

Human Pose Estimation using Partial Configurations and Probabilistic Regions

Timothy J. Roberts*, Stephen J. McKenna† and Ian W. Ricketts‡

April 4, 2006

Abstract

A method for recovering a part-based description of human pose from single images of people is described. It is able to perform estimation efficiently in the presence significant background clutter, large foreground variation, self-occlusion and occlusion by other objects. This is achieved through two key developments. Firstly, a new formulation is proposed that allows partial configurations, hypotheses with differing numbers of parts, to be made and compared. This permits efficient global sampling in the presence of self and other object occlusions without prior knowledge of body part visibility. Secondly, a highly discriminatory likelihood model is proposed comprising two complementary components. A boundary component improves upon previous appearance distribution divergence methods by incorporating high-level shape and appearance information and hence better discriminates textured, overlapping body parts. An inter-part component uses appearance similarity of body parts to reduce the number of false-positive, multi-part hypotheses, hence increasing estimation efficiency. Results are presented for challenging images with unknown subject and large variations in subject appearance, scale and pose.

1 Introduction and Motivation

Human pose estimation, if reliable and efficient, could form the basis of many important applications. In addition to being necessary for automated analysis when only images are available as input,

computer vision also provides a compelling alternative to other sensing modalities in terms of cost, intrusiveness, accuracy and reliability. These often competing requirements make it unlikely that a ‘one size fits all’ approach will be successful. Here we focus upon the largely ignored problem of automatic estimation of the transformation and visibility of a set of body parts from highly unconstrained *single* images. In particular, we introduce the partial configuration formulation that allows pose hypotheses with varying numbers of visible parts to be made and compared. This part-based approach can be contrasted with lower detail pose parametrisations such as global position and scale or a body contour, which are popular for highly unconstrained applications such as surveillance and smart environments. The advantage of a part based approach is that occlusion can be explicitly modelled and efficient part based global search techniques can be employed. The partial configuration approach is also applicable to scenes containing occlusion of people by other objects. It can also be contrasted with highly detailed pose descriptions, such as 3D surface structure, which occur in certain medical and professional sports analysis applications where the emphasis on accuracy and detail usually results in a more costly, intrusive system that requires highly constrained environments and off-line operation in order to be beneficial. This focus on a medium level of detail and highly unconstrained images is arguably the most promising in terms of future, large scale applications such as computer games, virtual reality and high bandwidth human computer interfaces and would allow more detailed automated interpretation of pose from pre-existing images.

The majority of current pose estimation methods

*troberts@computing.dundee.ac.uk

†stephen@computing.dundee.ac.uk

‡ricketts@computing.dundee.ac.uk

make strong assumptions regarding the background scene, subject’s appearance (clothing), viewpoints, temporal dynamics, self-occlusion and occlusion by other objects [40, 38, 16]. Indeed, in spite of considerable research into human tracking, most tracking systems remain limited to constrained scenes and rely upon strong temporal pose constraints and therefore manual (re)initialisation. In contrast, this paper focuses on the task of estimating human pose from *single* real-world images of poorly constrained scenes. This is clearly a more challenging task since temporal information is absent. Furthermore, given images containing restrictive, partial information due to occlusion, the aim is to estimate the body pose of the visible portion. It is assumed that the system is not required to recover personal metric details such as body sizes.

Most existing pose estimation systems follow an *analysis by synthesis* paradigm in which pose models are hypothesised and compared to images. Bayesian probability often forms the basis of such systems for a number of reasons. Firstly, the probabilistic logic provides a coherent framework for modelling the inherent uncertainty in the image and system. Secondly, Bayes rule allows the dependency between model and data to be reversed and thus the principled use of this paradigm. Thirdly, probability theory allows the construction of an efficient system since decision making can proceed in the presence of limited data and assumptions. Finally, Bayesian statistics allows additional, perhaps subjective, prior information such as pose constraints to be incorporated in a principled manner, something which is particularly important in light of the complex appearance of people in images. Many human tracking systems build upon this spatial framework by assuming a Markov relationship between frames and thereby obtain a temporal prior. Taken together these approaches are essentially the probabilistic manifestation of the model-based architecture proposed early on by O’Rourke and Badler [45].

In this Bayesian framework, components that must be developed in order to construct such a system are (i) a pose model that, by incorporating prior knowledge, describes the variation of human shape, (ii) a likelihood model that discriminates incorrect pose hypotheses from correct ones (i.e. those that correspond to people) based upon image measure-

ments, and (iii) a computationally feasible estimation scheme that searches for probable pose hypotheses. The aims of the work described here are:

1. Efficient discrimination of a single person with a complex, unknown appearance from a cluttered, unknown scene that possibly occludes body parts.
2. The development of a formulation that allows efficient, accurate global estimation in such conditions.

This emphasis on the formulation and likelihood is central to the spirit of this and previous work [50] and distinguishes it from other research that usually concentrates on prior constraints and efficient search techniques.

The paper is organized as follows. Section 2 reviews the most relevant previous work. Section 3 describes the pose estimation problem to be addressed, making explicit certain operational assumptions. Sections 4 and 5 describe the overall formulation and the likelihood model. An empirical investigation of the likelihood is presented in Section 6. Section 7 describes a pose estimation scheme and applies it to real-world images. Finally, Section 8 discusses the system as a whole, summarises its advantages and limitations and suggests some directions for future research.

2 Previous work

A sizeable literature exists that is concerned, in some way or another, with the estimation of human pose from images. Exhaustive coverage would be inappropriate here. Instead, work most relevant to the formulation of pose estimation systems and likelihood models is reviewed. Much of the literature is concerned with human tracking rather than estimation from a single image but certain components, such as likelihood models, are still relevant here.

2.1 Formulation

Body modelling for pose estimation has been formulated using a variety of approaches. Some concentrate on modelling the appearance in the image whilst others explicitly model the physical 3D

structure and acquisition system. Some model the body as a whole whilst others decompose the body into component parts.

Image contour models and silhouettes are particularly relevant in applications where the background is known or can be estimated. For example, silhouettes have been used to recover body parts from walking sequences [20]. An active contour model with interactive correction has been used to recover a 3D body model [64]. Baumberg and Hogg [3] used contour modes of variation to track pedestrians. Contour shape models have also been combined with 2D part models to estimate 3D pose [4] and over multiple viewpoints [44]. Such hybrid models help disambiguate self-occluding poses such as the hands moving over the body and thereby resolve a key disadvantage of the contour representation. Rosales *et al.* [52, 53] inferred possible 2D joint locations from Hu moments of silhouettes using a learning architecture that implicitly represented the part-based nature of the human body. Given multiple views, expectation-maximisation was used to find the most consistent 3D pose and the associated views.

In order to tackle the problem of contour *detection*, MacCormick and Blake [31] proposed a probabilistic discriminant based upon a likelihood ratio of edge measurements due to foreground and background clutter. This approach has similarities to the one adopted in this paper. When combined with importance sampling it allowed global sampling of an image containing low-dimensional targets (head and shoulders). This partially addressed the problem of automatically initialising contour models but was not practical for more varied objects. The approach was extended to discriminate occlusion events from weak measurements [32].

The structure of the human body is well understood and anthropometric data are readily available [18]. A part-based description is natural since it corresponds to our rigid bone structure. The shapes of individual parts are commonly modelled using geometric primitives. Cardboard people models use 2D rectangular patches [25, 6]. Others have used 2D ribbons [30] or 2D elliptical blobs [65]. Early work by Hogg [21] used 3D cylindrical primitives. Others have subsequently used cones with elliptical cross-sections [63], super-quadrics [36] or tapered super-quadrics [26,

17]. The shape parameters of these geometric part primitives are often fitted manually or to specific cases using multiple views (e.g. [26]). Inter-subject variability is usually ignored. A notable exception is [59] whose detailed, high-dimensional, human model used anthropometric data to specify a prior on part sizes. However, intra-scene variation due to non-rigid deformation and clothing motion was not addressed.

It is common to consider the body as a tree structure, with the torso as the root, and to chain the model-to-image transforms hierarchically, thereby capturing the kinematic structure of the human body. Various parameterisations of 3D part transformations have been proposed. Hierarchical 3D transformations with rotations parameterised using Euler angles have been used [63] but suffer from singularity problems which arise when changes in state become unobservable. A 2D scaled prismatic parameterisation avoids such problems [6, 43]. The 3D twists and exponential map formulation used in robotics linearises the kinematics and removes the singularity problems [5]. Such problems can also be avoided by using a random sample estimation scheme such as particle filtering that does not rely upon local derivatives [11].

