by making limiting assumptions but the resulting systems will not allow for accurate estimation in other, more realistic, scenes.

Here we propose to construct a likelihood model by propagating the foreground appearance as part of the state. The state consists of a shape (pose) component and a texture component, i.e. $x \equiv \{x_{shape}, x_{texture}\}$. Such a likelihood model should be highly discriminatory. However, problems arise when trying to estimate a complex, changing appearance given the limited amount of available data and computational resource. Furthermore, model initialisation is essential when using a likelihood of this form. Existing tracking systems typically require manual initialisation and this approach is similarly adopted in the experiments described here. The method presented should ultimately be considered to be part of a larger system that includes likelihood models that do not require manual initialisation in order to perform automatic (re)initialisation and efficient, iterative tracking.

A.2 Method

Many previous systems rely upon likelihood models that assume restrictive scene conditions such as tight, high contrast, textureless clothing or a static, simple or known background. The system presented here is less restrictive in that it copes with textured and loose-fitting clothing.

A.2.1 Shape Model

The body is highly deformable and exact modelling of its form is infeasible and unnecessary in this context. Its important properties can be captured using an articulated body model. A 3D articulated shape model is used since it has a low dimensionality, captures the kinematic structure of the body, allows for easy encoding of prior knowledge such as joint limits, automatically handles self-occlusion and enables changes in body part appearance due to rotation in depth to be handled explicitly.

The shape component of the state space, x_{shape}, then in general becomes the relative position and orientation of the primitives, their shapes and their sizes. Each of the shape primitives, indexed by $b \in \{1...B\}$, has a surface that is naturally described using some co-ordinate system, a point in which is denoted by ω_b. For example, the surface of a fixed size cylinder is conveniently described by a length and an angle, i.e. $\omega_b = (l, \theta)$. A point on the subject is then specified by the pair (b, ω_b). In order to project a surface point onto the image plane, the co-ordinates are first converted to Cartesian form. Homogeneous, relative transformations are then chained together to transform up the kinematic tree into world co-ordinates and finally, using a camera model, to project onto the image plane.

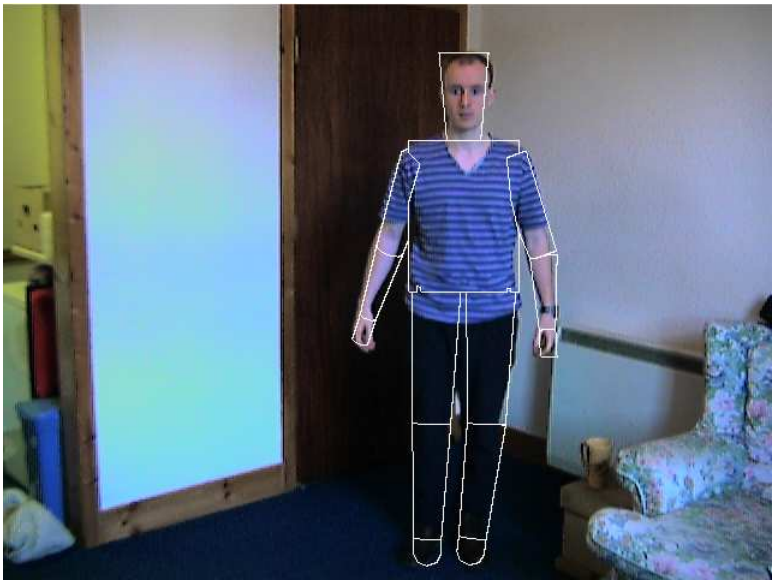

Figure A.1: The model overlaid on a frame from a waving gesture sequence used throughout this appendix to illustrate ideas. Notice the approximate alignment of the edges.

Figure A.2: Frames 0, 10 and 26 from the waving sequence.

In the particular implementation described here, the body was represented using elliptic cross-section cylinders with constant, manually initialised sizes and shapes. The camera was modelled using an orthographic projection since the sequences under consideration did not contain strong perspective effects and the likelihood model was relatively insensitive to small changes in shape. However, the extension to perspective projection is straightforward. Independent movement of the head, hands and feet was not modelled, leaving a total of 22 degrees of freedom, encoded as 3 root translations, T_{Torso}, and 19 rotational degrees of freedom, $\{R_b\}$, four for each limb and three for the torso. Prior knowledge on joint angles was encoded using a quadratic ramp function at boundaries. The shape component, which is

illustrated projected onto an example image in Figure A.1, can thus be written as:

$$x_{shape} = \{T_{Torso}, \{R_b\}\} \tag{A.4}$$

A.2.2 Likelihood Model

Likelihood evaluations are performed by projecting the model onto the image and comparing observed image features with expected values. Feature values are denoted here by $\vec{q}$. Expected features, such as pixel colour or local filter responses, will in fact have a (multi-modal) probability distribution rather than a single value. This is because of body model inaccuracies, discretisation and noise. In general, the feature distributions are not known *a priori*. Most likelihood models, including those using edges (e.g. edge strength), require parameters to be set in order for matching to be performed. What varies is the number and range of these parameters. Some cues, such as skin colour and edges, can be reliably configured off-line and the parameters fixed. Such *static* models can be used for more automatic (re)initialisation. Here we consider a likelihood model where the expected features for regions, Ω_b, on the surface of the 3D articulated model are propagated as part of the state:

$$x_{texture} = \{p_{(b|\omega_b)}(\vec{q}), \omega_b \in \Omega_b\}. \tag{A.5}$$

Hence we consider only colour distributions, the primary reasons being their quasi-invariance to viewpoint and ease of implementation. Since clothing is often textured these distributions can be multi-modal. Therefore, (non-robust) template tracking is inappropriate. Matching using distributions provides greater discrimination than matching individual measurements. We propose using normalised multi-dimensional

histograms to represent the feature distributions and denote them by $H_{(b,\omega_b)}$. Other possible distribution statistics include, for example, cumulative histograms and moments.

