Human Pose Estimation using Learnt Probabilistic Region Similarities and Partial Configurations ECCV 2004 PREPRINT

Timothy J. Roberts, Stephen J. McKenna and Ian W. Ricketts

Division of Applied Computing
University of Dundee
Dundee DD1 4HN, Scotland
{troberts,stephen,ricketts}@computing.dundee.ac.uk
http://www.computing.dundee.ac.uk

Abstract. A model of human appearance is presented for efficient pose estimation from real-world images. In common with related approaches, a high-level model defines a space of configurations which can be associated with image measurements and thus scored. A search is performed to identify good configuration(s). Such an approach is challenging because the configuration space is high dimensional, the search is global, and the appearance of humans in images is complex due to background clutter, shape uncertainty and texture.

The system presented here is novel in several respects. The formulation allows differing numbers of parts to be parameterised and allows poses of differing dimensionality to be compared in a principled manner based upon learnt likelihood ratios. In contrast with current approaches, this allows a part based search in the presence of self occlusion. Furthermore, it provides a principled automatic approach to other object occlusion. View based probabilistic models of body part shapes are learnt that represent intra and inter person variability (in contrast to rigid geometric primitives). The probabilistic region for each part is transformed into the image using the configuration hypothesis and used to collect two appearance distributions for the part's foreground and adjacent background. Likelihood ratios for single parts are learnt from the dissimilarity of the foreground and adjacent background appearance distributions. It is important to note the distinction between this technique and restrictive foreground/background specific modelling. It is demonstrated that this likelihood allows better discrimination of body parts in real world images than contour to edge matching techniques. Furthermore, the likelihood is less sparse and noisy, making coarse sampling and local search more effective. A likelihood ratio for body part pairs with similar appearances is also learnt. Together with a model of inter-part distances this better describes correct higher dimensional configurations. Results from applying an optimization scheme to the likelihood model for challenging real world images are presented.

1 Introduction

It is popular in the literature to match a high-level shape model to an image in order to recover human pose (see the review papers [1, 2]). Samples are drawn from the shape configuration space to search for a good match. The success of this approach, in terms of its accuracy and efficiency, depends critically on the choice of likelihood formulation and its implicit assumptions. This paper presents a strong likelihood model and a flexible, effectively low dimensional formulation that allows efficient inference of detailed pose from real-world images. Pose estimation is performed here from single colour images so no motion information is available. This method could however form an important component in an automatically (re)initialising human tracker.

1.1 Assumptions

Estimation of human body pose from poorly constrained scenes is made difficult by the large variation in human appearance. The system presented here aims to recover the variation due to body pose automatically and efficiently in the presence of other variations due to:

- *unknown* subject identity, clothing colour and texture
- *unknown*, significantly cluttered, indoor or outdoor scenes
- uncontrolled illumination
- general, other object occlusion

It is assumed that perspective effects are weak and that the scale is such that distributions of pixel values or local features can be estimated and used to characterise body parts. It is further assumed that the class of view point is known, in this case a side on view. These assumptions apply to a large proportion of real world photographs of people.

1.2 Formulation

There are two main approaches to human pose estimation. The ‘top-down’ approach makes samples in a high dimensional space and fully models self-occlusion (e.g. [3–6]). It does not incorporate bottom-up part identification and is inappropriate without a strong pose prior (and is therefore mostly used in trackers). The ‘bottom-up’ approach identifies the body parts and then assembles them into the best configuration. Whilst it does sample globally it does not model self-occlusion. Both approaches tend to rely on a fixed number of parts being parameterised (a notable exception being the recent work of Ramanan and Forsyth [7]). However, occlusion by other objects or weak evidence may make some parts unidentifiable. The approach of *partial configurations* presented here bridges these two approaches by allowing configurations of different dimensionalities to be compared. This is done by combining learnt likelihood ratios computed only from the parameterised, visible parts. The method has several advantages. Firstly, it allows general occlusion conditions to be handled. Secondly, it makes use of the fact that some parts might be found more easily than others. For example, it is often easier to locate parts that do not overlap. Thirdly, it makes use of the fact that configurations with small numbers of parts contain much of the overall pose information because of inter-part linking. For example, knowing the position of just the head and outer limbs greatly constrains the overall pose. The approach of partial configurations, along with a global stochastic optimization scheme, is more flexible than pictorial structures [8] since it allows a large range of occlusion conditions. When employed in a time-constrained optimization scheme, it allows the system to report lower dimensional solution(s) should a higher dimensional one not be found in time. A consequence of the formulation is that parts must be parameterised in their own co-ordinate system rather than hierarchically as is often the case in tracking systems, e.g. [3]. Whilst this might appear to increase the dimensionality of the pose parameter space, in practice an offset term is often required to model complex joints like the shoulder [6] making the difference one of mathematical convenience.