More flexible joint representations than simple relative orientation have been used with a resulting increase in state dimensionality. Relative translation of parts constrained by spring forces has been used for the shoulder. Alternative ball and socket joint parameterisations have been investigated [1]. A phase space representation was considered for the arm [39]. Finally, part-based transformations also allow existing inverse kinematics techniques to be applied [66].

Three-dimensional part models account for self-occlusion by representing depth and using hidden surface removal. In order to account for self-occlusion with 2D part models, depth ordering can be used [49]. It is possible to track through self-occlusions without predicting them explicitly by propagating multiple hypotheses [6]. Furthermore, even with a detailed 3D model it can be difficult to describe the appearance of partially occluded parts. This prompted work on actively choosing viewpoints from which to compute the likelihood based upon part visibility [26]. One significant advantage of physically motivated 3D models is that

hypotheses can easily be related between multiple views [17]. However, the focus of this paper is on monocular estimation. There are problems inherent in uncalibrated monocular estimation of 3D pose even when 2D joint point locations are *known* [62, 2, 41].

Part-based models 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 [27]. Physically motivated 3D models also allow constraints based upon part interpenetration to be expressed [59].

In conclusion, a part-based representation is natural given our prior knowledge of human body structure. In comparison to contour-based models, using this prior knowledge on the form of the body removes much of the burden of learning highly varied shape models, implicitly accounts for non-linear changes resulting from self-occlusion and allows easier encoding of pose constraints. Furthermore, a part-based parameter space is easier to interpret and allows constraints to be encoded more easily than contour descriptions. In comparison to 3D 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.

2.2 Likelihood

Due to the complexity and variation of human appearance, building a general but discriminatory likelihood model is difficult and still a topic of active research. Various likelihood models have been proposed. Those based on cues such as optical flow and background models are not mentioned because their use is excluded by the problem characteristics outlined in Section 3. It should be noted that strongly discriminatory likelihood models are not as important for trackers using temporal constraints on typically short, manually initialised sequences as they are for the global pose estimation problem considered here.

Any likelihood model should be evaluated within the context of the entire pose estimation system. Nevertheless two desiderata can be identified:

(i) it should be highly discriminatory, especially since the number of incorrect instances is much larger than the number of correct instances, and (ii) it should allow efficient sampling and search techniques to be applied. These goals can be competing. From the point of view of discrimination the ideal likelihood model would be a delta function on the correct model configuration. In reality, the likelihood is usually multi-modal and complex. Furthermore, there is a tradeoff between model generality and discrimination. For example, better discrimination becomes possible when prior knowledge of foreground appearance is available. However, a key problem in human pose estimation is that there is in general limited information available regarding the foreground and background appearance.

Two categories of likelihood model can be identified: (i) those that are based upon differences in appearance of the foreground and background around the boundary, and (ii) those that model the appearance of object foreground.

2.2.1 Boundary Models

Likelihood models based upon appearance differences across the model boundary are popular for human tracking since they can exhibit a good degree of invariance to changes in foreground and background appearance. Early work by Hogg [21] used a threshold on the magnitudes of Sobel filter responses to detect edges and projected model boundary segments were then inspected to find edges within a specified distance and relative orientation. Gavrilu and Davis [17] used a similar approach but employed a robust variant of a chamfer distance transform computed from detected edges in order to provide a smooth, broad search function. Wachter and Nagel improved on this approach by matching model edges directly to filter responses and actively selecting strong model candidates based on the overlap with similar parts [63]. Furthermore, due to the limitations of intensity edge cues in real world images, a foreground template was also employed to stabilise tracking.

A different approach to boundary modelling involves casting model normals and inspecting gradients along these ‘measurement lines’. McCormick and Blake [31] developed a probabilis-

tic formulation of this approach based on modelling distributions of features on the foreground and background. In order to deal with occlusion, which is manifested as groups of weak boundary measurements, this scheme was extended to incorporate a Markov random field learnt from previous occlusion instances [32].

Konishi *et al.* [28] emphasised the importance of both principled statistical modelling and describing filter responses from the non-boundary edges. To accomplish this, ground truth segmentations were used to learn the probability density functions (PDFs) of multi-scale filter responses both on and *off* object boundaries. Then a ratio between these PDFs, the likelihood ratio, was used to provide a nonlinear 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 [57] applied this approach to humans by learning PDFs of intensity edge features for points around human boundaries and for points on the background. Likelihood ratios were then combined by assuming independence. However, in contrast to previous work, it was 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. A conclusion of their work that is relevant here is that statistical models of colour and texture would improve results.

In contrast to the above approaches that compare model projections to simple local filters, Ronfard *et al.* [51] trained 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). However, this system was unable to account for self-occlusion. 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).

Much of the work relating to boundary detection relates to the gradient of intensity images. Human clothing however is usually colourful and making use of this information should improve discrimination. Furthermore, clothing is usually textured

which requires more complex models of boundaries. A generic approach to colour edge detection is provided by the compass operator (Ruzon and Tomasi [54]). Here the divergence between colour distributions either side of a circle’s oriented bisector is mapped, heuristically, to edge strength. Martin *et al.* [34] improved on this approach by combining intensity, colour and texture features and by learning the mapping from ground truth segmentations in a similar manner to the work of Konishi *et al.* described above. In particular, colour and texture gradients were formulated using the χ^2 measure between colour and texture [33] distributions either side of the boundary. Although this approach, whereby the image is filtered before fitting the high level geometric model, provided good performance in a statistically sound formulation it is unable to account for large scale texture changes often present in human clothing. The author comments the performance of localised bottom-up boundary models at finding boundaries in real-world images is significantly lower than human performance. This is not surprising given that texture can occur over large scales and that the high human performance depends upon having large regions either side of the boundary. Mori *et al.* [42] combined these statistical boundary detection methods with a normalised cuts segmentation scheme. Some of the resulting regions correspond to salient parts that ‘pop out’ and can drive pose estimation. When assembling poses from these salient parts inter region appearance is also used.

Research pertaining to ‘top down’ colour and texture boundary models has been surprisingly limited, especially for human pose estimation. Shahrokni *et al.* [56] used zero and first order Markov processes along measurement lines 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 limitations of intensity edge cues. The formulation allowed fast local tracking. However, using texture on measurement lines assumes that the texture can be described by this line, which can be violated when the surface undergoes non-rigid deformation, for example. In this paper, a boundary likelihood model is introduced at the level of body parts, along with a model of inter-part simi-

larity in order to improve localisation of large scale textures

2.2.2 Foreground Models

A description of the absolute appearance is usually unavailable, primarily due to the variability in clothing. In tracking scenarios, foreground appearance can be assumed to be known from either manual initialisation [6, 63] or previous frames but will vary due to lighting changes and clothing motion.

One case when absolute appearance is known with some certainty is skin although this is sensitive to illumination and is often only applicable to the head and hands. In relation to human pose estimation, template matching and texture distributions learnt off-line have been used to detect naked people, their limbs and torsos [14, 22]. Park *et al.* [46] used a segmentation scheme and then applied a colour classifier to detect skin coloured body parts. Pfister used clusters in colour-position space to find the head and hands in a real-time implementation [65]. However, foreground regions were first segmented using a background model. Face detection methods have been combined with other likelihood modules to improve performance. For example, [37] represented the face and upper body using local gradient orientation features.

In the case when clothing appearance information is available, Sidenbladh *et al.* [58] proposed learning a linear subspace representation of the surface texture for a particular subject from a set of views. 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. Regions on the surface were weighted by visibility. This model allowed rotation about the limb's major axis to be recovered if the limb's surface had distinguishing non-symmetric features such as an emblem. However, the appearance had to be learned off-line and was subject-specific. In contrast to this off-line appearance estimation, Ramanan and Forsyth [48] used motion, appearance consistency and kinematic constraints to find a colour representation of the foreground appearance of individual limbs automatically before tracking.

Human appearance has other properties that can be used to discriminate it from the background.

Sidenbladh *et al.* [57] learnt the distributions for correct and incorrect poses of ridge features formulated in terms of second derivative filters at the scale of the body part. A clear example of the discriminatory power of larger scale foreground structure is provided by similarity templates, concatenations of relationships between pairs of pixels in an image window [61]. These have been used for pedestrian detection but relied upon limited variation in pose and were slow to evaluate. Relationships between pairs of foreground features have not been used for more 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 paper, the similarity between the appearance of opposing body parts is used to improve discrimination of larger configurations and thereby constrain the estimation.

Finally it is important to note that foreground appearance models are of independent interest for tracking specific people through occlusions and group interactions [35, 19].