In order to find the likelihood of a hypothesised pose, rays are cast into the scene at each pixel to determine the point of intersection, if any, with the shape model. Hypothesized histograms, $H'_{(b,\omega_b)}$, are collected for each region on the articulated model from the current image. These are compared to the propagated appearance using a distribution similarity measure, S. The likelihood is then modelled, in Equation (A.6), as the sum of similarities weighted by the visibility of the region in the image, where V denotes the set of pixels corresponding to the body and $|V|$ denotes the number of visible pixels. In this way larger regions contribute more weight to the likelihood.

$$p_{divergence}(y_t|x_t) \propto \frac{\sum_{\{V\}} S(H'_{(b,\omega)}, H_{(b,\omega)})}{|V|} \tag{A.6}$$

Clustering using split and merge

Estimating the feature distributions at points on the body is difficult given the limited image data available. Furthermore, the distributions are varying over time due to illumination changes and clothing movement. However, we observe that many of the points on the surface of the body belong to the same piece of clothing and will therefore often have similar distributions. A clustering routine can group points on the 3D model to improve estimation. We use a computationally efficient, iterative grouping scheme. Consider estimating an unknown bin, $H_{(b,\omega_b)}(q)$, from the histograms for other body regions using a weighted sum:

$$H_{(b,\omega_b)}(q) \approx \sum_{b'} \sum_{\omega'_b} H_{(b',\omega'_b)}(q) p(H_{(b',\omega'_b)}|H_{(b,\omega_b)}) \tag{A.7}$$

The conditional probability can be modelled in a Bayesian fashion using a likelihood determined from a similarity measure, S, on the known histogram bins and a prior determined from knowledge of clothing structure:

$$p(H_{(b',\omega')}|H_{(b,\omega_b)}) \approx S(H_{(b',\omega')}, H_{(b,\omega_b)}) P_{(b,\omega),(b',\omega')} \tag{A.8}$$

Direct use of the sum in (A.7) is not computationally feasible since it involves summing over all points on the body for each unestimated histogram bin. However, it can be seen that large contributions to the sum must be similar to the histogram in question and therefore similar to each other. Therefore, the sum is reasonably well approximated by the average bin value taken from the group of similar regions.

To perform region merging, a threshold K is introduced. It controls the level of detail represented by the system and encodes the model order. When K is large, the system behaves like a template tracker by preserving individual regions. When K is small, the system behaves like a blob tracker, ultimately representing the person using a single distribution. For a particular sequence, with a particular image resolution, target size and level of noise, there will be an optimal choice of threshold for tracking that balances appearance estimation with excessive loss of local structure. The merging decision criterion then becomes:

$$S(H_{b,\omega_b}, H_{b',\omega'_b}) > \frac{K}{P_{(b,\omega_b),(b',\omega'_b)}} \tag{A.9}$$

The clothing structure prior, $P_{(b,\omega_b),(b',\omega'_b)}$, in (A.9) is learned from example images

b	ω_b	b'	ω'_b	$P_{(b,\omega_b),(b',\omega'_b)}$
Upper Arm	l, θ	Upper Arm	$l, \theta + \delta\theta$	0.9
Head	l, θ	Hand	-	0.7
Upper Arm	l, θ	Upper Leg	-	0.3

Table A.1: Example histogram merge priors

of differently clothed people by manually aligning the model to the image and performing an exhaustive pairwise comparison. The prior is set to the average of the observed similarities. Examples are shown in Table A.1. Currently, only a small data set has been used to learn the prior.

Merging is an $O(u^2)$ operation, where u is the number of unique regions. However, this cost is greatly outweighed by the improvement in computational efficiency due to reductions in the number of region comparisons and storage overhead. Figure A.3 illustrates a body part model with region grouping leading to shared feature distributions.