1.3 Outline

The remainder of the paper details the three components that make up the likelihood ratio used to find humans in real images. For ease of exposition, Section 2 begins by describing the likelihood ratio used to find single body parts. A probabilistic region template is transformed into image space and used to estimate foreground and adjacent background appearances. The hypothesised foreground and background appearances are compared and a likelihood ratio is computed, based upon learnt PDFs of the similarity for on-part responses and off-part responses. The performance of this technique is then demonstrated and compared to a competing method. Section 3.2 presents a method for comparing hypothesised pose configurations incorporating inter-part joint constraints in which subsets of the body parts are instantiated. Section 3.3 then introduces a constraint based on the *a priori* expectation that pairs of parts will have similar appearance. Finally, pose estimation results are presented and conclusions drawn.

2 Finding Single Parts using Probabilistic Regions

The model of body parts proposed here provides an efficient mechanism for the evaluation of hypothesised body parts in everyday scenes due to a highly discriminatory response and characteristics that support efficient sampling and search. This Section describes the method used for modelling body part shape and the use of image measurements to score part hypotheses. It concludes with an investigation of the resulting response.

2.1 Modelling Shape

Current systems often use 2D or 3D geometric primitives such as ellipses, rectangles, cylinders and tapered superquadrics to represent body parts (e.g. [3–5]). These are convenient but rather *ad hoc* approximations. Instead, *probabilistic region templates* are used here as body part primitives. Due to the limited presence of perspective effects and 3D shape variation, a 2D model with depth ordering is used to represent the body.

A variation of the scaled prismatic model [9] is used to parameterise the transformed appearance. This reduces the dimensionality compared to a 3D model and removes kinematic singularities [10].

A body part, labelled here by $i(i \in 1...N)$, is represented using a single probabilistic region template, M_i, which represents the uncertainty in the part's shape without attempting to enable shape instances to be accurately reconstructed [1]. This is particulary important for efficient sampling when the subject wears lose fitting clothing. The probability that an image pixel at position (x, y) belongs to a hypothesised part i is then given by $M_i(T_i(x, y))$ where T_i is a linear transformation from image coordinates to template coordinates determined by the part's centre, (x_c, y_c), image plane rotation, θ, elongation, e, and scale, s. The elongation parameter alters the aspect ratio of the template and is used to approximate rotation in depth about one of the part's axes. The probabilities in the template are estimated from example shapes in the form of binary masks obtained by manual segmentation of training images in which the elongation is maximal (i.e. in which the major axis of the part is parallel to the image plane). These training examples are aligned by specifying their centres, orientations and scales. Unparameterised pose variations are marginalised over, allowing a reduction in the size of the state space. Specifically, rotation about each limb's major axis is marginalised since these rotations are difficult to observe. The templates are also constrained to be symmetric about this axis. It has been found, due to the insensitivity of the likelihood model described below to precise contour location, that upper and lower arm and leg parts can reasonably be represented using a single template. This greatly improves the sampling efficiency. Some learnt probabilistic region templates are shown in Fig. 1. The uncertain regions in these templates arise because of (i) 3D shape variation due to change of clothing and identity, (ii) rotation in depth about the major axis, and (iii) inaccuracies in the alignment and manual segmentation of the training images.

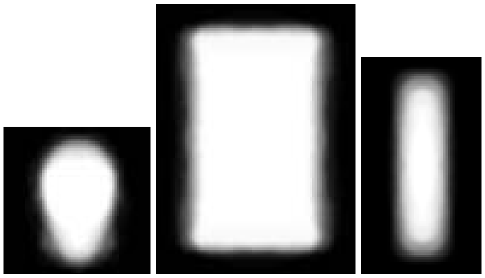

Fig. 1. Head, torso and limb probabilistic region templates. The upper and lower arm and legs are represented using a single mask (increasing sampling efficiency). Notice the masks' symmetries.

[1] Note that while it would be possible to represent the body parts using a set of basis regions, the mean was found to be sufficient here.

