Int J Comput Vis (2008) 79: 285–298
DOI 10.1007/s11263-007-0120-6

Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts

Pushmeet Kohli · Jonathan Rihan · Matthieu Bray · Philip H.S. Torr

Received: 7 April 2006 / Accepted: 26 December 2007 / Published online: 10 January 2008

Abstract This paper presents a novel algorithm for performing integrated segmentation and 3D pose estimation of a human body from multiple views. Unlike other state of the art methods which focus on either segmentation or pose estimation individually, our approach tackles these two tasks together. Our method works by optimizing a cost function based on a Conditional Random Field (CRF). This has the advantage that all information in the image (edges, background and foreground appearances), as well as the prior information on the shape and pose of the subject can be combined and used in a Bayesian framework. 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. 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 rigid, deformable or articulated object.

Keywords Pose estimation · Segmentation · Energy minimization

1 Introduction

Human pose inference is an important problem in computer vision. It stands at the crossroads of various important applications ranging from Human Computer Interaction (HCI) to surveillance. The importance and complexity of this problem can be gauged by the number of papers which have tried to deal with it (Agarwal and Triggs 2004; Kehl et al. 2005; Shakhnarovich et al. 2003; Gavrila and Davis 1996; Sidenbladh et al. 2000a, 2000b; Sminchisescu and Triggs 2001; Urtasun et al. 2005; Lan and Huttenlocher 2005; Deutscher et al. 2001; Mori et al. 2004; Ramanan and Forsyth 2003; Felzenszwalb and Huttenlocher 2000). Most algorithms which perform pose estimation require the segmentation of humans as an essential introductory step (Agarwal and Triggs 2004; Kehl et al. 2005; Shakhnarovich et al. 2003). This precondition limits the 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 (Stauffer and Grimson 1999). These approaches however cannot obtain good segmentations in challenging scenarios which have: complex foreground and background, multiple objects in the scene, and moving camera/background. Some pose inference methods exist which do not need segmentations. These rely on features such as chamfer distance (Gavrila and Davis 1996), appearance (Sidenbladh et al. 2000a, 2000b), or edge and intensity (Sminchisescu and Triggs 2001). 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 and nature of the object to be segmented.

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. (2005). Addressing the object segmentation problem, they report that "*samples from the Gibbs*

P. Kohli (✉) · J. Rihan · M. Bray · P.H.S. Torr
Department of Computing, Oxford Brookes University, Wheatley Campus, Oxford OX33 1HX, UK
e-mail: pushmeet.kohli@brookes.ac.uk