2.3 Estimation

Current human pose estimation methods can be broadly categorised as either combinatorial or full state space search. The combinatorial approach is adopted in most single image pose estimation systems and consists of finding candidate body parts and then grouping these as determined by inter-part constraints. This approach usually makes assumptions regarding the form of the optimisation function in order to efficiently combine the parts. In contrast, full state space search approaches attempt to find all possible body parts simultaneously and make no such assumptions. However, they are usually only applicable when strong prior information regarding pose is available, such as a motion estimate in a tracking scenario. Another point of differentiation is that of solution representation, some pose estimation systems estimate a single pose and a local estimate of uncertainty (e.g. [63, 22]) whilst others estimate multiple solutions, modelled either semi-parametrically [6] or non-parametrically as particle sets [59, 10].

A good general example of the combinatorial approach is pictorial structures [13] which have been

applied to human pose estimation in indoor scenes. 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. The solution is a MAP estimation of pose.

The combinatorial approach to human pose estimation has since been most actively developed by Forsyth *et al.* Early work described the use of a hierarchical grouping scheme, called a body plan, for detecting of naked humans and animals [14, 15]. These body plans employ a sequential classification approach 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, individual part detections with many false positives were fed into a pair-wise classifier which in turn fed into a three-part classifier, each stage further pruning the candidates. The topology of the classification network is fixed prior to learning the individual classifiers. A sequential classification approach was also described by Ioffe and Forsyth [23] but relied on binary classification of limbs rather than estimating likelihoods. A new approach to estimation was then proposed based upon drawing assemblies proportional to a likelihood of the full (fixed size) assembly. In order to efficiently draw samples from this likelihood, a set of marginal likelihoods was proposed and assumed to be independent of other parts (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 parts to be distinct, this model no longer has a tree structure and cannot be inferred using dynamic programming. In later work the single tree model was criticised as a representation due to the inability to represent self-occlusion (a frequent, significant phenomenon in human pose estimation). A mixture of trees representation was proposed to address these deficiencies without resort to full state space search [22].

[37] learnt a set of classifiers for combinations

of frontal and profile upper body views and frontal leg views. As expected using joint classification of parts greatly improved performance. Mori *et al* [42] used salient parts identified from a bottom up segmentation scheme to drive a combinatorial pose estimation method. Whilst this system provided good results on the cropped sports test images it is not clear how successful the approach will be with larger uncertainty in scale.

The second estimation approach, that of searching the full dimensional state space makes no assumptions regarding inter-part relations such as self-occlusion and is the most popular approach for human tracking where a temporal prior is available. However, this approach usually requires manual initialisation and re-initialisation upon failure to be computationally feasible. It is therefore only considered briefly. One set of approaches uses either local gradient (e.g. [63]) or hierarchical-best-first search schemes (e.g. [17]). Whilst efficient, such local approaches are not feasible when significant occlusion occur, fast irregular motions are present or large amount of background clutter is present. Furthermore, such models can have problems with singularities [43, 11]. Cham and Rehg [6] overcome these problems by recovering the multiple hypotheses and tracking these using local Gauss-Newton search. An alternative approach is that of sampling importance resampling and a non-parametric particle representation such as Condensation [24], whereby samples are drawn from the temporal prior for diffusion and resampling. However, blind application of such methods requires a large number of particles for good results. Therefore, many techniques have been proposed to increase efficiency by taking advantage of the structure of the human form. For example, Deutscher *et al* [9] used ideas from simulated annealing to more reliably estimate the global structure of the posterior distribution. The formulation resulted in an automatic soft partitioning of the search space and used genetic algorithm-style cross-over operators to take advantage of the (semi-) hierarchical nature of the problem [10]. Choo and Fleet [7] applied the hybrid Markov Chain Monte Carlo (MCMC) scheme whereby a local potential is defined that speeds acceptance without biasing the sampling behaviour. In addition, multiple Markov chains were used to efficiently explore the multi-modal poste-

rior. [29] used a data driven MCMC scheme to estimate the pose of the upper body. Sminchisescu *et al* developed a set of novel sampling schemes that explicitly model characteristics of the distributions found in monocular human tracking. One such technique, Covariance Scaled Sampling [59], samples deeply along the directions of largest uncertainty since in monocular tracking, the authors reason, it is in these directions that alternative maxima are likely to be found. Another algorithm, Kinematic Jumping [60], was proposed to flip the orientation of parts in order to escape local maxima that occur in monocular estimation. Such local maxima are related to poses that cannot be reliably discriminated even when joint positions are known [62].

3 Image and Scene Characteristics

To begin the discussion of the work presented in this paper the assumptions regarding the image and scene are made explicit. **Subject.** A single person is assumed to be present in an unknown pose that is to be recovered. The appearance (e.g. clothing) of this person is unknown. Clothing can be loose fitting and therefore have a complex outline. Furthermore, clothing can be textured in complex ways and at many scales. **Scene.** Images are of indoor or outdoor scenes. These scenes have unknown structure and can contain clutter at similar scales to the human body. Different types of scene lead to differences in the shape of typical objects that are visible (and presumably also differences in the pose of the person). Objects in the scene can have a textured appearance. **Occlusion.** In addition to self-occlusion, the person might be partially occluded by other objects in the scene, a key difficulty in visual pose estimation. **Viewpoint.** It is assumed that the scale of the person 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. **Perspective.** It is assumed that perspective effects are weak. In particular, it is assumed that these effects are small when compared to intra- and inter-person shape variability. Perspective effects could be modelled if the intrinsic camera parameters were known. **Modality.** Since the majority of existing images are in

colour the focus is upon this modality (instead of monochromatic, range or infrared modalities). The colour signal can be noisy. **Illumination.** The images can have complex illumination from multiple sources. Illumination might be so poor that certain body parts cannot be distinguished based on local visual appearance. Strong cast shadows and self-shadowing can occur.

4 Formulation

A central thesis of this work is that the limitations of current pose estimation systems when applied to real world images are symptomatic of a poor pose representation and that these limitations cannot be resolved efficiently, if at all, by simply improving the likelihood model and estimation scheme. As discussed in the review, the structure of the human body can be naturally described using part-based decomposition. The part parameters used here correspond to a level of detail suitable for many applications where the input is monocular, cluttered and low quality. In common with previous part based approaches our formulation has an easy to interpret parameter space, models self-occlusion and allows constraints on the body to be encoded. However, in contrast to previous part-based approaches the approach described here does not constrain the number of parts or rely upon knowledge of the part visibility prior to estimation. The key point here is that although using a fixed number of body parts 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* and *greater efficiency*. 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 in the same manner as self-occlusion.

4.1 Partial Configurations

A key element of the formulation developed here, coined partial configurations, is that the number of hypothesised body parts can vary. Possible partial configurations include single part hypotheses, full body hypotheses and everything in between. A partial configuration includes pose hypotheses

for some non-empty subset of the set of a person’s body parts. Its parameter space dimensionality varies with the number of parts in the configuration. It should be emphasised that the posterior distribution of partial configurations does not treat the parts independently (both the likelihood and prior constraints use pairwise part potentials). A configuration will be denoted \mathbf{C} and its parts indexed by i or j .

Clearly, for this approach to be useful it must be possible to compare partial configurations of differing dimensionality. Moreover, larger correct hypotheses should be preferred to smaller correct hypotheses. Consider how hypotheses in a fixed size state space are usually compared. The most popular approach is to find the maxima of the posterior $p(\mathbf{C}|\mathbf{I})$ where \mathbf{I} denotes the image. It usually suffices to compute the likelihood and prior, ignoring 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). Computing the normalising factor, the evidence, for each combination of parts and thereby computing posterior probabilities that can be compared is not computationally feasible.

Instead, the problem is treated as one of discriminating between people and background *at each point in the state space*. The state space is therefore augmented by a class label $v \in \{0, 1\}$ that labels the hypothesis as either for a person ($v = 1$) or for a background process ($v = 0$). An optimum classification for a particular pose is found by choosing the class with the highest probability (assuming uniform risk) [12]. This is equivalent to forming the logarithm of the posterior ratio, ρ , given in Equation (1) and classifying hypotheses as people when $\rho > 0$ and as background otherwise.

$$\begin{aligned} \rho &= \ln \frac{p(\mathbf{C}, v = 1|\mathbf{I})}{p(\mathbf{C}, v = 0|\mathbf{I})} \\ &= \ln \frac{p(\mathbf{I}|\mathbf{C}, v = 1)}{p(\mathbf{I}|\mathbf{C}, v = 0)} + \ln \frac{p(\mathbf{C}, v = 1)}{p(\mathbf{C}, v = 0)} \quad (1) \end{aligned}$$

Posterior ratios allow 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 estimation efficiency, the key point is that whilst allowing the number of parts to vary greatly increases the number of possible pose hypotheses, the relationship between configurations can be used to disproportionately increase sampling efficiency and perform global estimation. In general, configurations with large numbers of parts are more strongly discriminated from the background than configurations with small numbers of parts (although finding larger configurations can be more difficult). Therefore, large correct configurations tend to have higher posterior ratios than small correct configurations. This is due to the structured form of appearance and pose of people which is modelled using inter-part pose and appearance constraints.