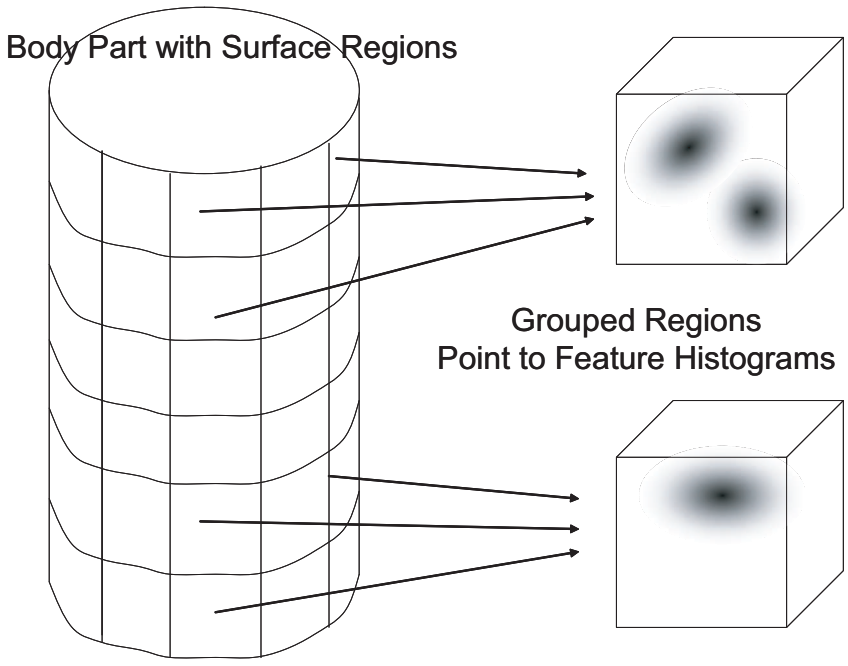

Figure A.3: A body part where grouped regions have associated feature distributions.

Since regions can erroneously merge we also introduce a splitting operation. How-

ever, performing this in a similar manner to merging requires unique histograms to be stored for every atomic region, resulting in a large storage overhead. Therefore, we currently use an heuristic splitting criterion based upon a threshold on the sum of histogram lookups in an atomic region from the current image. This rule is particularly efficient which is important since it is performed for every atomic region in every frame.

Region Comparison Techniques

Many histogram similarity measures have been proposed. These include inter-bin measures such as the Bhattacharyya, Jeffrey, Minkowski, Intersection, χ^2, and Kullback Leibler, and intra-bin measures such as QBIC and the Earth Movers distance, see e.g. Puzicha et al. [1999]. Inter-bin measures are favoured here because of their lower computational cost. Sum of histogram back-projections, which is quicker to calculate online, can also be used but allows less discriminatory power since it uses each of the measurements independently, ignoring how these are distributed.

Background Model

The distribution induced by the likelihood model described so far cannot be used to disambiguate certain poses. For example, consider the waving sequence, where the lower arm, which is uniformly coloured, rotates in depth. As the arm foreshortens the hypothesised histograms will remain approximately constant and therefore so will the likelihood. This problem is illustrated in Figure A.4. To overcome this problem multiple solutions could be propagated using a semi-parametric or non-parametric density representation (Cham and Rehg [1999], Isard and Blake [1996]). However, this approach is particularly expensive when propagating an appearance

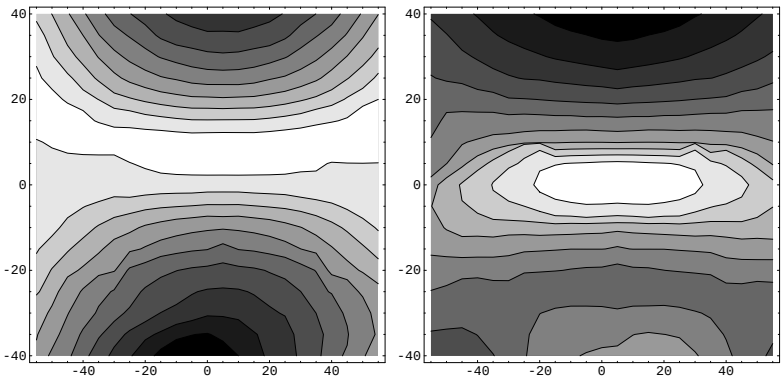

Figure A.4: Visualisation of the likelihood whilst rotating the lower right model arm against frame 10 of the waving sequence. The solution is centralised. Abscissa: out of plane rotation, ordinate: in plane rotation. The central ridge in the first plot has a large likelihood and illustrates the inability of the model to resolve out of plane rotations. The second plot illustrates how conditioning the likelihood to maximise foreground usage results in a single solution.

estimate and only delays decision making to a higher-level stage. For these reasons the approach taken in this work was to condition the likelihood to maximise the foreground usage as determined by a statistical background model. The background is modelled using a multi-variate Gaussian in chromaticity-intensity space for each pixel i. These Gaussian densities are recursively updated as described by McKenna et al. [2000]. The resulting modified likelihood is illustrated in Figure A.4 and is defined as:

$$p(y_t|x_t) = p_{divergence} \times p_{background} \tag{A.10}$$

$$p_{background}(y_t|x_t) = \frac{\sum_{i \in V} p_f(i)}{\sum_{i \in I} p_f(i)} \tag{A.11}$$

where $p_f(i)$ denotes the foreground probability density at pixel i, I is the set of all pixels in the image, and V is the set of pixels corresponding to the body.

Appearance Update

Figure A.5 shows a cropped region from the back projection of the arm histogram onto two frames, two seconds apart. It can be seen that the foreground appearance changes over time, sometimes quite quickly. The histograms are recursively updated to account for such changes using Equation (A.12). The rate of adaptation is controlled by the empirical constant, c. In the current implementation, the appearance model is updated using only those pixels that are sufficiently different from the background. This reduces the chance of the tracker diverging.