P.H.S. Torr
e-mail: philiptorr@brookes.ac.uk

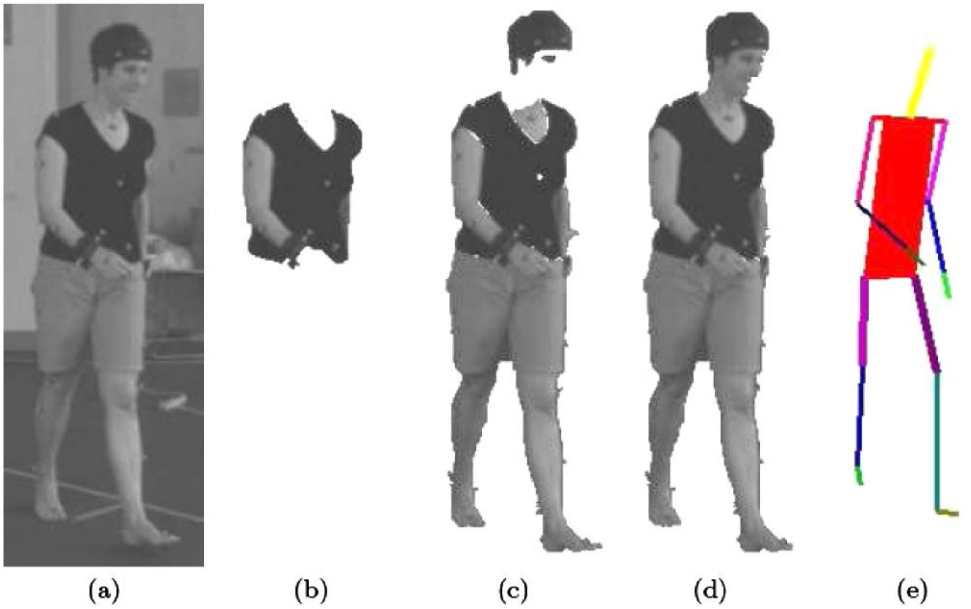

Fig. 1 Improving segmentation results by incorporating more information in the CRF. (**a**) Original image. (**b**) The segmentation obtained corresponding to the MAP solution of a CRF consisting of colour likelihood and contrast terms as described in (Boykov and Jolly 2001). We give the exact formulation of this CRF in Sect. 2.2. (**c**) The result obtained when the likelihood term of the CRF also takes into account the Gaussian Mixture Models (GMM) of individual pixel intensities as described in Sect. 2.2. (**d**) Segmentation obtained after incorporating a 'pose-specific' shape prior in the CRF as explained in Sect. 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 Sect. 3). Observe how incorporating the individual pixel colour models in the CRF (**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**)

distribution defined by the Markov Random Field (MRF) very rarely give rise to realistic shapes". As an illustration of this statement, Fig. 1(b) shows the segmentation result corresponding to the maximum a posteriori (MAP) solution of the Conditional Random Field (CRF) 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 or CRFs used for modeling the image segmentation problem, with information about the nature and shape of the object to be segmented (Kumar et al. 2005; Huang et al. 2004; Freedman and Zhang 2005; Zhao and Davis 2005). One of the earliest methods for combining MRFs with a shape prior was proposed by Huang et al. (2004). They incrementally found the MAP solution of an extended MRF[1] integrated with a probabilistic deformable model. They were able to obtain a refined estimate of the object contour by using belief propagation in the area surrounding the contour of this deformable model. This process was iterated till convergence.

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

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. (2005). They modeled the segmentation problem by combining CRFs 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 using the Expectation-Maximization (EM) algorithm. During this optimization procedure, a graph cut had to be computed in order to obtain the segmentation score each time any parameter of the CRF 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 provide 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. Recently, Zhao and Davis (2005) exploited the idea of object-specific segmentation in improving object recognition or detection. Their method worked by coupling the twin problems of object detection and segmentation in a single framework. They

matched exemplars to objects in the image using chamfer matching and thus like (Kumar et al. 2005) also suffered from the problem of maintaining a huge exemplar set for complex objects.

Shape Priors in Level Sets Prior knowledge about the shape to be segmented has also been used in level set methods for obtaining an object segmentation. Like (Kumar et al. 2005) these methods learn the prior using a number of training shapes. Leventon et al. (2000) performed principal component analysis on these shapes to get a embedding function which was integrated in the evolution equation of the level set. More recently, Cremers et al. (2006) have used kernel density estimation and intrinsic alignment to embed more complex shape distributions. Compared to Kumar et al. (2005) and Zhao and Davis (2005) these methods have a more compact representation of the shape prior. However, they suffer from the disadvantage that equations for level set evolution may not lead to the globally optimal solution.

In the next section we will describe how we overcome the problem of maintaining a huge exemplar set 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 computationally efficient we use the dynamic graph cut algorithm which was recently proposed in Kohli and Torr (2005)). 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 Random Fields using Dynamic Graph Cuts Inferring the most probable solution of a MRF or CRF involves minimizing the energy function which characterizes it. This energy is defined by some CRF parameters and the data. A change in any of the two causes a change in the energy. If these changes are minimal, then intuitively the change in the MAP solution of the CRF 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 or CRFs that vary minimally from one time instant to the next (Kohli and Torr 2005). The underlying idea of our paper was 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 solution after every change from scratch. In the case of large problem instances and few changes, dynamic computation yields a substantial speed-up.