4.2 Modelling Part Pose

Partial configurations can be applied to 2D or 3D part parameterisations. A depth-ordered 2D model is adopted here since the application is to monocular data and this avoids the genuine ambiguities that exist with 3D monocular estimation and gives a more compact state space, both of which ease estimation. As mentioned previously it is assumed that the uncertainty in 2D shape due to limited perspective effects and 3D shape variation is comparable to the uncertainty due to clothing and intra-personal variability.

We first emphasise that the partial configurations formulation is non-hierarchical: there is no single root body part and parts in a kinematic chain are often missing. Although this removes the automatic kinematic behaviour, it is not clear whether this behaviour eases estimation anyway. Whilst it might be argued that a non-hierarchical representation results in a higher dimensional state space it is worth noting that complex joints like the shoulder cannot be adequately modelled using relative orientation alone and pose models often use relative translation to allow a better fit.

The transformation of the 2D body part model into image space is a restricted affine transformation and is similar to the scaled prismatic parameterisation [6, 43]. However, in this system the parts share a common scale parameter. Equation (2) gives the transformation, $T_i(\mathbf{x})$, for a point on the

i^{th} part. The interpretation of this transformation is straightforward: a part model is translated so that its centre is at (a_i, b_i) and then rotated by θ_i in the image plane. An extension, denoted by e_i , is applied to model rotation out of the image plane and the part is scaled by a common scale factor s .

$$T_i(\mathbf{x}) = s \begin{bmatrix} \cos \theta_i & e_i \sin \theta_i \\ -\sin \theta_i & e_i \cos \theta_i \end{bmatrix} \mathbf{x} + \begin{pmatrix} a_i \\ b_i \end{pmatrix} \quad (2)$$

4.3 Probabilistic Region Templates for Parts

The use of *ad hoc* geometric primitives as part models limits generality and does not address uncertainty due to inter-person variability, intra-person variability, clothing and non-rigid deformation. However, modelling such variation explicitly is not necessary in order to solve for the desired pose description. *Probabilistic region templates* are proposed here as an alternative. They encode shape uncertainty explicitly and do not make hard distinctions between the foreground and background of hypothesised parts. In addition to the variations identified above, part shape uncertainty is also introduced by un-modelled perspective effects and 3D shape variation. In general, shape model uncertainty is due to un-parameterised variation (in contrast to pose uncertainty which is inherent in the problem).

A part’s shape is represented as a probabilistic region template, denoted \mathbf{M}_i , computed from aligned training data by estimating a mean image from manually specified binary segmentation images (see Fig. 2). Each point in \mathbf{M}_i represents the probability of that point being on the part. These non-parametric templates thus encode shape uncertainty based upon marginalisation over un-parameterised shape variation. Opting not to parameterise a degree of freedom, and marginalising over it, reduces the size of the state space. For convenience, probabilistic region templates are treated as having infinite extent although they are illustrated and implemented as finite masks that include all the non-zero probabilities.

Specifically, probabilistic region templates were estimated from manually specified part segmentations aligned using the 2D transformation of Equa-

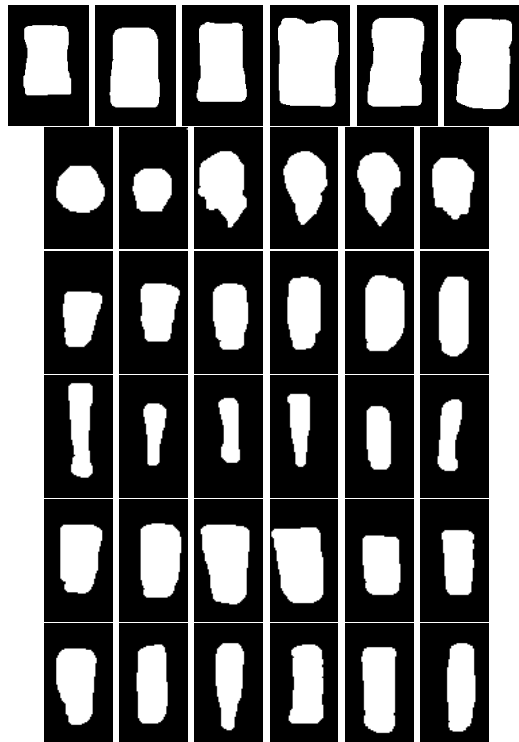


Figure 1: Examples of manually segmented part foreground. The rows correspond to torsos, heads, upper arms, lower arms, upper legs and lower legs.

tion (2). Limb training examples were extracted from fully extended parts (i.e. with the major axis of the part approximately parallel to the image plane). Fig. 1 illustrates typical segmentations. Twenty training examples were used for each limb part (making a total of 160) along with 20 for the torso and 40 for the head. Note that training segmentations were deliberately not re-scaled to normalise for changes in the physical size of the subject, in order to account for this variability.

The rotation about each limb part’s major axis was not parameterised since these rotations changed the shape and appearance very little. Furthermore, the limb part templates were constrained to be symmetric about their major axis. This was achieved by flipping the training segmentations and using the flipped versions for learning as well.

4.4 Probabilistic Self-occlusion

Depth ordering is used to account for self-occlusion. Since the depth ordering is not known prior to estimation in most cases, the ordering must

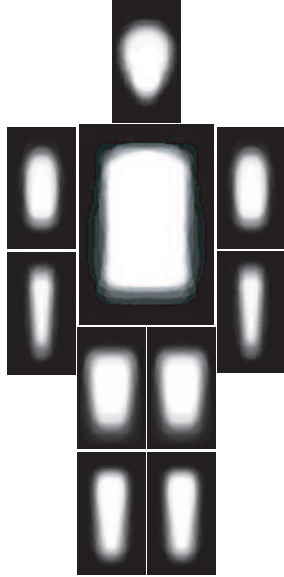


Figure 2: The probabilistic region templates, all at the same scale, that result from marginalising over the foreground segmentations and enforcing symmetry around the vertical axis.

be included as part of the pose hypothesis. Since shape is modelled using probabilistic region templates, a hypothesised configuration gives rise to a probabilistic assignment of part labels to points in image space. More specifically, let \mathcal{P}_i denote the set of image points at which the i^{th} part is visible. Furthermore, let $j \uparrow i$ denote that the j^{th} part is closer to the camera than the i^{th} part according to the hypothesised depth ordering. The probability that the i^{th} part will be visible at a point \mathbf{x} in the image plane (i.e. the part’s foreground probability at that point) is determined by the inverse part transform, as shown in Equation (3).

$$P(\mathbf{x} \in \mathcal{P}_i | \mathbf{C}) = \mathbf{M}_i(T_i^{-1}(\mathbf{x})) \times \prod_{j:j \uparrow i} (1 - \mathbf{M}_j(T_j^{-1}(\mathbf{x}))) \quad (3)$$

The probability that no part will be visible at a point \mathbf{x} given the hypothesised configuration can be computed using Equation (4) where i indexes the parts in the configuration.

$$P(\nexists i : \mathbf{x} \in \mathcal{P}_i | \mathbf{C}) = 1 - \sum_i P(\mathbf{x} \in \mathcal{P}_i | \mathbf{C}) \quad (4)$$

These equations will be used to form likelihood measurements. In particular, the probability of a part being visible at a point in the image can be used to determine a weight that is used to form a distribution that describes the appearance of the body part.

4.5 Pose Prior

A simple hard constraint prior can be based on upper and lower bounds on the relative pose of anchor points defined on pairs of body parts. Such a prior can be learned from data. This prior embodies scale and translational invariance. It does not directly incorporate constraints upon the absolute or relative orientation of body parts. A more specific model, defined in terms of global pose, would improve discrimination but such models require more data to estimate and are not the focus of this work.

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 coordinates) that connects the anchor points between parts i and j be $(m_{i,j}, n_{i,j})$. 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 prior over background poses is also considered to be uniform, but over the entire image. The prior probabilities of being a person or background, $P(v)$, 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 by Equation (5):

$$p(\mathbf{C}, v = 1) \propto \prod_{i,j:j > i} p(m_{i,j} | v = 1) p(n_{i,j} | v = 1) \quad (5)$$

where i and j index the parts in the configuration, \mathbf{C} . Body parts are also constrained to lie within the image.

