# PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans Using Dynamic Graph-Cuts*

Matthieu Bray, Pushmeet Kohli, and Philip H.S. Torr

Dept. of Computing, Oxford Brookes University
{mbray, pushmeet.kohli, philiptorr}@brookes.ac.uk

**Abstract.** We present a novel algorithm for performing integrated segmentation and 3D pose estimation of a human body from multiple views. Unlike other related state of the art techniques which focus on either segmentation or pose estimation individually, our approach tackles these two tasks together. Normally, when optimizing for pose, it is traditional to use some fixed set of features, e.g. edges or chamfer maps. In contrast, our novel approach consists of optimizing a cost function based on a Markov Random Field (MRF). This has the advantage that we can use all the information in the image: edges, background and foreground appearances, as well as the prior information on the shape and pose of the subject and combine them in a Bayesian framework. Previously, optimizing such a cost function would have been computationally infeasible. However, our recent research in dynamic graph cuts allows this to be done much more efficiently than before. We demonstrate the efficacy of our approach on challenging motion sequences. Note that although we target the human pose inference problem in the paper, our method is completely generic and can be used to segment and infer the pose of any specified rigid, deformable or articulated object.

## 1 Introduction

Human pose inference is an important problem in computer vision standing at the crossroads of various applications ranging from Human Computer Interaction (HCI) to surveillance. The importance and complexity of this problem can be guaged by observing the number of papers which have tried to deal with it [1, 2, 3, 4, 5, 6]. In the last few years, several techniques have been proposed for tackling the pose inference problem, some of which have obtained decent results. In particular, the work of Agarwal and Triggs [1] using relevance vector machines and that of Shakhnarovich *et al.* [3] based on parametric sensitive hashing induced a lot interest and have been shown to give good results.

Most algorithms which perform pose estimation require the segmentation of humans as an essential introductory step [1, 2, 3]. This precondition limits the

* This work was supported by the EPSRC research grant GR/T21790/01(P) and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 642–655, 2006.

use of these techniques to scenarios where good segmentations are made available by enforcing strict studio conditions like blue-screening. Otherwise a preprocessing step must be performed in an attempt to segment the human, such as [7]. These approaches however cannot overcome the complexity of the problem of producing good segmentations for the general case of complex foreground and backgrounds (as will be seen in section 4), and where there are multiple objects in the scene or the camera/background is not stationary. Some pose inference methods exist which do not need segmentations. These rely on features such as chamfer distance [4], appearance [5], or edge and intensity [6]. However, none of these methods is able to efficiently utilize all the information present in an image and fail if the feature detector they are using fails. This is partly because the feature detector is not coupled to the knowledge of the pose of the object.

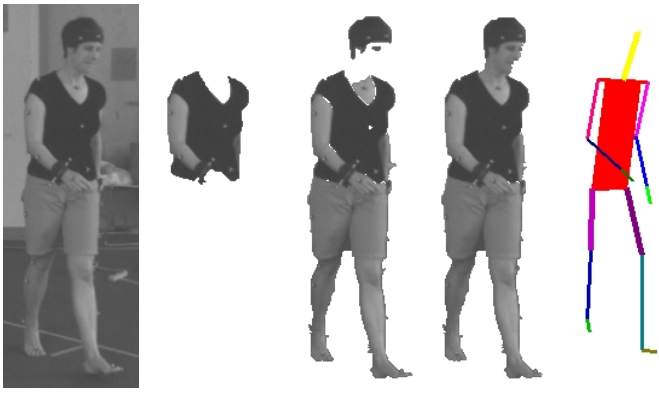
The question is then, how to simultaneously obtain the segmentation and human pose using all available information contained in the images?

Some elements of the answer to this question have been described by Kumar *et al.* [8]. Addressing the object segmentation problem, they report that the ***"samples from the Gibbs distribution defined by the MRF very rarely give rise to realistic shapes"***. As an illustration of this statement, figure 1(*b*) shows the segmentation result corresponding to the maximum a posteriori (MAP) solution of the Markov random Field (MRF) incorporating information about the image edges and appearances of the object and background. It can be clearly seen that this result is nowhere close to the ground truth.

**Shape priors and segmentation.** In recent years, a number of papers have tried to couple MRFs used for modelling the image segmentation problem, with information about the nature and shape of the object to be segmented [8, 10, 11]. One of the initial methods for combining MRFs with a shape prior was proposed by Huang *et al.* [10]. They incrementally found the MAP solution of an extended MRF[1] integrated with a probabilistic deformable model. By using belief propagation in the area surrounding the contour of this deformable model in an iterative manner, they were able to obtain a refined estimate of the contour. Their work however did not address the crucial problem of obtaining a object-like segmentation using prior information about the object which was later addressed by [8, 11].