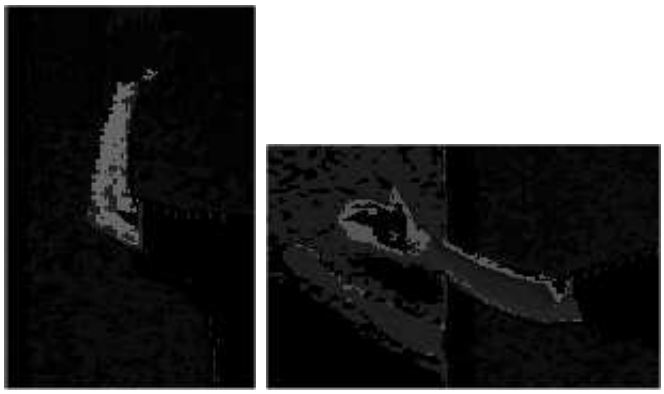

Figure A.5: Probability map for a lower arm histogram for frames from the waving sequence two seconds apart. It can be seen that the distribution has changed. The background has also changed.

$$H_t = cH'_t + (1 - c)H_{t-1} \tag{A.12}$$

A.3 Empirical Evaluation

In the implementation described here, a mixed iterative optimization scheme was employed. In the first frame, the pose was manually initialised and the surface feature distributions extracted. Then for each subsequent frame a semi-hierarchical, best-first search was first performed by coarsely sampling the state space around

an estimate given by a constant velocity motion model. The number and spacing of samples was chosen empirically using the likelihood response. For example, in the case of the upper arm with three degrees of freedom, sampling at four half limb widths in all directions at two half limb width intervals requires 64 samples. The components of the hierarchical samples with highest likelihoods were combined and used to seed local gradient-based search of the likelihood space. Including the hierarchical sampling has the effect of reducing the chance of getting trapped in local maxima and thereby allowing larger inter-frame motion. It is particularly useful when self-occlusion of the human body causes gradient information to be lost.

As previously mentioned, zeroth-order chromaticity-intensity statistics were used as features. A good histogram size was found empirically to be $12 \times 12 \times 8$ bins. No prior colour information was used. The system was implemented in C++. Preprocessing to find the histogram bins, an efficient model projection implementation and loop unrolling resulted in efficient likelihood calculations, the main computational burden for most trackers. The system required around $100MB$ to store the appearance model and made of the order of $10,000$ samples per frame at around $10ms$ per sample.

A.3.1 Likelihood Investigation

Figure A.6 shows different similarity measures as the model upper right arm undergoes image-plane rotation. It can be seen that grouping has a large effect on the profile of the response. A large amount of grouping causes local detail to be lost and localization suffers. In the case of too little grouping, the histograms are poorly estimated and the response is less smooth and has significant secondary maxima. It can also be seen that some similarity measures produce smoother responses and are less sensitive to the amount of grouping. In particular, the Bhattacharyya coefficient

was found to work well for tracking and grouping. The back-projection worked well when all background pixels were sufficiently different from the foreground.

A.3.2 Grouping Results

Figure A.7 illustrates the result of applying the grouping scheme to the first frame in the waving sequence. The number of unique regions is plotted against time. It can be seen that the system quickly converges to a stable region representation. It can be seen that occasionally split and merges are performed after the system has reached a steady state, this is primarily due to new regions becoming visible.

A.3.3 Tracking Results

Figure A.8 shows the result from successfully tracking in an everyday indoor scene. The subject is wearing loose-fitting clothes with both textured and plain regions. The background contains a significant amount of clutter, similarly coloured objects and uneven, natural lighting. The sequence was captured at 12 frames per second and at a resolution of 640×480 pixels. The frame-rate was lower than is usual, making tracking more difficult. The sequence contains 72 frames (6 seconds) which compares favourably to the length of sequences used in related published results. The update constant c in Equation (A.12) was set to 0.2.

It can be seen that the tracker maintains lock throughout the sequence including during the frequent self-occlusions. The hierarchical sampling along with a low frame rate makes the result a little jumpy. Furthermore, in some frames the upper arm is localised only approximately due to constraints on the model deformations. A strength of this region-based formulation is that the tracker degrades gracefully

under such conditions. The most significant error is that it switches the legs halfway through the sequence. This is due to the low frame rate and highly symmetric appearance and could be overcome by using a better motion model. In the final frames, tracking of a foot is inaccurate because of the heavy shadowing and the similarity of the background and foreground distribution.

A.4 Conclusions

A likelihood formulation was presented to allow for detailed, accurate pose estimation in unknown scenes. The model was based upon estimating the feature statistics of regions on the surface of a 3D articulated body model. Two problems with this approach are density estimation and computational efficiency. A region grouping algorithm was presented to overcome these difficulties and its benefits were illustrated.

The tracker worked well in real world scenes of moderate complexity and due to the properties of the likelihood model inference was more efficient. However, in more complex scenes, most notably outdoors, we found that the tracker diverged due to the rapid appearance adaptation. We believe that combining this model with a static cues would alleviate such problems. In this regard the method was a step towards addressing how to estimate human appearance and use this estimate for efficient iterative tracking.