Human Pose Estimation In the last few years, several techniques have been proposed for tackling the pose inference problem. In particular, the works of Agarwal and Triggs (2004) using relevance vector machines and that of Shakhnarovich et al. (2003) based on parametric sensitive hashing induced a lot of interest and have been shown to give good results. Some methods for human pose estimation in monocular images use a tree-structured model to capture the kinematic relations between parts such as the torso and limbs (Mori et al. 2004; Ramanan and Forsyth 2003; Felzenszwalb and Huttenlocher 2000). They then use efficient inference algorithms to perform exact inference in such models. In their recent work, Lan and Huttenlocher (2005) show how the tree-structured restriction can be overcome while not greatly increasing the computational cost of estimation.

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.[2] 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 CRF (Kumar et al. 2005) and provide an efficient method for its solution called POSECUT. We include a human *pose-specific* shape prior in the CRF used for image segmentation, to obtain high quality segmentation results. We refer to this integrated model as a *pose-specific* CRF. Unlike Kumar et al. (2005), our approach does not require the laborious process of learning exemplars. Instead we use a simple articulated stickman model, which together with an CRF is used as our shape prior. The experimental results show that this model suffices to ensure human-like segmentations.

Given an image, the solution of the pose-specific CRF 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 CRF to obtain a *pose-specific* CRF 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 CRF and the different terms

[2] A shorter version of this paper earlier appeared as (Bray et al. 2006). This extended version contains a more thorough discussion of the behavior of the optimization algorithm and additional quantitative and qualitative results.

Fig. 2 Interactive Image Segmentation. The figure shows how good segmentation results can be obtained using a set of rough region cues supplied by the user. (**a**) An image with user specified segmentation cues (shown in *blue* and *red*). These cues were used to obtain the segmentation shown in image (**b**). This segmentation is not perfect and can be improved by specifying additional cues which are shown in (**b**). The final segmentation result is shown in image (**c**)

used in its construction are introduced in the same section. In Sect. 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 Sect. 4. These include qualitative and quantitative results on challenging data sets. We compare 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. Section 5 discusses the extension of our work for object detection. The conclusions and the directions for future work are listed in Sect. 6.

2 Pose Specific CRF for Image Segmentation

In this section we define a CRF based energy function that gives the cost of any pose of a subject. This energy function is minimized using the Press et al. (1988) minimization algorithm and graph cuts to obtain the pose and segmentation of the human as described in Sect. 3. The optimization of the energy is made efficient by the use of the dynamic graph cut algorithm (Kohli and Torr 2005).

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. 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 Fig. 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 (2001) in their work on interactive image segmentation. They showed how users could refine segmentation results by specifying additional constraints. This can be done by labeling particular regions of the image as 'object' or 'background' and then computing the MAP solution of the CRF again. The interactive image segmentation process is illustrated in Fig. 2. 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 Kumar et al. (2005) to generate shape priors were in-fact an overkill and could be replaced by much simpler models. Motivated by these observations we decided against using a sophisticated shape prior. We have used two simple models in our work which are described below.

Stickman Model We used a simple articulated stickman model for the full body human pose estimation problem. The model is shown in Fig. 1(e). It is used to generate a rough pose-specific shape prior on the segmentation. As can been seen from the segmentation results in Fig. 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.

The Upper body Model The second model was primarily designed for the problem of segmenting the human speaker in video conference scenarios. The model can be seen in Fig. 3. It is parameterized by 6 parameters which encode the x and y location of the two shoulders and the length and angle of the neck.