2.2 Single Part Likelihood

Several methods for body part detection have been proposed although in the opinion of the authors much work remains to be done. Matching geometric primitives to an edge field is popular, e.g. [11]. Wachter and Nagel [3] used only the edges that did not overlap with other parts. Sidenbladh *et al*. [12] emphasised learning the distribution of foreground and background filter responses (edge, ridge and motion) rather than forming *ad hoc* models. Ronfard *et al*. [8] learned part detectors from Gaussian derivative filters. Another popular method is modelling the background, but this has the obvious limitation of requiring knowledge of the empty scene. Matching model boundaries to local image gradients often results in poor discrimination. Furthermore, edge responses provide a relatively sparse cue which necessitates dense sampling. In order to achieve accurate results in real world scenes the authors believe that a description that takes account of colour or texture is necessary. To accomplish this the high-level shape model can be used earlier in the inference process. One might envisage learning a model that described the wide variation in the foreground appearance of body parts present in a population of differently clothed people. Such a model would seek to capture regularities due to the patterns typically used in clothing. However, such an approach would require a high dimensional model and prohibitively large amounts of training data. Furthermore, it would not be strongly discriminatory because most clothing and image regions are uniformly textured.

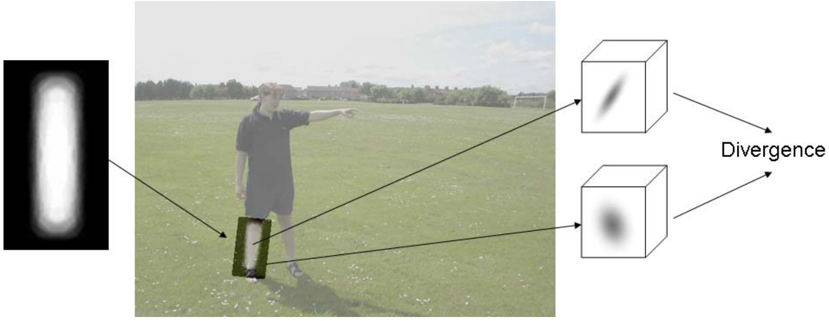

Fig. 2. The flow of data: A lower leg body part probabilistic region template is transformed into the image. The spatial extent of the template is such that the areas (in the probabilistic sense) of the foreground and background regions are approximately equal. The probabilistic region is used to estimate the foreground appearance and adjacent background appearance histograms. A likelihood is learnt based upon the divergence of the two histograms.

The approach taken here is to use the dissimilarity between the appearance of the foreground and background of a transformed probabilistic region as illustrated in Fig. 2. These appearances will be dissimilar as long as a part is not completely camouflaged. The appearances are represented here as PDFs of intensity and chromaticity image features, resulting in 3D distributions. In general, local filter responses could also be used to represent the appearance (c.f. [13]). Since texture can often result in multi-modal distributions, each PDF is encoded as a histogram (marginalised over position). For scenes in which the body parts appear small, semi-parametric density estimation methods such as Gaussian mixture models would be more appropriate. The foreground appearance histogram for part i, denoted here by F_i, is formed by adding image features from the part's supporting region proportional to $M_i(T_i(x, y))$. Similarly, the adjacent background appearance distribution, B_i, is estimated by adding features proportional to $1 - M_i(T_i(x, y))$.

It is expected that the foreground appearance will be less similar to the background appearance for configurations that are correct (denoted by on) than incorrect (denoted by $\overline{on}$). Therefore, a PDF of the Bhattacharya measure given by Equation (1) is learnt for on and $\overline{on}$ configurations [14]. The on 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 $\overline{on}$ distribution was estimated by generating random alignments elsewhere in 100 images of outdoor and indoor scenes. The on PDF can be adequately represented by a Gaussian (although in fact the distribution is skewed). Equation (2) defines $SINGLE_i$ as the ratio of these two distributions. This is the response used to score a single body part configuration and is plotted in Fig. 3.