The problem however was still far from being completely solved since objects in the real world change their shapes constantly and hence it is difficult to ascertain what would be a good choice for a prior on the shape. This complex and important problem was addressed by the work of Kumar *et al.* [8]. They modelled the segmentation problem by combining MRFs with layered pictorial structures (LPS) which provided them with a realistic shape prior described by a set of latent shape parameters. Their cost function was a weighted sum of the energy terms for different shape parameters (samples). The weights of this energy function were obtained by optimizing the labelling solution (background/foreground) using the Expectation-Maximization (EM) algorithm. During this optimization

[1] It is named an *extended* MRF due to the presence of an extra layer in the MRF to cope with the shape prior.



(a) (b) (c) (d) (e)

**Fig. 1.** Segmentation results corresponding to MRFs incorporating increasingly more information. (a) Original image. (b) The segmentation obtained corresponding to the MAP solution of a MRF consisting of colour likelihood and contrast terms as described in [9]. We give the exact formulation of this MRF in section 2.2. (c) The result obtained when the likelihood term of the MRF also takes into account the Gaussian Mixture Models (GMM) of individual pixel intensities as described in section 2.2. (d) Segmentation obtained after incorporating a 'pose-specific' shape prior in the MRF as explained in Section 2.3. The prior is represented as the distance transform of a stickman which guarantees a human-like segmentation. (e) The stickman model after optimization of its 3D pose (see Section 3). Observe how incorporating the individual pixel colour models in the MRF (c) gives a considerably better result than the one obtained using the standard appearance and contrast based representation (b). However the segmentation still misses the face of the subject. The incorporation of a stickman shape prior ensures a human-like segmentation (d) and provides simultaneously (after optimization) the 3D pose of the subject (e).

procedure, a graph cut had to be computed in order to obtain the segmentation score each time any parameter of the MRF was changed. This made their algorithm extremely computationally expensive.

Although their approach produced good results, it had some shortcomings. It was focused on obtaining good segmentations and did not furnish the pose of the object explicitly. Moreover, a lot of effort had to be spent to learn the exemplars for different parts of the LPS model. In the next section we will describe how we overcome the second limitation by using a simple articulated stickman model, which is not only efficiently renderable, but also provides a robust human-like segmentation and accurate pose estimate. To make our algorithm further computationally efficient we use the dynamic graph cut algorithm which was recently proposed in [12]. This new algorithm enables multiple graph cut computations, each computation taking a fraction of the time taken by the conventional graph cut algorithm if the change in the problem is small.

**Solving markov random fields using dynamic graph cuts.** A MRF is defined by its parameters and the observed data. A change in any of the two

thus causes a change in the MRF. If these changes are minimal, then intuitively the change in the MAP solution of the MRF should also be small. We made this observation and showed how dynamic graph cuts can be used to efficiently find the MAP solutions for MRFs that vary minimally from one time instant to the next [12]. The underlying idea of our paper was that of dynamic computation, where an algorithm solves a problem instance by dynamically updating the solution of the previous problem instance. Its goal is to be more efficient than a re-computation of the problem solution after every change from scratch. In the case of enormous problem instances and few changes, dynamic computation yields a substantial speed-up.

**Overview of the paper.** The paper proposes a novel algorithm for performing integrated segmentation and 3D pose estimation of a human body from multiple views. We do not require a feature extraction step but use all the data in the image. We formulate the problem in a Bayesian framework building on the object-specific MRF [8] and provide an efficient method for its solution called PoseCut. We include a human *pose-specific* shape prior in the MRF used for image segmentation, to obtain high quality segmentation results. We refer to this integrated model as a *pose-specific* MRF. As opposed to Kumar *et al.* [8], our approach does not require the laborious process of learning exemplars. Instead we use a simple articulated stickman model, which together with an MRF is used as our shape prior. Our experimental results show that this model suffices to ensure human-like segmentations.

Given an image, the solution of the pose-specific MRF is used to measure the quality of a 3D body pose. This cost function is then optimized over all pose parameters using dynamic graph cuts to provide both a object-like segmentation and the pose. The astute reader will notice that although we focus on the human pose inference problem, our method is in-fact general and can be used to segment and/or infer the pose of any object. We believe that our methodology is completely novel and we are not aware of any published methods which perform simultaneous segmentation and pose estimation. To summarize, the novelties of our approach include:

- An efficient method for combined object segmentation and pose estimation (PoseCut).
- Integration of a simple 'stickman prior' based on the skeleton of the object in a MRF to obtain a *pose-specific* MRF which helps us in obtaining high quality object pose estimate and segmentation results.

In the next section we give an intuitive insight into our framework. The pose-specific MRF and the different terms used in its construction are introduced in the same section. In section 3 we formulate the pose inference problem and describe the use of dynamic graph cuts for optimization in our problem construction. We present the experimental results obtained by our methods in section 4. These include comparison of our segmentation results with those obtained by some state of the art methods. We also show some results of simultaneous 3D pose estimation and segmentation. Our conclusions and the directions for future work are listed in Section 5.