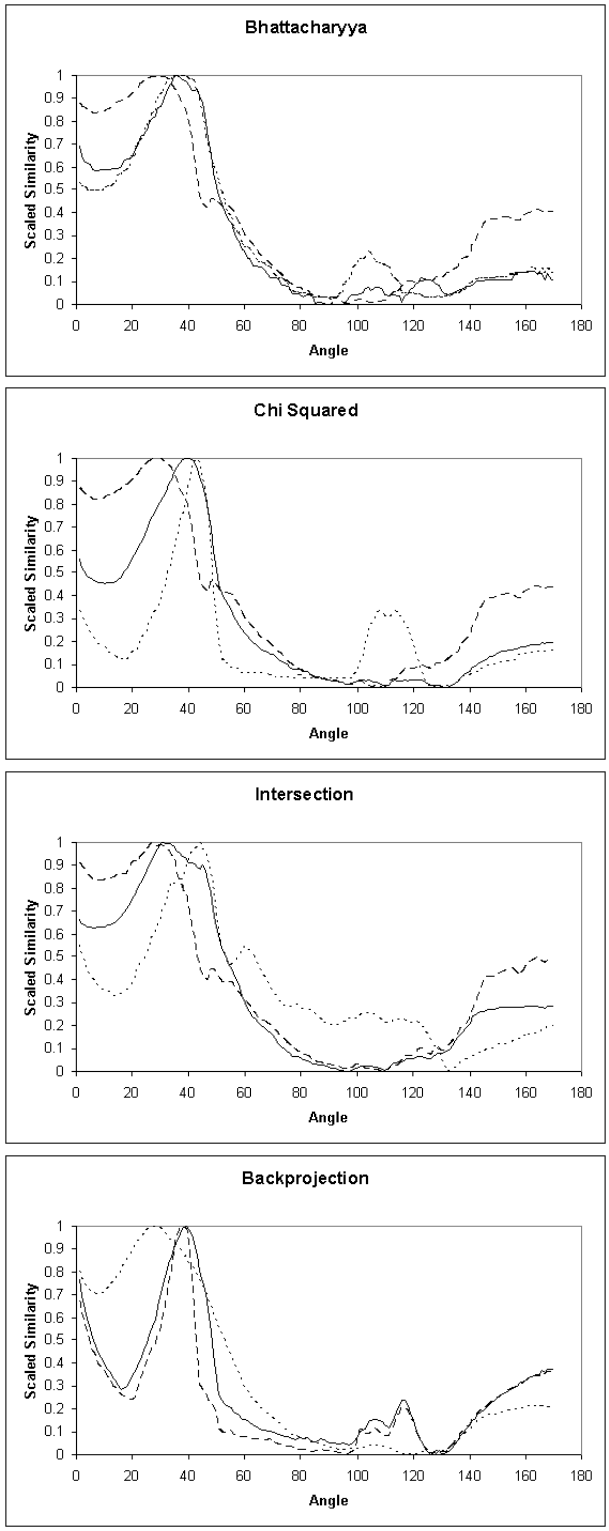

Figure A.6: Investigating the effects of region grouping and different similarity measures. Plots show the likelihood, $p_{divergence}$, for upper arm rotation against frame 10 of the waving sequence for three levels of grouping: dashed= 5 regions, solid= 20 regions, dotted= 120 regions. The true rotation angle was 39°.