$$I(F_i, B_i) = \sum_{\mathbf{f}} \sqrt{F_i(\mathbf{f}) \times B_i(\mathbf{f})} \tag{1}$$

$$SINGLE_i = \frac{p(I(F_i, B_i)|on)}{p(I(F_i, B_i)|\overline{on})} \tag{2}$$

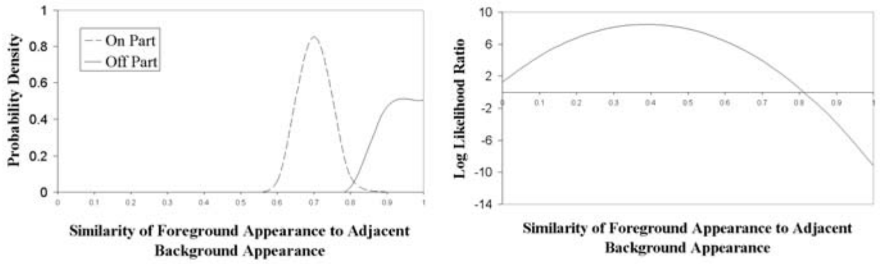

Fig. 3. Left: A plot of the learnt PDFs of foreground to background appearance similarity for the on and $\overline{on}$ part configurations of a head template. Right: The log of the resulting likelihood ratio. It can be seen that the distributions are well separated.

2.3 Enhancing Discrimination using Adjoining Regions

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. For example, due to the structure of clothing, when detecting an upper arm, *adjoining* background areas around the shoulder joint are often similar to the foreground appearance (as determined by the *structural* model used here to gather appearance data). The histogram model proposed thus far, which marginalises appearance over position, does not use this information optimally. To enhance discrimination, two separate adjacent background histograms are constructed, one for adjoining regions and another for non-adjoining regions. It is expected that the non-adjoining region appearance will be less similar to the foreground appearance than the adjoining region appearance. Currently, the adjoining and non-adjoining regions are specified manually during training by a hard threshold. A probabilistic approach, where the regions are estimated by marginalising over the relative pose between adjoining parts (to get a low dimensional model), would be better, but requires large amounts of training data. 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 formulation presented later in Section 3.1 is used.

2.4 Single Part Response Investigation

The middle column of Fig. 4 shows the projection of the likelihood ratio computed using Equation (2) onto typical images containing significant clutter. The top image shows the response for a head while the other two images show the response of a vertically-oriented limb filter. It can be seen that the technique is highly discriminatory, producing relatively few false maxima. Note the false response in between the legs in the second image: the space between the legs is itself shaped like a leg. Although images were acquired using various cameras, some with noisy colour signals, system parameters were fixed for all test images.

In order to provide a comparison with an alternative method, the responses obtained by comparing the hypothesised part boundaries with edge responses were computed in a similar manner to Sidenbladh and Black [12]. These are shown in the rightmost column of Fig. 4. Orientations of significant edge responses for foreground and background configurations were learned (using derivatives of the probabilistic region template), treated as independent and normalised for scale. Contrast normalisation was not used. Other formulations (e.g. averaging) proved to be weaker on the scenes under consideration. The responses using this method are clearly less discriminatory.

Fig. 5 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. The proposed part likelihood is more expensive to compute than the edge-based filter (approximately an order of magnitude slower in our implementation). However, it is far more discriminatory and as a result, fewer samples are needed when performing pose search, leading to an overall performance benefit. Furthermore, the collected foreground histograms are useful for other likelihood measurements as discussed below.

3 Body Pose Estimation with Partial Configurations

Since any single body part likelihood will result in false positives it is important to encode higher order relationships between body parts to improve discrimination. In this