The parameters of the prior, namely the minimum and maximum relative horizontal and vertical displacements, were determined from 150 images

of standing, walking, pointing, waving and sitting poses from various viewpoints (i.e. not always face on). People were always upright in these images. Part elongation was constrained such that $e_i \geq 0.7$ for any hypothesised part.

5 Likelihood

In this section the probabilistic part based shape model is used to inspect the image and determine support for the partial configuration hypothesis. Clearly, it is not feasible to learn the joint PDF of measurements conditioned upon model parameters, $p(\mathbf{I}|\mathbf{C}, v)$, due to the large number of parameters. In the interests of generalisation a highly parameterised model must be established that encodes conditional independencies, representing, for example, invariance to position and foreground appearance. This Section develops two such models. The boundary model discriminates people using short range appearance differences that occur along the model boundary. In contrast, the inter-part model discriminates people based upon long range appearance similarities between body parts. Both likelihood models are formulated in terms of the divergence of appearance distributions formed from regions in the image induced by the high-level shape model.

5.1 Foreground Appearance

Let $f(\mathbf{I}, \mathbf{x})$ denote a local image feature computed at location \mathbf{x} in image \mathbf{I} . A foreground appearance model for a part is based upon marginalising features over its foreground region, weighted by visibility:

$$F_i(\mathbf{z}) = \sum_{\mathbf{x}: f(\mathbf{I}, \mathbf{x})=\mathbf{z}} P(\mathbf{x} \in \mathcal{P}_i | \mathbf{C}) \quad (6)$$

The appearance distributions are represented using joint intensity-chromaticity histograms (3D distributions). A histogram representation is used for speed of computation and because the distributions are often multi-modal. 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. [55].

5.2 Part Boundary Model

As remarked in the review section, the majority of boundary based object localisation methods rely upon bottom up boundary detection (i.e. edge or localised filter). Furthermore, much consideration has been given to how to compare feature distributions and what features to use. In contrast, the focus here is on using the high-level model to capture large scale texture and predicting where strong contrasting boundaries are likely to occur. This can be seen as a natural progression of the feature divergence approaches to larger scales by conditioning upon the high level model. This difference in emphasis is particularly important for human appearance modelling where large scale textures often occur and where there is often weak boundaries between body parts (i.e. there is a discrepancy between the structural part model and the visual boundaries in the image). For example, whilst the kinematic structure is described by boundaries at the neck, elbow, shoulders, hips and knees the visual boundaries often occur on a part (e.g. a T-shirt ending midway along the upper arm) or not at all (e.g. the usual lack of a boundary between the upper arm and torso).

In order to account for the difference between a kinematic and visual boundary the notion of spatially dependent contrast is introduced. Many models of contrast are possible, here a straightforward model based upon the expected contrast between body parts is developed. This model assumes all points on two body parts are equally contrasting (ignoring visibility). The extent to which a pixel \mathbf{x} is expected to contrast in appearance with the i^{th} part is given by Equation (7):

$$\Gamma_i(\mathbf{x}) = P(\nexists j : \mathbf{x} \in \mathcal{P}_j | \mathbf{C}) + \sum_j \alpha_{ij} P(\mathbf{x} \in \mathcal{P}_j | \mathbf{C}) \quad (7)$$

where α_{ij} are weights that encode prior expectation of part contrast with other parts. Note that $\alpha_{ij} = \alpha_{ji}$ and $\alpha_{ii} = 0$. These weights could be estimated from representative data or, as in the case of experiments reported in Section 7, subjectively (specifically, $\alpha_{ij} = 0.1$ when i and j were adjoining limb segments, $\alpha_{ij} = 0.5$ between other combinations). It is important to understand that using more discriminatory features, such as orientation

dependent texture features, will result in higher expected contrast between points.

As one might expect, varying the size of the contrasting region gives a tradeoff between shape specificity and obtaining a good estimate of the contrasting appearance. For the results presented here the contrasting appearance distribution is extracted from a region of approximately equal area (in the probabilistic sense) to the foreground region. 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). In particular, this region is formed for each part by finding the Euclidean distance d such that all points with $P(\mathbf{x} \in \mathcal{P}_i|\mathbf{C}) = 0$ that are within a distance d of a point with $P(\mathbf{x} \in \mathcal{P}_i|\mathbf{C}) > 0$, in addition to those with $P(\mathbf{x} \in \mathcal{P}_i|\mathbf{C}) < 1$, 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.

$$B_i(\mathbf{z}) = \sum_{\mathbf{x}:f(\mathbf{I},\mathbf{x})=\mathbf{z}} \Gamma_i(\mathbf{x}) \quad (8)$$

Finally, to complete the model of a contrasting region, consideration must be given to the evaluation of partial configurations that do not describe all possible body parts. In such cases the *expected* pose of the missing body parts can be used to obtain contrasting regions. For example, when detecting single body parts, performance can be improved by distinguishing positions where the background appearance is most likely to differ from the foreground appearance. For example, a region at the top of the lower arm where it usually joins the upper arm can be identified as an adjoining region and would be expected to have similar appearance to the part. 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 above is employed.

Once the foreground and contrasting appearance distributions (histograms) are formed there are many measures that could be used to compare them including χ^2 , the Kullback Leibler divergence, the

Jeffrey distance, the Bhattacharyya measure and the Minkowski metric. Alternatives have been proposed specifically for comparing histograms including histogram intersection, the quadratic form and the Earth mover’s distance, the latter two being global measures of histogram similarity. (See Puzicha *et al.* [47] for a comparison of distribution similarity measures in the context of texture regions.) Based on its success in colour based tracking [8], the Bhattacharyya measure is adopted here. The divergence measure between the i^{th} part’s foreground and background appearance is given by Equation (9).

$$\mathcal{D}_i = \sum_{\mathbf{z}} \sqrt{F_i(\mathbf{z}) \times B_i(\mathbf{z})} \quad (9)$$

The Bhattacharyya measure is related to the likelihood non-linearly. Therefore, the distributions of divergence between foreground and background appearance are learned for correct ($v = 1$) and incorrect ($v = 0$) configurations in a supervised fashion. In particular, a $v = 1$ distribution is 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 $v = 0$ distribution, which encodes the likelihood of observing a part-shaped object in the class of scenes under consideration, is estimated by generating random alignments elsewhere. The ratio of these two distributions, Equation (10), defines a log-likelihood ratio, l_1 , for a partial configuration based on the region divergence measures for its parts.

$$l_1 = \sum_i (\ln p(\mathcal{D}_i|v = 1) - \ln p(\mathcal{D}_i|v = 0)) \quad (10)$$

Any single-part 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 be a body part.

In order to obtain a smooth log likelihood function, l_1 , and interpolate/extrapolate the learnt data, a parametric function was fitted to the data. In particular, it was expected, and confirmed empirically, that a Boltzmann sigmoid function, with a functional form given in Equation (11), would provide a good fit for the boundary ratio (correlation coefficient 0.96). This is the function used to ‘score’

a single body part configuration and is plotted in Fig. 3. This learnt sigmoid function acts as a soft classifier for body parts based upon the divergence measure.

$$S(x) = a + \frac{b - a}{1 + e^{\frac{c-x}{d}}} \quad (11)$$

5.3 Inter-Part Model

Since the part boundary likelihood ratio will usually result in many false positives, it is useful to encode relationships *between* the appearance of body parts to improve discrimination. For example, a person’s upper left arm will often 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.

The model of inter-part similarity encodes dissimilarity using the divergence between pairs of parts. 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, image texture 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 part boundary model, colour histograms are used to represent appearance and divergence between two parts’ foregrounds as given by Equation (12). Future work might investigate texture features in addition to colour to enhance discrimination of body parts, especially overlapping parts with oriented texture.

$$\mathcal{F}_{ij} = \sum_{\mathbf{z}} \sqrt{F_i(\mathbf{z}) \times F_j(\mathbf{z})} \quad (12)$$

PDFs of the divergence measure were learnt similarly to the part boundary case. Equation (13) gives the inter-part log likelihood ratio, l_2 , that results from these two distributions.

$$l_2 = \sum_{(i,j)} (\ln p(\mathcal{F}_{ij}|v = 1) - \ln p(\mathcal{F}_{ij}|v = 0)) \quad (13)$$

Specifically, inter-part divergence PDFs for pairs of opposing limb parts expected to have similar appearance were learned from examples with correct and incorrect pose. In particular, 20 pairs of upper and lower arms and legs were used. Their correct pose was manually specified and incorrect poses were generated by drawing from the pose prior of Equation (5). Fig. 4 shows plots of these two PDFs modelled as Gaussians. The resulting likelihood ratio function is also shown. It can be seen that this model strongly penalises opposing part pairs that are not similar in appearance.