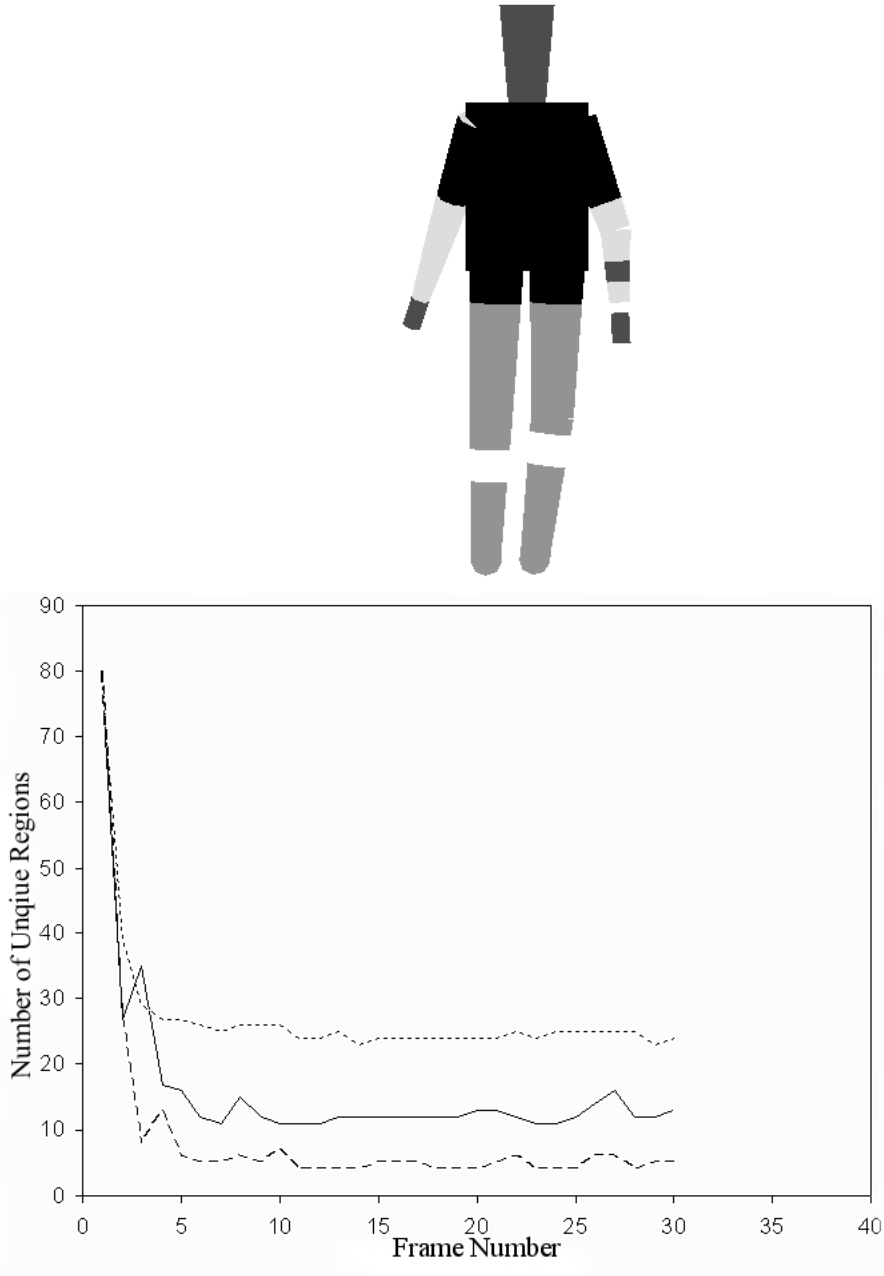

Figure A.7: Results from applying region merging to the first frame of the waving sequence. Top: visualisation of the largest grouped regions. Bottom: plot showing the behaviour of the grouping algorithm for three different merge thresholds.

Figure A.8: Tracking a highly textured subject through a few walking cycles containing self-occlusion in a cluttered indoor scene without a specific motion model.

Bibliography

J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999. 3.2.2

P. Baerlocher and R. Boulic. *Deformable Avatars*, chapter Parametrization and range of motion of the ball-and-socket joint, pages 180–190. Kluwer Academic, 2001. 3.2.3

C. Barron and I. A. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding*, 81(3):269–284, March 2001. 3.2.3

A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, 1994. 3.2.3

A. Blake and M. Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion.* Springer-Verlag, 1998. 3.2.3, 4.2.2

R. Bowden. *Learning non-linear Models of Shape and Motion.* PhD thesis, Dept Systems Engineering, Brunel University, 2000. 3.2.3

R. Bowden, T. A. Mitchell, and M. Sarhadi. Non-linear statistical models for the

3D reconstruction of human pose and motion from monocular image sequences. *Image and Vision Computing*, 18(9):729–737, June 2000. 3.2.3

C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8–15, Santa Barbara, CA, 1998. 3.2.3

C. Cedras and M. Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, March 1995. 3.2.2, 4.2.4

T. J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 239–245, Fort Collins, Colorado, USA, 1999. 3.2.3, 3.4, 4.2.3, 5.2, 5.2.2, A.2.2

K. Choo and D. J. Fleet. People tracking using hybrid Monte Carlo filtering. In *IEEE International Conference on Computer Vision*, pages 321–328, Vancouver, 2001. 5.2.2

D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of nonrigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 673–678, 2000. 3.2.3, 4.2.3, 4.3.1

T. F. Cootes and C. J. Taylor. Statistical model of appearance for computer vision. Technical report, University of Manchester, 2001. 3.2.3, 4.2.2

R. T. Cox. Probability, frequency and reasonable expection. *American Journal of Physics*, 14:1–13, 1946. 3.2.1

T. Darrell, G. G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, June 2000. 4.2.4

J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through singularities and discontinuities by random sampling. In *IEEE International Conference on Computer Vision*, pages 1144–1149, September 1999. 3.2.3, 5.2.2

J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133, South Carolina, USA, 2000. 4.2.4, 4.3, 5.2.2

J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 669–676, Hawaii, 2001. 5.2, 5.2.2

R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2nd edition, 2001. 3.2.1, 3.3.2, 4.3.1

R. Fablet. Automatic detection and tracking of human motion with a view-based representation. In *European Conference on Computer Vision*, volume I, page 476 ff, 2002. 3.2.3, 4.2.4

P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 66–73, 2000. 5.2.1

R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2003. 4.3.1

D. A. Forsyth and M. M. Fleck. Body plans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–683, Puerto Rico, 1997. 4.2.3, 5.2.1

D. A. Forsyth and M. M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77, August 1999. 5.2.1

W. Frey, M. Zyda, R. McGhee, and B. Cockayne. Off-the-shelf, real-time, human body motion capture for synthetic environments. Technical report, Computer Science Department Naval Postgraduate School, 1995. 2.2

A. Galata, N. Johnson, and D. Hogg. Learning structured behaviour models using variable length markov models. In *IEEE International Workshop on Modeling People*, 1999. 3.2.4

D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999. 3.2.2

D. M. Gavrila and L. S. Davis. 3D model-based tracking of humans in action: A multi-view approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, USA, 1996. 3.2.3, 4.2.2, 5.2.2

R. Gross and J. Shi. The CMU motion of body (mobo) database. Technical report, CMU, 2001. 6.2

M. Grosso, R. Quach, and N. Badler. Anthropometry for computer animated human figures. In *Proceedings of Computer Animation Human Figures*, pages 83–96. Spring Verlag, 1989. 3.2.3, 3.2.3

I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis. Backpack: Detection of people carrying objects using silhouettes. In *IEEE International Conference on Computer Vision*, pages 102–107, 1999a. 3.2.3, 4.2.4

I. Haritaoglu, D. Harwood, and L. S. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *IEEE International Workshop on Visual Surveillance*, page 613, June 1999b. 3.2.3, 4.2.3, 4.2.4

I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, August 2000. 3.2.3, 4.2.4