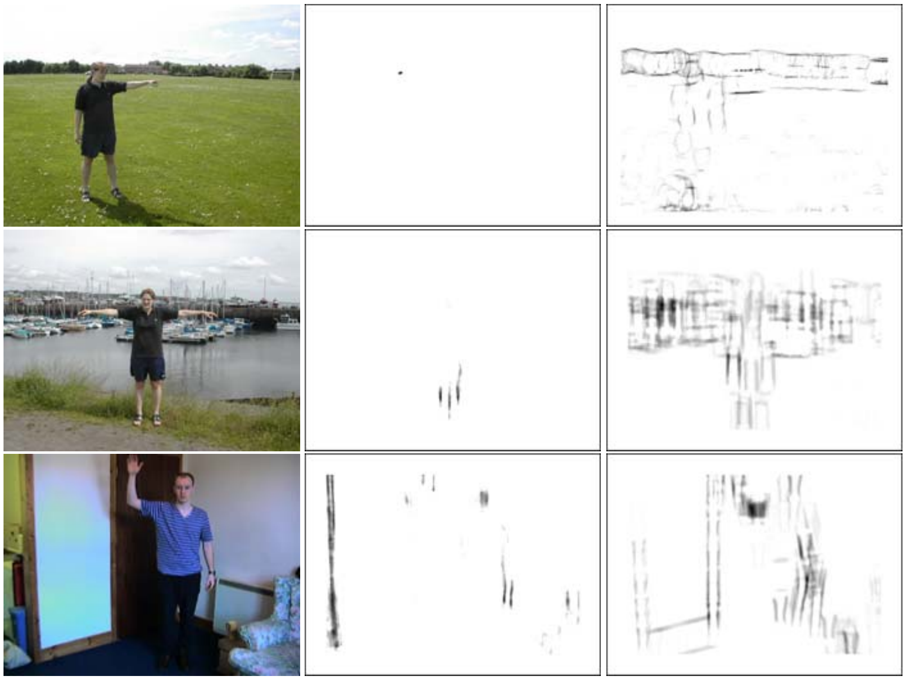

Fig. 4. First column: Typical input images from both outdoor and indoor environments. Second column: projection of the log likelihood (positive only, re-scaled) from the part filters. Third column: projection of the log likelihood ratio (positive only, re-scaled) for an edge-based model. First row: head model. Second and third rows: limb model (vertical orientation).

system this is accomplished by encoding an expectation of structure in the foreground appearance and the spatial relationship of body parts.

3.1 Extending Probabilistic Regions to Multi-Part Configurations

Configurations containing more than one body part can be represented using a straightforward extension of the probabilistic region approach described above. In order to account for self-occlusion, the pose space is represented by a depth ordered set, V, of probabilistic regions with parts sharing a common scale parameter, s. When taken together, the templates determine the probability that a particular image feature belongs to a particular parts foreground or background. More specifically, the probability that an image feature at position (x, y) belongs to the foreground appearance of part i is given by $M_i(T_i(x, y)) \times \prod_j (1 - M_j(T_j(x, y))$ where j labels closer, instantiated parts. Forming the background appearance is more subtle since some parts often have a similar appearance. Therefore, a list of paired body parts is specified manually and the background appearance histogram is constructed from features weighted by $\prod_k (1 - M_k(T_k(x, y))$ where k labels all instantiated parts other than i and those paired with i. Thus, a single image feature can contribute to the foreground and adjacent

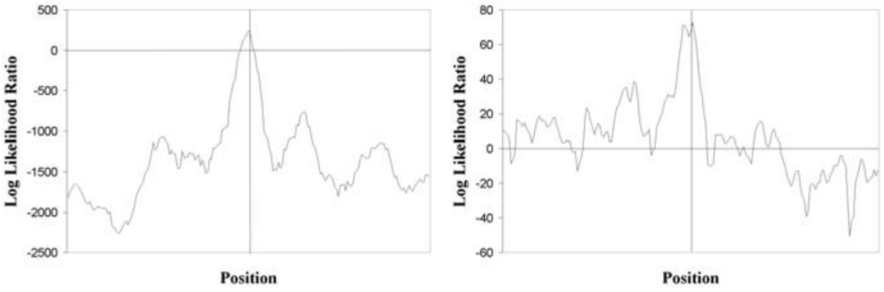

Fig. 5. Comparison of the spatial variation (plotted for a horizontal change of 200 pixels) of the learnt log likelihood ratios for the model presented here (left) and the edge-based model (right) of the head in the first image in Fig. 4. The correct position is centered and indicated by the vertical bar. Anything above the horizontal bar, corresponding to a likelihood ratio of 1, is more likely to be a head than not.

background appearance of several parts. When insufficient data is available to estimate either the foreground or the adjacent background histogram (as determined using an area threshold) the corresponding likelihood ratio is set to one.

3.2 Inter-Part Joint Constraints

A link is introduced between parts i and j if and only if they are physically connected neighbours. Each part has a set of control points that link it to its neighbours. A link has an associated value $LINK_{i,j}$ given by:

$$LINK_{i,j} = \begin{cases} 1 & \text{if } \delta_{i,j}/s < \varDelta_{i,j} \\ e^{(\delta_{i,j}/s - \varDelta_{i,j})/\sigma} & \text{otherwise} \end{cases} \qquad (3)$$

where $\delta_{i,j}$ is the image distance between the control points of the pair, $\varDelta_{i,j}$ is the maximum un-penalised distance and σ relates to the strength of penalisation. If the neighbouring parts do not link directly, because intervening parts are not instantiated, the un-penalised distance is found by summing the un-penalised distances over the complete chain. This can be interpreted as a force between parts equivalent to a telescopic rod with a spring on each end.