5.4 Combining the Models

The log-likelihood ratios, l_1 and l_2 , are combined assuming conditional independence. Notice that the relative importance of the models is implicit in the likelihood ratio which allows principled fusion of the different models. The overall log-likelihood ratio is the sum of the boundary and inter-part components.

6 Empirical Investigation of Likelihood

Before considering full pose estimation it is useful to investigate the likelihood models by computing projections resulting from the variation of parts from interesting configurations. This allows an understanding of the sensitivity to different types of clutter. Furthermore, an intensity edge model to allow a more quantitative evaluation of the new boundary likelihood model.

For these tests and the pose estimation results that follow the images were obtained using a range of still and video cameras. It was assumed that the intrinsic camera parameters were unknown and constant (between scenes and cameras). The typical resolution of the input images was 640×480 pixels. The colour appearance histograms had 8^3 bins.

6.1 Comparison to Intensity Edge Model

The proposed part boundary likelihood ratio was compared to an intensity edge-based model for

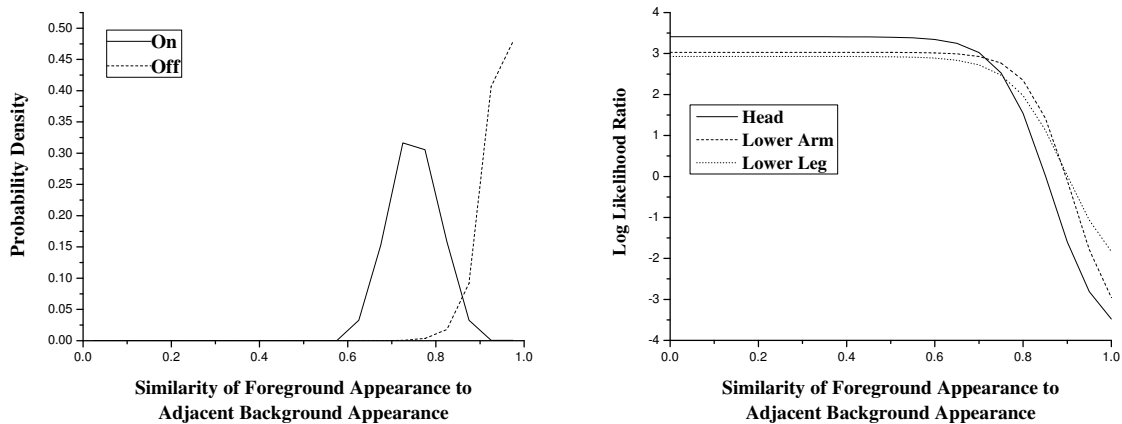


Figure 3: Left: A plot of the learnt PDFs of foreground to background appearance similarity for the $v = 1$ and $v = 0$ part configurations of a head template. Right: A plot of a Boltzmann sigmoid function fit to the log of the likelihood ratio data for head, lower arm and lower leg parts. It can be seen that the distributions are well separated.

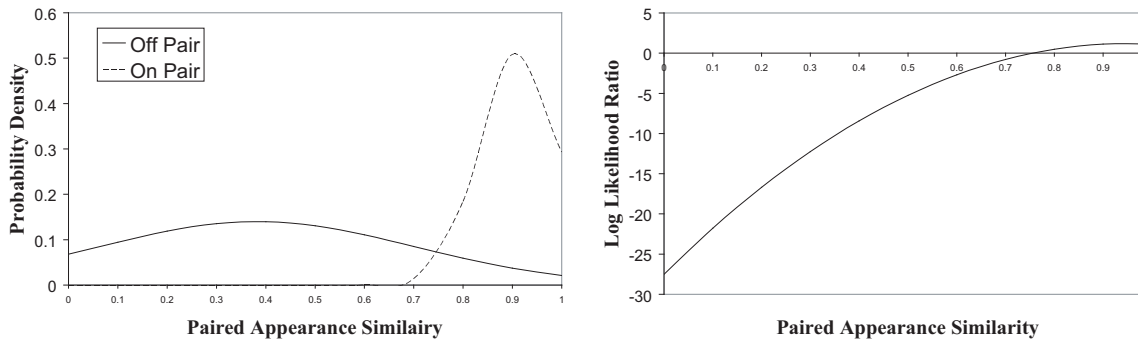


Figure 4: Left: A plot of the learnt PDFs of foreground appearance similarity for paired and non-paired configurations. Right: The log of the resulting likelihood ratio. It can be seen, as would be expected, that more similar regions are more likely to be a pair.

single body parts. Rather than make assumptions about the form of the model boundary to be matched to image edges (for example, a fixed variance Gaussian [63]), the learned probabilistic region templates were used. The spatial gradient of a probabilistic region template provides an estimate of the mean spatial gradient for a part (since the derivative is a linear operator). A response is formed by convolving the derivative of the probabilistic mask with the image. Image edge magnitude and orientation were computed using 3×3 Sobel filters.

The magnitudes of edge responses corresponding to a body part boundary will vary in a structured

and partly predictable way. For example, boundary segments around the joints neighbouring similarly clothed parts will have low magnitudes. A model of this structure can be used to improve discrimination. The single part response was investigated here using manually selected segments of expected high magnitude, similarly to some other systems (e.g. [57]). Part-specific response distributions were learned in a supervised fashion from body part training data for correct ($v = 1$) and incorrect ($v = 0$) part pose. Only the magnitude of the component of the filter response orthogonal to the model was used for discrimination. The likelihood ratio for the part as a whole was computed by

assuming independence of the individual measurements.

6.2 Single Part Likelihood Ratios

Fig. 5 shows the projections of log-likelihood ratios for single part configurations onto typical images containing significant clutter. Results are shown for both the part boundary model (computed using Equation (10)) and the comparative edge-based model. 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 proposed part boundary method is highly discriminatory and produces relatively few false maxima. Fig. 6 illustrates the typical spatial variations of the part boundary and the edge-based likelihood ratios. The edge-based response, whilst indicative of the correct position, has significant false, positive log-likelihood ratios.

Although the part boundary ratio is more expensive to compute than the edge-based ratio (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. The edge-based method did not use contrast normalisation or a multi-scale approach as expounded in Ref. [57] and this may partly explain its relatively poor performance.

6.3 Inter-part Likelihood Ratios

Fig. 7 shows the projection of the inter-part likelihood ratio for a typical image and shows it to be highly discriminatory. It limits the possible pose configurations if one limb part can be found reliably and helps reduce the likelihood of incorrect large assemblies. This enables larger incorrect configurations to be pruned, making deterministic, combinatorial search more feasible.

7 Pose Estimation

A central thesis of this work is that by improving the formulation and likelihood models, the estimation problem can be eased. In particular, by better discriminating individual parts and using rela-

tions between parts to prune larger incorrect configurations, relatively simple estimation schemes can yield useful pose estimation results. The partial configuration formulation allows bottom-up sampling to focus pose search, making global estimation feasible whilst still allowing for self-occlusion (and inter-part appearance relations). 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 [51] and mixtures of trees [22], is more flexible and it is feasible because of the strong likelihood model developed. The search scheme presented here is relatively straightforward. Nevertheless it succeeds in obtaining interesting pose estimation results thus demonstrating the strength of the formulation and likelihood model.

Recall that it is assumed that exactly one person is present in the image. The pose estimation problem can thus be treated as one of global maximisation. It is also assumed, as is common with other pose estimation systems, that the scale parameter is known to a good approximation. The model based approach makes incorporate this information straightforward.

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 search scheme used here aims to identify the head and outer limbs (i.e. the lower arms and lower legs). This also makes labelling of parts simpler. Future work could improve the search scheme to find the remaining parts.