A. Hilton. Towards model-based capture of a persons shape, appearance and motion. In *IEEE International Workshop on Modelling People*, 1999. 2.2

D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983. 3.2.3, 4.2.2

S. Ioffe and D. A. Forsyth. Human tracking with mixtures of trees. In *IEEE International Conference on Computer Vision*, volume 1, pages 690–695, 2001a. 3.3.2, 4.2.3, 5.2, 5.2.1, 5.3, 5.4

S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43:45–68, June 2001b. 5.2.1, 5.4

M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision*, volume 1, pages 343–356, Cambridge, 1996. 3.2.3, 4.2.2, 5.2, 5.2.2, A.2.2

E. T. Jaynes. Bayesian methods: A general background. *Maximum Entropy and Bayesian Method in Statisitics*, pages 1–25, 1986. 3.2.1

N. Jojic, J. Gu, H. C. Shen, and T. S. Huang. Computer modeling, analysis, and synthesis of dressed humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 528–534, Santa Barbara, CA, 1998. 2.2.3

S. Ju, M. Black, and Y. Yacoob. Cardboard people: a parameterized model of articulated image motion. In *IEEE International Conference on Face and Gesture Recognition*, pages 38–44, Killington, VT, USA, 1996. 3.2.3, 4.2.4

I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–87, San Francisco, USA, 1996. 3.2.3

R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 5.2.2

I. Karaulova, P. Hall, and A. Marshall. A hierarchical model of dynamics for tracking people with a single video camera. In *British Machine Vision Conference*, pages 352–361, Bristol, 2000. 3.2.3

M. Kass, A. Withkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988. 3.2.3

R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995. 3.3.2

S. Konishi, A.L. Yuille, J.M. Coughlan, and S.C. Zhu. Statistical edge detection: learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):57–74, January 2003. 4.2.2, 4.2.2

M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 359–377, April 1995. 3.2.3

T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–156, November 1998. 4.2.3

S. Loncaric. A survey of shape analysis techniques. In *Pattern Recognition*, pages 983–1001, 1998. 3.2.3

J.P. MacCormick and A. Blake. A probabilistic contour discriminant for object localisation. In *IEEE International Conference on Computer Vision*, pages 390–395, 1998a. 3.2.3, 3.3.2, 4.2.2

J.P. MacCormick and A. Blake. Spatial dependence in the observation of visual contours. In *European Conference on Computer Vision*, 1998b. 3.2.3, 4.2.2

J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, June 2001. 4.2.2

D.R. Martin, C.C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, May 2004. 4.2.2, 4.3

P. Maybank. *Stochastic Models, Estimation and Control*. Academic Press, 1979. 5.2.2

S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, October 2000. 3.2.3, 4.2.3, 4.2.4, A.2.2

D. Metaxas and D. Terzopoulos. Shape and non rigid motion estimation through physics based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):580–591, June 1993. 3.2.3

M. Mikic, I. Trivedi, E. Hunter, and P. Cosman. Articulated body posture estimation from multi-camera voxel data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 455–460, 2001. 3.2.3, 4.2.4

T. B. Moeslund and F. Bajers. Summaries of 107 computer vision-based human motion capture papers. Technical Report LIA99-01, University of Aalborg, 1999. 3.2.2

T. B. Moeslund and E. Granum. 3D human pose estimation using 2D-Data and an alternative phase space representation. In *IEEE Workshop on Human Modeling, Analysis and Synthesis*, 2000. 3.2.3

T. B. Moeslund and E. Granum. A survey of computer vision-based human motion

capture. *Computer Vision and Image Understanding*, 81(3):231–268, March 2001. 1, 3.2.2

G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, pages 666–680, 2002. 3.2.3

D. D. Morris and J. M. Rehg. Singularity analysis for articulated object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 289–296, 1998. 3.2.3, 3.4, 5.2.2

Eadweard Muybridge. *The Human Figure in Motion*. Dover, 1989. 5.2.1

R. Okada, Y. Shirai, and J. Miura. Tracking a person with 3D motion by integrating optical flow and depth. In *IEEE International Conference on Face and Gesture Recognition*, pages 336–341, Grenoble, 2000. 4.2.4

E. Ong and S. Gong. A dynamic human model using hybrid 2D-3D representations in hierarchical PCA space. In *British Machine Vision Conference*, pages 33–42, Nottingham, 1999. 3.2.3

C. Oren, M. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–199, 1997. 3.2.3

D. Ormoneit, H. Sidenbladh, and M. Black. Tracking human motion using functional analysis. In *IEEE Workshop on Human Modeling*, Hilton Head, SC, USA, June 2000. 3.2.4

J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):522–536, November 1980. 3.2.1, 3.2.3

J. Park, O. Hwang-Seok, D. Chang, and E. Lee. Human posture recognition using curve segments for image retrieval. In *SPIE Conference on Storage and Retrieval for Media Databases*, volume 3972, pages 2–11, 2000. 4.2.3

V. Pavlovic and J. M. Rehg. Impact of dynamic model learning on classification of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 788–795, 2000. 3.2.4

V. Pavlovic, J. Rehg, and J. MacCormick. Learning switching linear models of human motion. Technical report, Compaq, 1999a. 3.2.4

V. Pavlovic, J. M. Rehg, T. J. Cham, and K. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *IEEE International Conference on Computer Vision*, pages 94–101, 1999b. 3.2.4