3.3 Learnt Paired Part Similarity

Certain pairs of body parts can be expected to have a similar foreground appearance to one another. For example, a person's upper left arm will nearly always have a similar colour and texture to the upper right arm. In the current system, the limbs are paired with their opposing parts. To encode this knowledge, a PDF of the divergence measure (computed using Equation (1)) between the foreground appearance histograms of paired parts and non-paired parts is learnt. Equation (4) shows the resulting likelihood ratio and Fig. 6 graphs this ratio. Fig. 7 shows a typical image projection of this ratio and shows

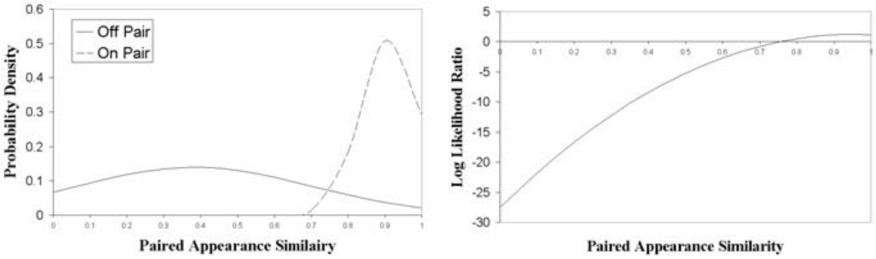

Fig. 6. Left: A plot of the learnt PDFs of foreground appearance similarity for paired and non-paired configurations. Right: The log of the resulting likelihood ratio. It can be seen, as would be expected, that more similar regions are more likely to be a pair.

the technique to be highly discriminatory. It limits possible configurations if one limb can be found reliably and helps reduce the likelihood of incorrect large assemblies.

$$PAIR_{i,j} = \frac{p(I(F_i, F_j)|on_i, on_j)}{p(I(F_i, F_j)|\overline{on_i, on_j})} \quad (4)$$

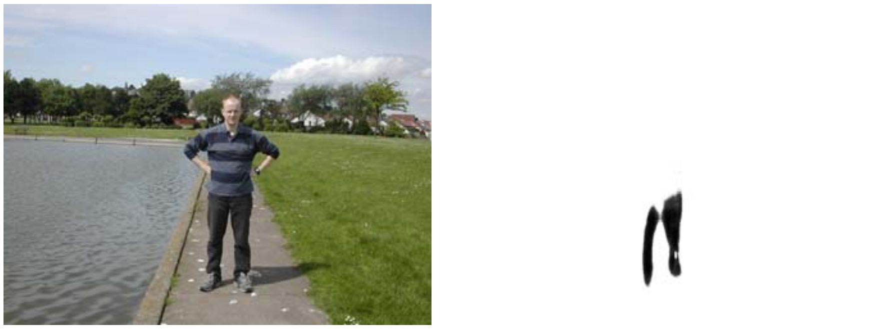

Fig. 7. Investigation of a paired part response. Left: an image for which significant limb candidates are found in the background. Right: the projection of the likelihood ratio for the paired response to the person's lower right leg in the image.

3.4 Combining the likelihoods

Learning the likelihood ratios allows a principled comparison of the various cues. The individual likelihood ratios are combined by assuming independence and the overall likelihood ratio is given by Equation(5). This rewards correct higher dimensional configurations over correct lower dimensional ones.

$$R = \prod_{i \in V} SINGLE_i \times \prod_{i,j \in V} PAIR_{i,j} \times \prod_{i,j \in V} LINK_{i,j} \quad (5)$$

3.5 Pose Estimation Results

The sampling scheme is described only briefly here as the emphasis of this paper is on a new formulation and likelihood. The search techniques will be more fully developed in future work. It is emphasised that the aim of the sampler is treated as one of maximisation rather than density estimation. The system begins by making a coarse regular scan of the image for the head and limbs. These results are then locally optimised. Part configurations are sampled from the resulting distribution and combined to form larger configurations and then optimised (in the full dimensional pose space, including the body part label) for a fixed period of time. It is envisaged that, due to the flexibility of the parametrisation, a set of optimization methods such as genetic style combination, prediction, local search, re-ordering and re-labelling can be combined using a scheduling algorithm and a shared sample population to achieve rapid, robust, global, high dimensional pose estimation. The system was implemented using an efficient, in-house C++ framework. Histograms with $8 \times 8 \times 8$ bins were used to represent a part's foreground and adjacent background appearance. The system samples single part configurations at the scale shown in Fig. 2 at approximately $3KHz$ from an image with resolution 640×480 on a 2GHz PC. Fig. 8 shows results of searching for partial pose configurations. It should be emphasised that although inter-part links are not visualised here, these results represent estimates of *pose configurations* with inter-part connectivity as opposed to independently detected parts. The scale of the model was fixed and the elongation parameter was constrained to be above 0.7.