## 2 Pose Specific MRF for Image Segmentation

In this section, we define an MRF-based energy function that gives the cost of any pose of a subject. We will optimize over this MRF using the Powell [13] minimization algorithm to infer the pose, and graph cuts to solve the segmentation as described in Section 3. The optimization of the energy is made efficient by the use of the dynamic graph cut algorithm [12].

Image segmentation has always remained an iconic problem of computer vision. The past few years have seen rapid progress made on it driven by the emergence of powerful optimization algorithms such as graph cuts. The early methods for performing image segmentation worked by coupling colour appearance information about the object and background with the edges present in an image to obtain good segmentations. However, this framework does not always guarantee good results. In particular, it fails in cases where the colour appearance models of the object and background are not discriminative as seen in figure 1($b$). The problem becomes even more pronounced in the case of humans where we have to deal with the various idiosyncracies of human clothing.

A semi-automated solution to this problem was explored by Boykov and Jolly [9] in their work on interactive image segmentation. They showed how users could refine segmentation results by specifying additional constraints. This can be done by labelling particular regions of the image as 'object' or 'background' and then computing the MAP solution of the MRF again. From their work, we made the following interesting observations: ***Simple user supplied shape cues used as rough priors for the object segmentation problem produced excellent results. The exact shape of the object can be induced from the edge information embedded in the image.*** Taking these into consideration, we hypothesized that the accurate exemplars used in [8] to generate shape priors were in-fact an overkill and could be replaced by a much simpler model.

**Stickman model.** Motivated by the observations made above, we decided against using a sophisticated shape prior. Instead, we used a simple articulated stickman model (shown in figure 1($e$)) to generate a rough pose-specific shape prior on the segmentation. As can been seen from the segmentation results in figure 1($d$), the stickman model helped us to obtain excellent segmentation results. The model has 26 degrees of freedom consisting of parameters defining absolute position and orientation of the torso, and the various joint angle values. There were no constraints or joint-limits incorporated in our model.

We now formally describe how the image segmentation problem can be modeled using a *pose-specific* MRF.

### 2.1 Markov Random Fields

A random field comprises of a set of discrete random variables $\{X_1, X_2, \ldots, X_n\}$ defined on the index set $\mathcal{V}$, such that each variable $X_v$ takes a value $x_v$ from the label set $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_l\}$ of all possible labels. We represent the set of all values $x_v, \forall v \in \mathcal{V}$ by the vector $\mathbf{x}$ which takes values in $\mathcal{X}^n$, and is referred to as

the configuration of the MRF. Further, we use $\mathcal{N}_v$ to denote the set consisting of indices of all variables which are neighbours of the random variable $X_v$ in the graphical model. This random field is said to be a MRF with respect to a neighborhood system $\mathcal{N} = \{\mathcal{N}_v | v \in \mathcal{V}\}$ if and only if it satisfies the positivity property: $\Pr(\mathbf{x}) > 0 \quad \forall \mathbf{x} \in \mathcal{X}^n$, and the Markovian property:

$$\Pr(x_v | \{x_u : u \in \mathcal{V} - \{v\}\}) = \Pr(x_v | \{x_u : u \in \mathcal{N}_v\}) \qquad \forall v \in \mathcal{V}. \tag{1}$$

Here we refer to $\Pr(X = \mathbf{x})$ by $\Pr(\mathbf{x})$ and $\Pr(X_i = x_i)$ by $\Pr(x_i)$. The MAP-MRF estimation problem can be formulated as an energy minimization problem where the energy corresponding to configuration $\mathbf{x}$ is the negative log likelihood of the joint posterior probability of the MRF and is defined as

$$E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{D}) + \text{const.} \tag{2}$$

where $\mathbf{D}$ is the observed data.

### 2.2 Image Segmentation as MAP-MRF Inference

In the context of image segmentation, $\mathcal{V}$ corresponds to the set of all image pixels, $\mathcal{N}$ is a neighbourhood defined on this set[2], the set $\mathcal{X}$ comprises of the labels representing the different image segments (which in our case are 'foreground' and 'background'), and the value $x_v$ denotes the labeling of the pixel $v$ of the image. Every configuration $\mathbf{x}$ of such an MRF defines a segmentation. The image segmentation problem can thus be solved by finding the least energy configuration of the MRF. The energy corresponding to a configuration $\mathbf{x}$ consists of a likelihood and a prior term as:

$$\Psi_1(\mathbf{x}) = \sum_{i \in \mathcal{V}} \left( \phi(\mathbf{D}|x_i) + \sum_{j \in \mathcal{N}_i} \psi(x_i, x_j) \right) + \text{const}, \tag{3}$$

where the prior $\psi(x_i, x_j)$ takes the form of a Generalized Potts model:

$$\psi(x_i, x_j) = \begin{cases} K_{ij} \text{ if } x_i \neq x_j, \\ 0 \quad \text{if } x_i = x_j. \end{cases} \tag{4}$$