7.1 Search Algorithm

7.1.1 Coarse sampling

First, the parameter spaces of single part pose configurations was uniformly sampled. Specifically, translation parameters were sampled every 3 pixels and part orientations every $\pi/4$ radians (which was sufficient because limb part templates were sym-

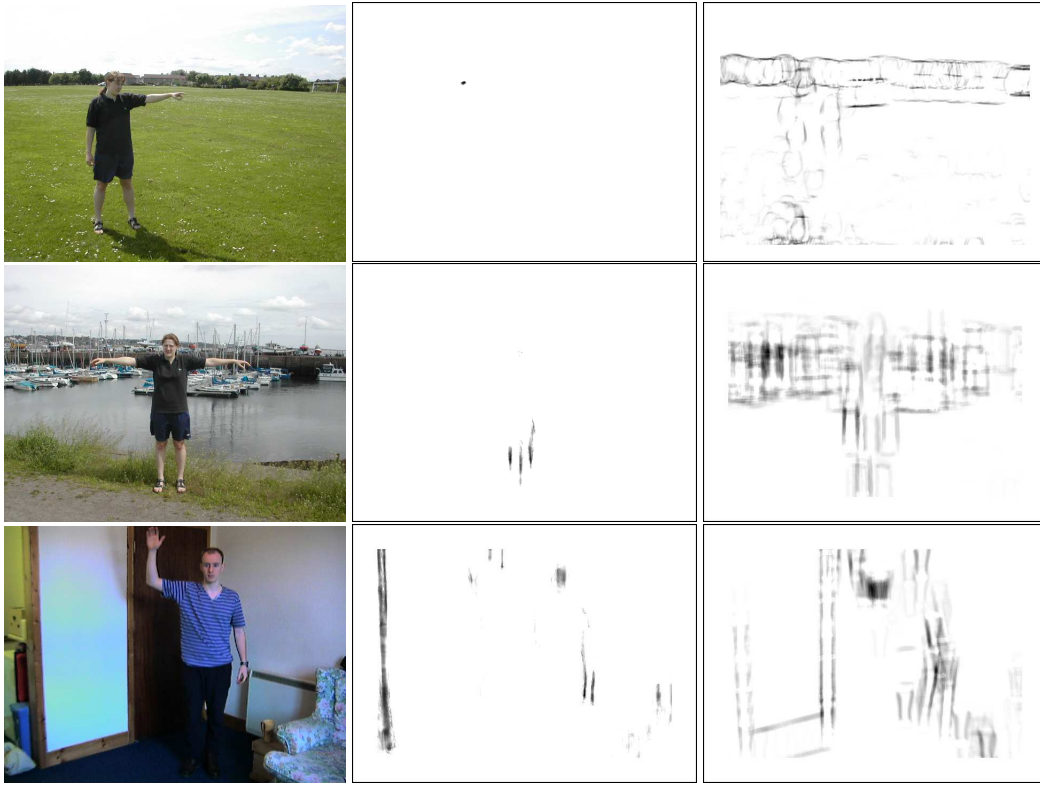


Figure 5: Various input images (left), the projection of boundary divergence likelihood (middle) and intensity edge likelihoods (right) for correctly sized and oriented part configurations. The top row shows the projection for the head model. Here the divergence method greatly reduces the number of false positives. The middle row shows the projection for the lower leg model. Here the clutter from the sail masts distracts the localised intensity edge model. The bottom row shows the projection of an arm model to an indoor scene. Here the door frame provides strong clutter for both models.

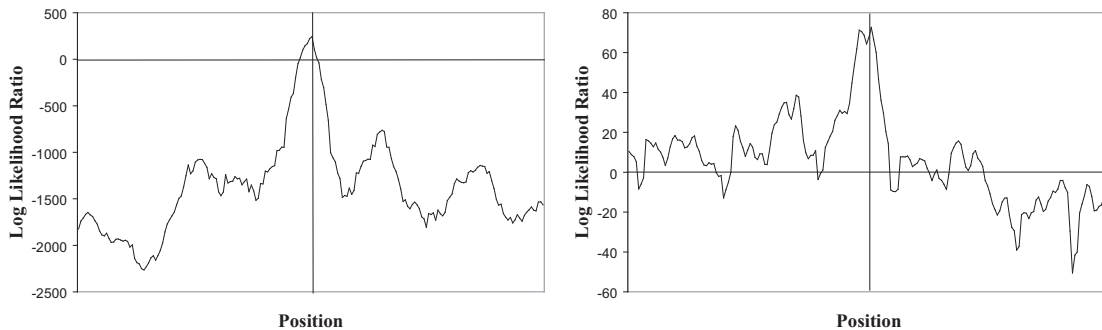


Figure 6: Comparison of the spatial variation (plotted for a horizontal change of 200 pixels) of the learnt log likelihood ratios for the model proposed here (left) and the edge-based model (right) of the head in Fig. 5. The correct position is centered and indicated by the vertical bar. Anything above the horizontal bar, corresponding to a likelihood ratio of 1, is more likely to be a head than not.

metric about their major axis). Part configurations with log-likelihood ratios greater than, T , were retained. The main purpose of the threshold T is to reduce the number of hypotheses in highly cluttered environments and achieve reasonable search times (possibly at the expense of incorrect an-



Figure 7: Investigation of a paired part response. Left: an image for which significant limb candidates are found in the background. Right: the projection of the log likelihood ratio for the paired response to the person's lower right leg in the image.

swers). For most experiments it was kept constant at a conservative value. Parts were allocated initial labels based upon the division of the image into quadrants. However, this initial labelling is unimportant because as subsequent search identifies larger configurations labels are re-hypothesised and better constrained. At this stage the head and some outer limbs were often found (if unoccluded and not camouflaged) along with false positives due to background clutter.

7.1.2 Local Optimisation

Each part configuration was locally optimised by iteratively proposing random perturbations and accepting them when the posterior ratio increased (this was found to perform better than a numerical gradient based local search). Translation was perturbed by up to 2 pixels in each direction, orientation by up to 12° and elongation by up to 10%. Depth order was also searched in order to account for self-occlusion by proposing part movements up and down a layer. Part configurations with similar pose were merged.

7.1.3 Combinatorial Search

Part candidates are then iteratively grouped to build larger pose configurations. This began with evaluation of all possible pairs of (locally optimised) parts. Those pairs with likelihood ratio lower than T were discarded. Triples were then formed from all the parts in the retained pairs. This continued

to some maximum configuration size, 5 parts in the results reported here. Evaluation of a configuration at any stage in this combinatorial search involved part labelling, evaluation of the prior and, for those configurations with non-zero prior, evaluation of the likelihood ratio, l . Part labelling was performed by first selecting an anchor part, the head if present and otherwise the upper left part. Other parts were labelled based on their position relative to this anchor part. Parts are labelled as left or right according to their relative positions in the image frame. The search was elitist in that the best configuration was kept irrespective of whether it passed later grouping stages. At each stage of grouping, the inter-part pose and appearance relations reduced the number of possible parts under consideration. After each grouping stage local optimisation was performed. After the final grouping stage the best configuration was locally optimised for 200 iterations prior to output, including the global scale factor which was perturbed by up to 5%. Every stage in the search involves a re-hypothesising of part labels. For very small configurations (e.g. an upper arm) the labelling is highly unconstrained and prone to error. However, for larger configurations the prior constraints make the labelling more likely to be correct. The labels at any stage (e.g. single part stage) in the search are only important for that stage to be able to return a MAP estimate of pose. Representing distributions over part labels (and pose) is deferred for future work.

A limitation of this 'feed-forward' inference scheme is that parts that are significantly occluded

by other part(s), and therefore have weak likelihood responses, are never hypothesised later in the search. Future work could develop sampling schemes that, given the pose of a set of body parts, form larger configurations that account for the lack of contrast and self-occlusion.

7.2 Empirical Evaluation

Pose estimation was evaluated on a set of images with characteristics described in Section 3 (i.e. cluttered indoor and outdoor scenes, various unknown subjects in various poses without constraining either clothing or lighting). Figures 8 and 9 show the pose configurations with highest posterior ratio found within a fixed maximum run-time. Although inter-part links are not visualised here, we emphasise that these results represent estimates of pose configurations with inter-part relationships in terms of appearance and pose. Closer inspection of the input images reveals the presence of JPEG encoding artifacts.

Histograms were built efficiently by projecting scan-line segments and iteratively computing the mask co-ordinates inside these segments. The colours in the image were preprocessed into histogram bins. The implementation sampled single part configurations with scale $s = 1$ at approximately $3KHz$ from an typical image with resolution 640×480 on a 2GHz PC. Run-time for a complete pose estimation ranged from 2 minutes, when limited part candidates were identified, to the maximum 2 hours, when many part candidates were identified in heavily cluttered scenes.

7.3 Discussion of Pose Estimation Results

The results support the 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 system was able to recover pose in the presence of other-object occlusion. The largest configurations presented in the results happen to have four parts. Although five part configurations were hypothesised they gave lower responses. For all the test images the head, arms and legs were correctly labelled. It was difficult to identify a single ‘correct’ scale for some images however, the sys-

tem was not overly sensitive to the correct choice. In particular, baggy clothing and perspective effects cause changes in relative scale between body parts (breaking the assumption of a common single scale). The model showed good generalisation over changes in scale (600% variation). It was observed from these results that pose estimation in indoor scenes was more problematic due to clutter from man-made objects such as door frames. Scenes with large amounts of clutter caused long run-times due to combinatorial explosion in the number of larger configurations.