A. P. Pentland and B. Horowitz. Recovery of non-rigid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):730742, July 1991a. 4.2.4

A.P. Pentland and B. Horowitz. Recovery of nonridid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):730–742, July 1991b. 3.2.3

R. Plankers. *Human Body Modelling From Image Sequences*. PhD thesis, EPFL, Switzerland, 2001. 3.2.3

R. Plankers and P. Fua. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding*, 81(3):285–302, March 2001. 3.2.3

R. Plankers and P. Fua. Articulated soft objects for multi-view shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (10):1182–1187, 2003. 3.2.3

R. Polana and R.C. Nelson. Low level recognition of human motion. In *Workshop on Non-Rigid Motion and Articulated Objects*, 1994. 3.2.3

J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *IEEE International Conference on Computer Vision*, pages 1165–1173, 1999. 4.3.1, 4.3.1, A.2.2

D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 467–474, Madison, Wisconsin, June 2003. 4.2.3, 5.2.1

J.M. Rehg and T. Kanade. Model based tracking of self occluding articulated objects. In *IEEE International Conference on Computer Vision*, pages 612–617, 1995. 3.2.3, 3.5.3

R. Ronfard, C. Schud, and B. Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714, Copenhagen, 2002. 4.2.2, 5.3

R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 721–727, 2000. 3.2.3

R. Rosales, M. Siddiqui, J. Alon, and S. Sclaroff. Estimating 3D body pose using uncalibrated cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 821–827, 2001. 3.2.3

M.A. Ruzon and C. Tomasi. Color edge detection with the compass operator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 160–166, 1999. 4.2.2

B. Schiele and J. L. Crowley. Recognition without correspondence using multidi-

mensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000. 4.3.1, 4.3.1

H. Segawa, N. Hiraki, H. Shioya, and T. Totsuka. Constraint-conscious smoothing framework for the recovery of 3D articulated motion from image sequences. In *IEEE International Conference on Face and Gesture Recognition*, 2000. 5.2.2

A. Shahrokni, T. Drummond, and P. Fua. Texture boundary detection for real-time tracking. In *European Conference on Computer Vision*, volume II, pages 566–577, 2004. 4.2.2, 4.3

H. Sidenbladh. Detecting human motion with support vector machines. In *International Conference on Pattern Recognition*, 2004. 3.2.3

H. Sidenbladh and M. J. Black. Learning image statistics for Bayesian tracking. In *IEEE International Conference on Computer Vision*, volume 2, pages 709–716, Vancouver, 2001. 3.3.2, 4.2.2, 4.2.3, 4.3, 4.3.1

H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *European Conference on Computer Vision*, pages 702–718, Dublin, 2000a. 3.2.1, 3.2.4, 4.2.4

H. Sidenbladh, F. de la Torre, and M. J. Black. A framework for modeling the appearance of 3D articulated figures. In *IEEE International Conference on Face and Gesture Recognition*, pages 368–375, Grenoble, 2000b. 4.2.3

C. Sminchisescu. Consistency and coupling in human model likelihoods. In *IEEE International Conference on Face and Gesture Recognition*, pages 27–32, Washington, 2002. 4.2.4

C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 447–454, Hawaii, 2001. 3.2.3, 4.2.4, 5.2

C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 69–76, 2003. 5.2.2

Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 810–817, Hawaii, 2000. 4.2.4

C. Stauffer and E. Grimson. Similarity templates for detection and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 221–228, 2001. 4.2.3, 4.3

K. Tabb, N. Davey, R. Adams, and S. George. Analysis of human motion using snakes and neural networks. In *Proceedings of the Joint IAPR Workshop on Articulated Motion and Deformable Objects*, pages 48–57, 2000. 3.2.3

C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80:349–363, September 2000. 3.2.3, 5.2.2

P. Tissainayagam and D. Suter. Visual tracking of multiple objects with automatic motion model switching. In *International Conference on Pattern Recognition*, volume 3, pages 1146–1149, 2000. 3.2.4

M. Turk and A.P. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991. 4.2.3

M. Varma and A. Zisserman. Texture classification: Are filter banks neccessary? In *IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 691–698, 2003. 6.2

S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences.

Computer Vision and Image Understanding, 74(3):174–192, June 1999. 3.2.3, 4.2.2, 4.2.3, 4.3.1, 4.3.1, 5.2, 5.2.2

J. Wilhelms, A. van Gelder, L. Atkinson-Derman, and A. Luo. Human motion from active contours. In *IEEE Workshop on Human Motion*, pages 155–160, 2000. 3.2.3

C. R. Wren, A. Azarbayejani, T. J. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997. 3.2.3, 4.2.3

J. Zhao and N.I. Badler. Inverse kinematics positioning using nonlinear programming for highly articulated figures. *ACM Trans. Graph.*, 13(4):313–336, 1994. ISSN 0730-0301. doi: 10.1145/195826.195827. 3.2.3

T. Zhao, R. Nevatia, and F. Lv. Segmentation and tracking of multiple humans in complex situations. In *IEEE Conference on Computer Vision and Pattern Recognition*, page II:194201, 2001. 3.2.3, 4.2.4

T. Zhao, T. Wang, and H. Shum. Learning a highly structured motion model for 3D human tracking. In *Asian Conference on Computer Vision*, Melbourne, 2002. 3.2.4