4 Summary

A system was presented that allows detailed, efficient estimation of human pose from real-world images. The focus of the paper was the investigation of a novel likelihood model. The two key contributions were (i) a formulation that allowed the representation and comparison of partial (lower dimensional) solutions and modelled other object occlusion and (ii) a highly discriminatory learnt likelihood based upon probabilistic regions that allowed efficient body part detection. It should be stressed that this likelihood depends only on there being differences between a hypothesised part's foreground appearance and adjacent background appearance. It does not make use of scene-specific background models and is, as such, general and applicable to unconstrained scenes. The results presented confirm that it is possible to use partial configurations and a strong likelihood model to localise the body in real-world images. To improve the results, future work will need to address the following issues. A limited model of appearance was employed based on colour values. Texture orientation features should be employed to disambiguate overlapping parts (e.g. the arm lying over the torso). The model should be extended through closer consideration of the distinction between structural (kinematic) and visual segmentation of the body. The assumptions of independence between the individual likelihoods, particularly for the link and paired appearance likelihoods, needs investigation. Lastly, and perhaps most importantly, future work needs to improve the sampler to allow high dimensional configurations that contain self occlusion and visually similar neighbouring parts to be localised.

Acknowledgments: This work was funded by the UK EPSRC.

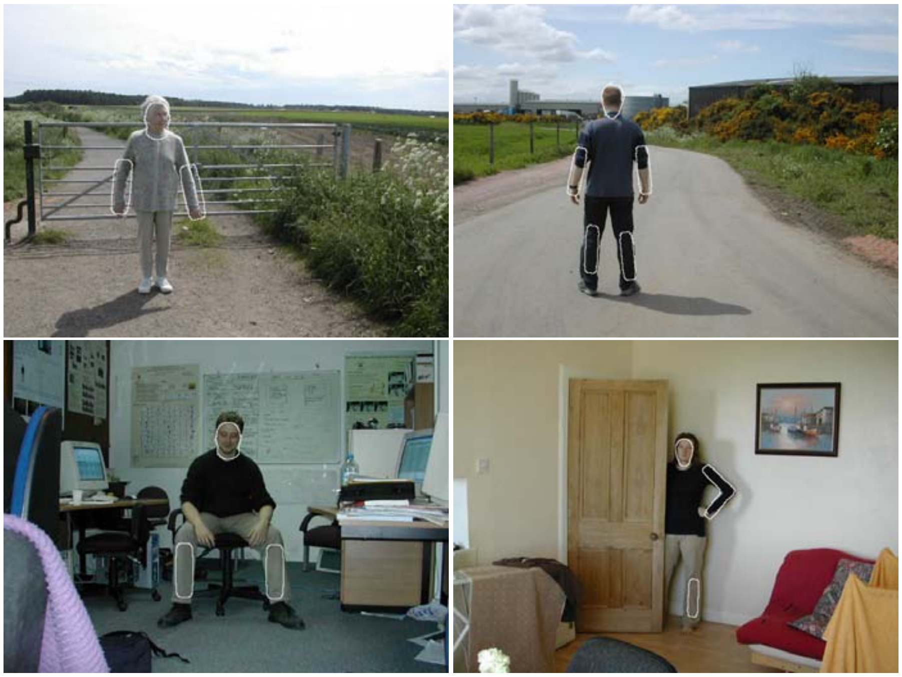

Fig. 8. Results from a search for partial pose configurations. The images are of both indoor and outdoor scenes and contain a significant amount of background clutter and in one case a door which partially occludes the subject. The samples with maximum score after searching for 2 minutes are shown.

References

1. D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999.
2. 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.
3. S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174–192, June 1999.
4. 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. 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.
6. 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.
7. D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, June 2003.
8. R. Ronfard, C. Schud, and B. Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714, Copenhagen, 2002.

9. 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.
10. 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.
11. 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.
12. 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.
13. B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.
14. 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.