The MRF used to model the image segmentation problem also contains a contrast term which favours pixels with similar colour having the same label [9, 14]. This is incorporated in the energy function by reducing the cost within the Potts model for two labels being different in proportion to the difference in intensities of their corresponding pixels. In our experiments, we use the term:

$$\gamma(i, j) = \lambda \exp\left( \frac{-g^2(i, j)}{2\sigma^2} \right) \frac{1}{\text{dist}(i, j)}, \tag{5}$$

[2] In this paper, we have used the standard 8-neighbourhood i.e. each pixel is connected to the 8 pixels surrounding it.

where $g^2(i,j)$ measures the difference in the RGB values of pixels $i$ and $j$ and $\text{dist}(i,j)$ gives the spatial distance between $i$ and $j$. This is a likelihood term (not prior) as it is based on the data, and hence has to be added separately from the smoothness prior. The energy function of the MRF now becomes

$$\Psi_2(\mathbf{x}) = \sum_{i \in \mathcal{V}} \left( \phi(\mathbf{D}|x_i) + \sum_{j \in \mathcal{N}_i} (\phi(\mathbf{D}|x_i, x_j) + \psi(x_i, x_j)) \right) \tag{6}$$

The contrast term of the energy function is defined as

$$\phi(\mathbf{D}|x_i, x_j) = \begin{cases} \gamma(i,j) & \text{if } x_i \neq x_j \\ 0 & \text{if } x_i = x_j. \end{cases} \tag{7}$$

The term $\phi(\mathbf{D}|x_i)$ in the MRF energy is the data log likelihood which imposes individual penalties for assigning any label $\mathcal{X}_k$ to pixel $i$. If we only take the appearance model into consideration, the likelihood is given by

$$\phi(\mathbf{D}|x_i) = -\log \Pr(i \in \mathcal{V}_k|\mathcal{H}_k) \qquad \text{if } x_i = \mathcal{X}_k \tag{8}$$

where $\mathcal{H}_k$ is the RGB (or for grey scale images, the intensity value) distribution for $\mathcal{S}_k$, the segment denoted by label $\mathcal{X}_k$[3]. The probability of a pixel belonging to a particular segment i.e. $\Pr(i \in \mathcal{S}_k|\mathcal{H}_k)$ is proportional to the likelihood $\Pr(I_i|\mathcal{H}_k)$, where $I_i$ is the colour intensity of the pixel $i$. As can be seen from figure 2($b$), this term is rather undiscriminating as the colours (grey intensity values in this case) included in the foreground histogram are similar to the ones included in the background histogram.

**Modeling pixel intensities as GMMs.** The MRF defined above for image segmentation performs poorly when segmenting images in which the appearance models of the foreground and background are not highly discriminative. When working on video sequences, we can use a background model developed using the Grimson-Stauffer [7] algorithm to improve our results. This algorithm works by representing the colour distribution of each pixel position in the video as a Gaussian Mixture Model (GMM). The likelihoods of a pixel for being background or foreground obtained by this technique are integrated in our MRF. Figure 1($c$) shows the segmentation result obtained after incorporating this information in our MRF formulation.

### 2.3 Incorporating the Pose-Specific Shape Prior

Though the results obtained from the above formulation look decent, they are not perfect. Note that there is no prior on the segmentation to look human like. Intuitively, incorporating such a constraint in the MRF would improve the final result obtained. In our case, this prior should be *pose-specific* as it depends on what pose the object (the human) is in. Kumar *et. al.* [8] in their work on

[3] In our problem, we have only 2 segments i.e. the foreground and the background.