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 initially assume. Moreover, these results compare favorably with other state-of-the-art pose estimation systems that require more restricted scenes and assume that more is known about the appearance (e.g. [22, 23]). In particular, these other systems also often only find three to five body parts.

8 Conclusions

Two fundamental problems of visual human pose estimation were focussed upon: (i) discriminating a subject with complex, unknown appearance from a cluttered, unknown scene that possibly occludes parts of the subject using a single image, and (ii) formulating the pose estimation problem such that efficient, accurate global estimation is possible in such conditions. This is in contrast to the majority of published research on human tracking and human pose estimation which has focused upon the estimation problem. The strategy adopted here was to ease the estimation problem by improving the formulation and likelihood model.

The first main contribution 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 of people but has been largely ignored in previous work. Sec-



Figure 8: Pose estimation results. Notice the large variation in clothing appearance, the loose fit of some clothing and degree of background clutter. Also notice the presence of other object occlusion. The only complete failure was the final image where a reflection in the background provided a strong part response.

only, 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 without making restrictive assumptions about self-occlusion, other-object occlusion or inter-part relations. This was combined with a novel shape model that, when transformed into the image using the pose parameters, encodes the uncertainty in visibility of parts at points in the image and 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 was further developed to model self-occlusion and expected contrast.

The second main contribution was an efficient, highly discriminatory, spatial likelihood model composed of two complementary components. The boundary component allowed good discrimination of body parts based upon divergence between feature distributions in regions induced by the high-level shape model encoded using probabilistic region templates and taking into account inter-part contrast. The use of high-level shape allowed better discrimination in the presence of complex, textured appearance than models based upon bottom-up boundary measurements. The inter-part component enabled efficient discrimination of larger con-

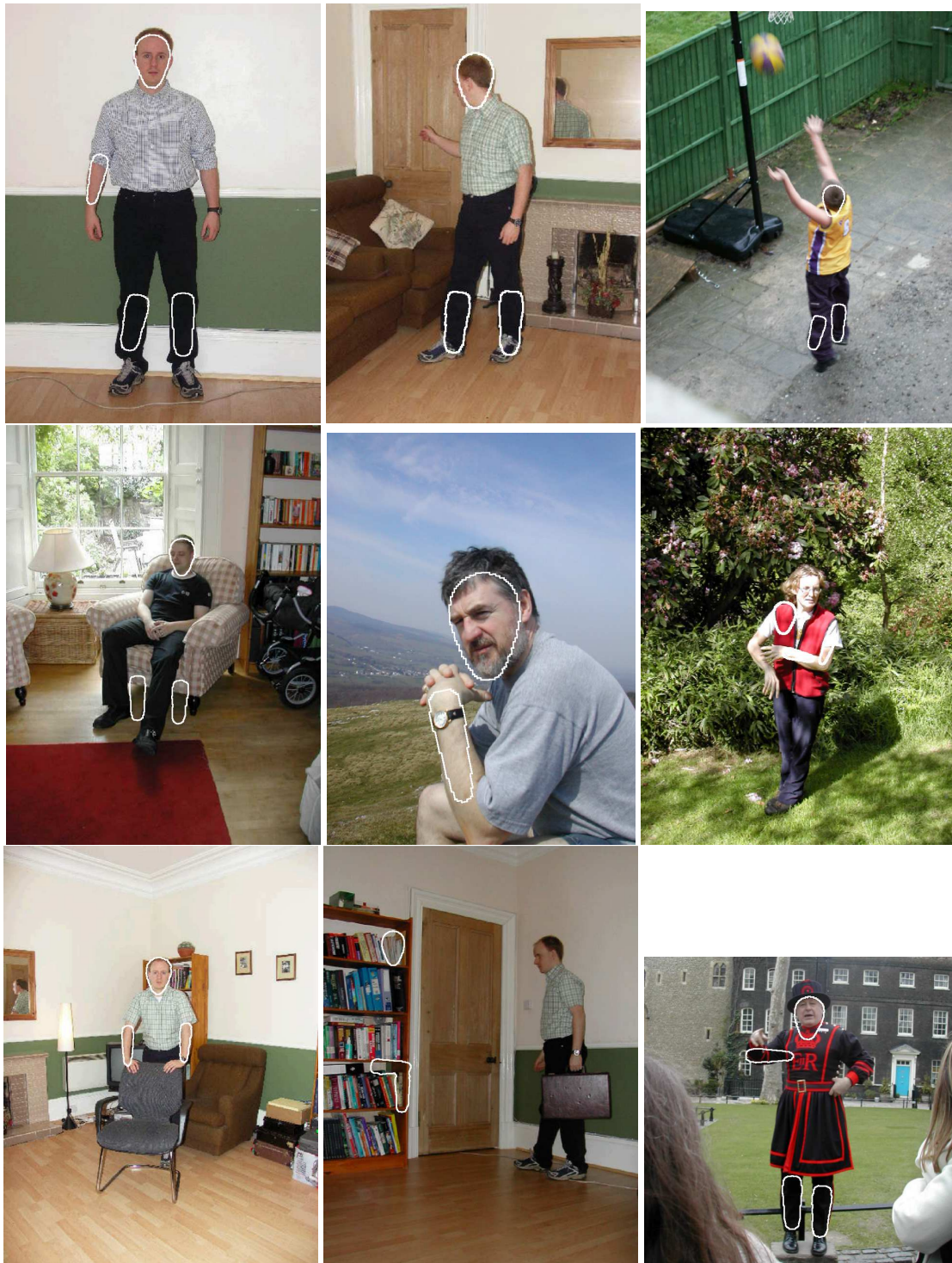


Figure 9: Pose estimation results. Notice the large scale changes that are present as well as the variation in pose (sitting, standing, jumping). Also notice the presence of other object occlusion. Two of the images show partially incorrect results and one is completely incorrect (due to strong clutter from the bookcase).

figurations by exploiting the long range appearance similarity between parts.

It was demonstrated that by building upon this foundation a straightforward search scheme is able

to recover a large amount of information about the pose efficiently and globally. Such an approach could be used to automatically initialise and recover human trackers without relying upon ad-hoc models or assumptions of visibility.

References

- [1] 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.
- [2] 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] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2nd edition, 2001.
- [13] 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.
- [14] D. A. Forsyth and M. M. Fleck. Body plans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–683, Puerto Rico, 1997.
- [15] D. A. Forsyth and M. M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77, August 1999.
- [16] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999.
- [17] 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.
- [18] M. Grosso, R. Quach, and N. Badler. Anthropometry for computer animated human figures. In *Proceedings of Computer Animation*

- Human Figures*, pages 83–96. Springer Verlag, 1989.
- [19] 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 1999.
- [20] 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.
- [21] D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.
- [22] S. Ioffe and D. A. Forsyth. Human tracking with mixtures of trees. In *IEEE International Conference on Computer Vision*, volume 1, pages 690–695, 2001.
- [23] S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43:45–68, June 2001.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] M. W. Lee and Cohen I. Human upper body pose estimation from static images. In *European Conference on Computer Vision*, pages 126–138, 2004.
- [30] 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.
- [31] J. MacCormick and A. Blake. A probabilistic contour discriminant for object localisation. In *IEEE International Conference on Computer Vision*, pages 390–395, 1998.
- [32] J. MacCormick and A. Blake. Spatial dependence in the observation of visual contours. In *European Conference on Computer Vision*, 1998.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilis-

- tic assembly of robust part detectors. In *European Conference on Computer Vision*, volume I, pages 69–81, 2004.
- [38] T. B. Moeslund and F. Bajers. Summaries of 107 computer vision-based human motion capture papers. Technical Report LIA99-01, University of Aalborg, 1999.
- [39] 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.
- [40] 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.
- [41] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, pages 666–680, 2002.
- [42] G. Mori, X. Ren, A.A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. pages II: 326–333, 2004.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] R. Ronfard, C. Schud, and B. Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714, Copenhagen, 2002.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.
- [56] A. Shahrokni, T. Drummond, and P. Fua. Texture boundary detection for real-time track-

- ing. In *European Conference on Computer Vision*, volume II, pages 566–577, 2004.
- [57] 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.
- [58] 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, 2000.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80:349–363, September 2000.
- [63] S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174–192, June 1999.
- [64] 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.
- [65] 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.
- [66] J. Zhao and N.I. Badler. Inverse kinematics positioning using nonlinear programming for highly articulated figures. *ACM Trans. Graph.*, 13(4):313–336, 1994.