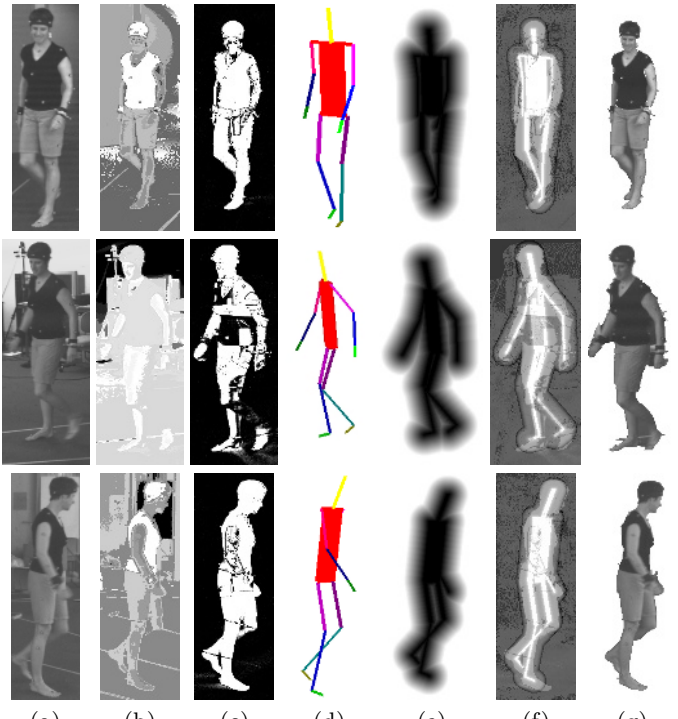



**Fig. 2.** (a) Original image. (b) The ratios of the likelihoods of pixels being labelled foreground/background ($\phi(\mathbf{D}|\mathbf{x}_i = \text{'fg'}) - \phi(\mathbf{D}|\mathbf{x}_i = \text{'bg'})$). These values are derived from the colour intensity histograms (see Section 2.2). (c) The segmentation results obtained by using the GMM models of pixel intensities. (d) The stickman in the optimal pose (see Sections 2.3 and 3). (e) The shape prior (distance transform) corresponding to the optimal pose of the stickman. (f) The ratio of the likelihoods of being labelled foreground/background using all the energy terms (colour histograms defining appearance models, GMMs for individual pixel intensities, and the pose-specific shape prior (see Sections 2.2, 2.2 and 2.3)) $\Psi_3(x_i = \text{'fg'}, \boldsymbol{\Theta}) - \Psi_3(x_i = \text{'bg'}, \boldsymbol{\Theta})$. (g) The segmentation result obtained from our algorithm which is the MAP solution of the energy $\Psi_3$ of the pose-specific MRF.

interleaved object recognition and segmentation, used the result of the recognition to develop a shape prior over the segmentation. This prior was defined by a set of latent variables which favoured segmentations of a specific pose of the object. They called this model the Object Category Specific MRF, which had the following energy function:

$$\Psi_3(\mathbf{x}, \boldsymbol{\Theta}) = \sum_i (\phi(\mathbf{D}|x_i) + \phi(x_i|\boldsymbol{\Theta}) + \sum_j (\phi(\mathbf{D}|x_i, x_j) + \psi(x_i, x_j))) \qquad (9)$$

with posterior $p(\mathbf{x}, \boldsymbol{\Theta}|\mathbf{D}) = \frac{1}{Z_3}\exp(-\Psi_3(\mathbf{x}, \boldsymbol{\Theta}))$. Here $\boldsymbol{\Theta}$ is used to denote the vector consisting of the object pose parameters. The shape-prior term of

the energy function for a particular pose of the human is shown in figure 2($e$). This is a distance transform generated from the stick-man model silhouette using the fast implementation of Felzenszwalb and Huttenlocher [15].

The function $\phi(x_i|\mathbf{\Theta})$ was choosen such that given an estimate of the location and shape of the object, pixels falling near to that shape were more likely to be labelled as 'foreground' and vice versa. It has the form: $\phi(x_i|\mathbf{\Theta}) = -\log p(x_i|\mathbf{\Theta})$. We follow the formulation of [8] and define $p(x_i|\mathbf{\Theta})$ as

$$p(x_i = \text{figure}|\mathbf{\Theta}) = 1 - p(x_i = \text{ground}|\mathbf{\Theta}) = \frac{1}{1 + \exp(\mu * (d(i, \mathbf{\Theta}) - d_r))}, \quad (10)$$

where $d(i, \mathbf{\Theta})$ is the distance of a pixel $i$ from the shape defined by $\mathbf{\Theta}$ (being negative if inside the shape). The parameter $d_r$ decides how 'fat' the shape should be, while parameter $\mu$ determines the ratio of the magnitude of the penalty that points outside the shape have to face compared to the points inside the shape.

### 2.4 MAP-MRF Inference Using Graph Cuts

Energies like the one defined in (9) can be solved using graph cuts if they are *sub-modular* [16]. The condition for sub-modularity is given as:

$$E(0,0) + E(1,1) \leq E(0,1) + E(1,0) \quad (11)$$

which implies that the energy for two labels taking similar values should be less than the energy for them taking different values. In our case, this is indeed the case and thus we can find the optimal configuration $\mathbf{x}^* = \min_{\mathbf{x}} \Psi_3(\mathbf{x}, \mathbf{\Theta})$ using a single graph cut. The labels of the latent variable in this configuration give the segmentation solution.

## 3 Formulating the Pose Inference Problem

Since the segmentation of an object depends on its estimated pose, we would like to make sure that our shape prior reflects the actual pose of the object. This takes us to our original problem of finding the pose of the human in an image. In order to solve this, we start with an initial guess of the object pose and optimize it to find the correct pose. When dealing with videos, a good starting point for this process would be the pose of the object in the previous frame. However, more sophisticated methods could be used based on object detection [17] at the expense of increasing the computation time.

One of the key contributions of this paper is to show how given an image of the object, the pose inference problem can be formulated in terms of a optimization problem over the MRF energy given in (9). Specifically, we solve the problem:

$$\mathbf{\Theta_{opt}} = \arg\min_{\mathbf{\Theta}}(\min_{\mathbf{x}} \Psi_3(\mathbf{x}, \mathbf{\Theta})). \quad (12)$$

Fig. 3 shows how $\min_{\mathbf{x}} \Psi_3(\mathbf{x}, \mathbf{\Theta})$ changes with rotation and translation of our shape prior. It can be clearly seen that the energy surface is uni-modal and hence can

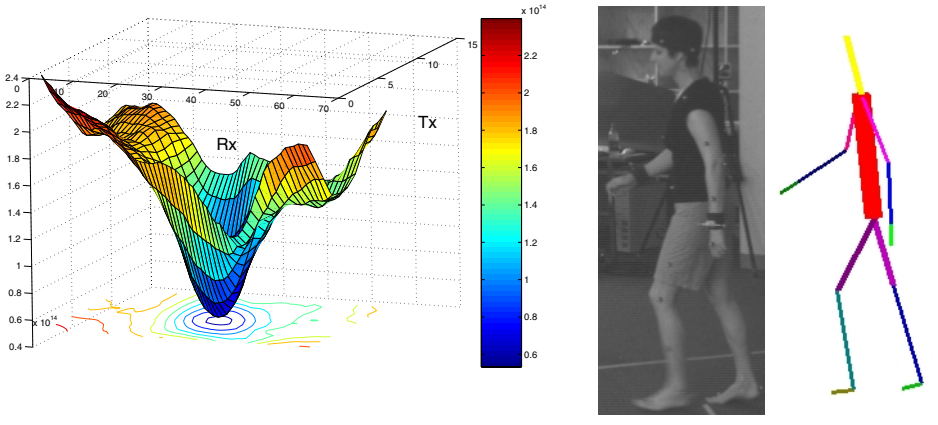



**Fig. 3.** a) The values of $\min_{\boldsymbol{x}} \Psi_3(\mathbf{x}, \boldsymbol{\Theta})$ obtained by varying the global translation and rotation of the shape prior in the x-axis. b) Original image. c) The pose obtained corresponding to the global minimum of the energy.

be optimized using any standard optimization algorithm like gradient descent. However, for more subtle joint angles, the energy is multi-modal, containing local minima. In our experiments, we used the Powell minimization [13] algorithm for optimization. When dealing with multiple views we solve the problem:

$$\boldsymbol{\Theta}_{\mathbf{opt}} = \arg\min_{\boldsymbol{\Theta}}(\min_{\mathbf{x}} \sum_{\text{views}} (\Psi_3(\mathbf{x}, \boldsymbol{\Theta})). \tag{13}$$

**Minimizing energies using dynamic graph cuts.** As explained earlier global minima of energies like the one defined in (9) can be found by graph cuts [16]. The time taken for computing a graph cut for a reasonably sized MRF is of the order of seconds. This would make our optimization algorithm extremely slow since we need to compute the global optimum of $\Psi_3(\mathbf{x}, \boldsymbol{\Theta})$ with respect to $\mathbf{x}$ multiple number times for different values of $\boldsymbol{\Theta}$. The graph cut computation can be made significantly faster by using the dynamic graph cut algorithm proposed recently in [12]. This algorithm works by using the solution of the previous graph cut computation for solving the new instance of the problem. We obtained a speed-up in the range of 15-20 times by using the dynamic graph cut algorithm.

## 4 Experiments

We now discuss the results obtained by PoseCut.

**Segmentation.** In order to demonstrate the performance of our method, we compare our segmentation results with those obtained by using the methods proposed in [7] and [18]. Bhatia *et al.* [18] learn a pixelwise background model

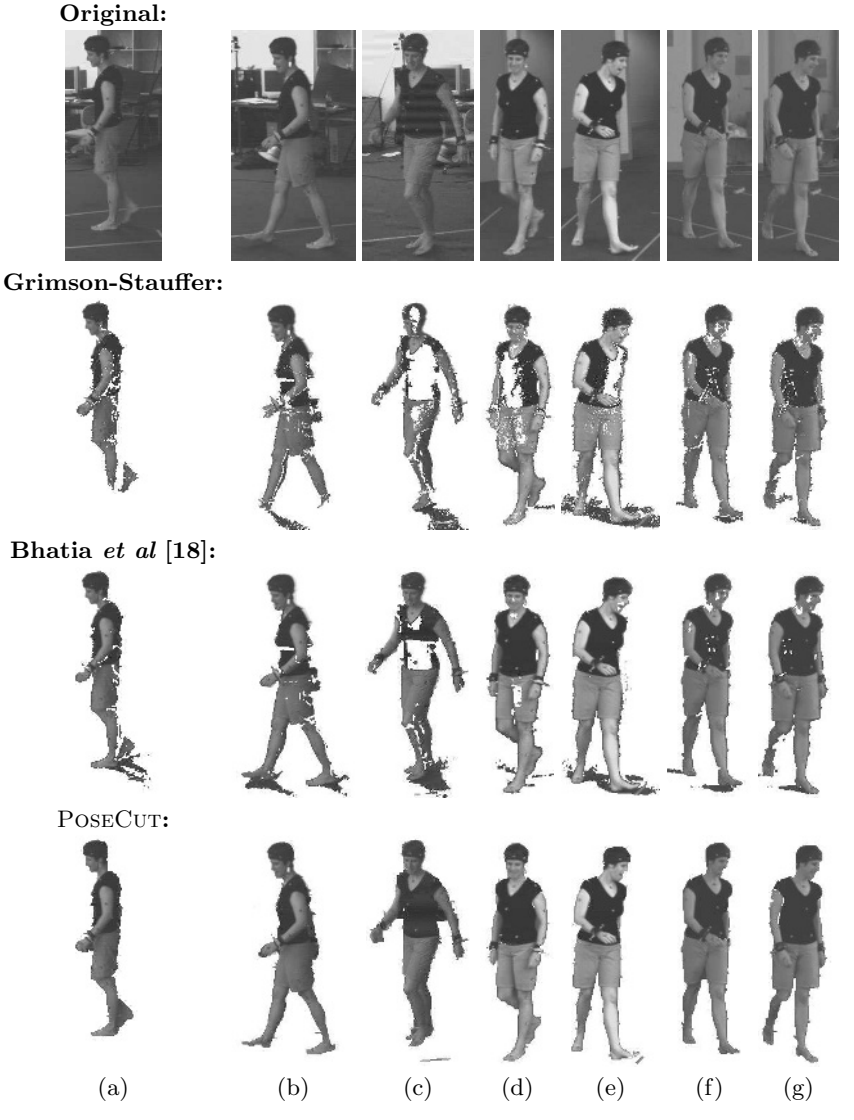



**Fig. 4.** Segmentation results obtained by Grimson-Stauffer, the method proposed by Bhatia *et al* [18] and PoseCut

represented by 3 Gaussians whose parameters are estimated by the Expectation-Maximization algorithm. They assume a uniform distribution for the likelihood of foreground pixels. It can be seen from the results in figure 4 that the segmentations obtained by using the methods of [7] and [18] are not accurate: They contain "speckles" and often segment the shadows of the feet as foreground. This is expected as they use only a pixelwise term to differentiate the background from the foreground and do not incorporate any spatial term which could offer a

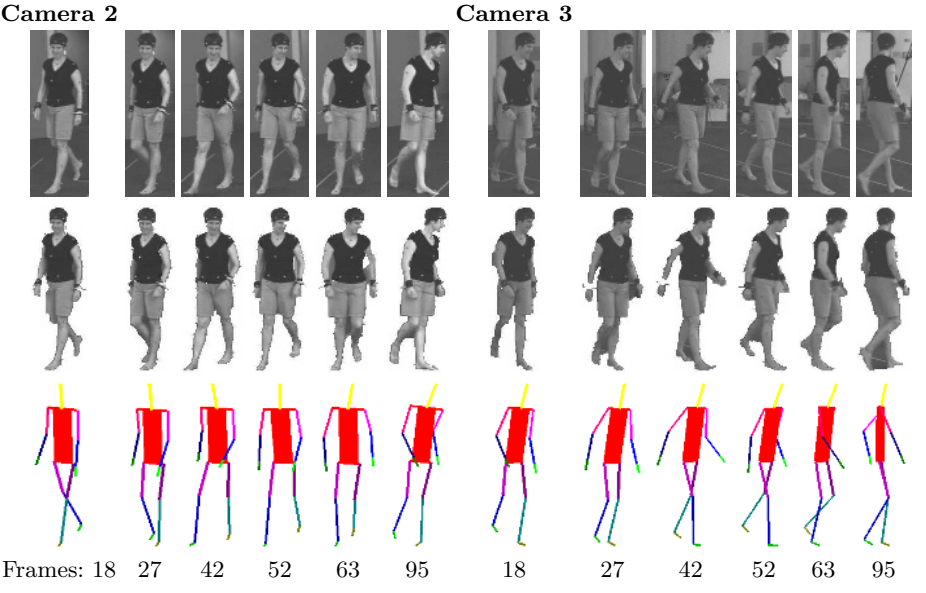



**Fig. 5.** Segmentation (middle) and pose estimation (bottom) results from PoseCut

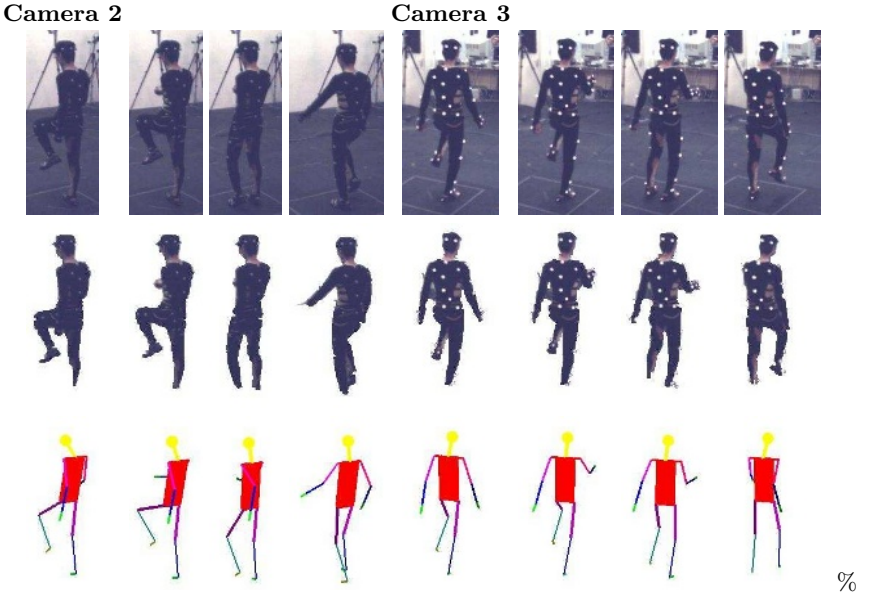



**Fig. 6.** Segmentation (middle row) and pose estimation (bottom row) results obtained by PoseCut. Observe that although the foreground and background appearances are similar, our algorithm is able to obtain good segmentations.

better "smoothing". In contrast, PoseCut which uses a pairwise potential term (as any standard graph cut approach) and a shape prior (which guarantees a human-like segmentation), is able to provide accurate results.

**Segmentation and pose estimation.** Figures 5 and 6 present the segmentations and the pose estimates obtained using PoseCut. The first data set comprises of three views of human walking circularly. The time needed for computation of the 3D pose estimate, on a PM 2GHz machine, when dealing with 644×484 images, is about 50 seconds per view[4]. As shown in these figures, the pose estimates match the original images accurately. In Figures 5 and 6, it should be noted that the appearance models of the foreground and background are quite similar: for instance, in Figure 6, the clothes of the subject are black in colour and the floor in the background is rather dark. The accuracy of the segmentation obtained in such challenging conditions demonstrates the robustness of PoseCut. An interesting fact to observe in Figure 5 about frame 95 is that the torso rotation of the stickman does not exactly conform with the original pose of the object. However, the segmentation of these frames is still accurate.

## 5 Conclusions and Future Work

The paper sets out a novel method for performing simultaneous segmentation and 3D pose estimation (PoseCut). The problem is formulated in a Bayesian framework which has the capability to utilize all information available (prior as well as observed data) to obtain good results. We showed how a rough pose-specific shape prior could be used to improve segmentation results significantly. We also gave a new formulation of the pose inference problem as an energy minimization problem and showed how it could be efficiently solved using dynamic graph cuts. The experiments demonstrate that our method is able to obtain excellent segmentation and pose estimation results.

It is common knowledge that the set of all human poses constitutes a low-dimensional manifold in the complete pose space. Optimizing over a parametrization of this low dimensional space instead of the 26D pose vector would intuitively improve both the accuracy and computation efficiency of our algorithm. Thus the use of dimensionality reduction algorithms is an important area to be investigated. The directions for future work also include using an appearance model per limb, which being more discriminative could help provide more accurate segmentations and pose estimates.

## References

1. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. In: CVPR. Volume II. (2004) 882–888
2. Kehl, R., Bray, M., Van Gool, L.: Full body tracking from multiple views using stochastic sampling. In: CVPR. Volume II. (2005) 129 – 136

[4] However, this could be speed up by computing the parameters of the MRF in an FPGA (Field-programmable gate array).

3. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: ICCV. (2003) 750–757
4. Gavrila, D., Davis, L.: 3D model-based tracking of humans in action: a multi-view approach. In: CVPR. (1996) 73–80
5. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3D human figures using 2D image motion. In: ECCV. (2000) 702–718
6. Sminchisescu, C., Triggs, B.: Covariance scaled sampling for monocular 3D body tracking. In: CVPR. (2001) 447–454
7. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR. (1999) 246–252
8. Kumar, M., Torr, P., Zisserman, A.: Obj cut. In: CVPR. Volume I. (2005) 18–25
9. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: ICCV. (2001) 105–112
10. Huang, R., Pavlovic, V., Metaxas, D.: A graphical model framework for coupling mrfs and deformable models. In: CVPR. Volume II. (2004) 739–746
11. Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: CVPR. Volume I. (2005) 755–762
12. Kohli, P., Torr, P.: Efficiently solving dynamic markov random fields using graph cuts. In: ICCV. (2005)
13. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: Numerical recipes in C. Cambridge Uni. Press (1988)
14. Blake, A., Rother, C., Brown, M., Pérez, P., Torr, P.: Interactive image segmentation using an adaptive gmmrf model. In: ECCV. Volume I. (2004) 428–441
15. Felzenszwalb, P., Huttenlocher, D.: Distance transforms of sampled functions. Technical Report TR2004-1963, Cornell University (2004)
16. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? In: ECCV. Volume III. (2002) 65 ff.
17. Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Filtering using a tree-based estimator. In: ICCV. (2003) 1063–1070
18. Bhatia, S., Sigal, L., Isard, M., Black, M.: 3d human limb detection using space carving and multi-view eigen models. In: ANM Workshop. Volume I. (2004) 17