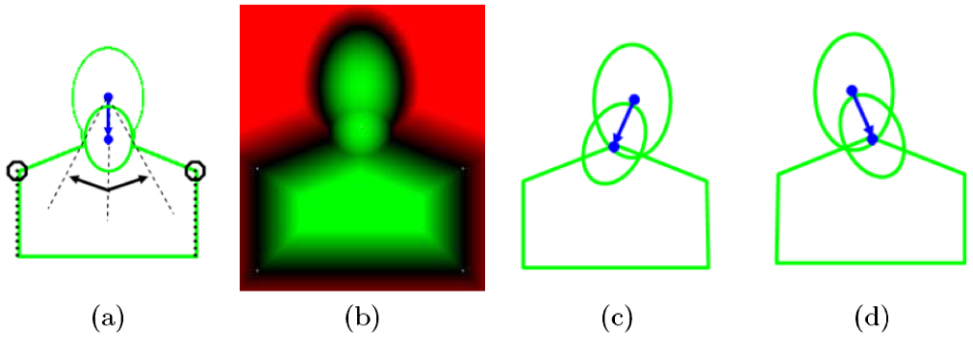

Fig. 3 Upper Body Model. (**a**) The model parameterized by 6 parameters encoding the x and y location of the two shoulders and the length and angle of the neck. (**b**) The shape prior generated using the model. Pixels more likely to belong to the foreground/background are green/red. (**c**) and (**d**) The model rendered in two poses

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

2.1 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 random field. Further, we use $\mathcal{N}_v$ to denote the set consisting of indices of all variables which are neighbors of the random variable X_v in the graphical model. A 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: $\Pr(\mathbf{x}) > 0\ \forall \mathbf{x} \in \mathcal{X}^n$, and the Markovian property:

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

Here we refer to $\Pr(X = \mathbf{x})$ by $\Pr(\mathbf{x})$ and $\Pr(X_i = x_i)$ by $\Pr(x_i)$. A conditional random field (CRF) may be viewed as an MRF globally conditioned on the data.

The problem of finding the most probable solution of the CRF 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 CRF and is defined as

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

where $\mathbf{D}$ is the observed data. The minimization problem is independent of the partition function of the probability distribution.

2.2 CRFs for Image Segmentation

I n the context of image segmentation, $\mathcal{V}$ corresponds to the set of all image pixels, $\mathcal{N}$ is a neighborhood defined on this set,[3] 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 CRF defines a segmentation. The image segmentation problem can thus be solved by finding the least energy configuration of the CRF. The energy function characterizing the CRFs used for image segmentation can be written as a sum of likelihood ($\phi(\mathbf{D}|x_i)$) and prior ($\psi(x_i, x_j)$) terms as:

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

The term $\phi(\mathbf{D}|x_i)$ in the CRF 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

$$\phi(\mathbf{D}|x_i) = -\log \Pr(i \in \mathcal{V}_k | \mathcal{H}_k) \quad \text{if } x_i = \mathcal{X}_k \tag{4}$$

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 $\mathcal{X}_k$.[4] 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 Fig. 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.

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

$$\psi(x_i, x_j) = \begin{cases} K_{ij} & \text{if } x_i \neq x_j, \\ 0 & \text{if } x_i = x_j. \end{cases} \tag{5}$$

The CRF used to model the image segmentation problem also contains a contrast term which favors pixels with similar colour having the same label (Boykov and Jolly 2001;

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

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

Blake et al. 2004). This is incorporated in the energy function by increasing the cost within the Potts model for two neighboring variables being different in proportion to the similarity in intensities of their corresponding pixels. In our experiments, we use the function:

$$\gamma(i,j) = \lambda \exp\left(\frac{-g^2(i,j)}{2\sigma^2}\right)\frac{1}{\text{dist}(i,j)}, \tag{6}$$

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 CRF now becomes

$$\Psi_2(\mathbf{x}) = \sum_{i\in\mathcal{V}}\left(\phi(\mathbf{D}|x_i) + \sum_{j\in\mathcal{N}_i}(\phi(\mathbf{D}|x_i,x_j) + \psi(x_i,x_j))\right). \tag{7}$$

The contrast term of the energy function is defined as

$$\phi(\mathbf{D}|x_i,x_j) = \begin{cases}\gamma(i,j) & \text{if } x_i \neq x_j \\ 0 & \text{if } x_i = x_j.\end{cases} \tag{8}$$

By adding this term to the energy, we have diverged from the strict definition of a MRF. The resulting energy function now characterizes a Conditional Random Field (Lafferty et al. 2001).

Modeling Pixel Intensities as GMMs The CRF 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 (Stauffer and Grimson 1999) 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 CRF. Figure 1(c) shows the segmentation result obtained after incorporating this information in our CRF 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 CRF would improve the segmentation. In our case, this prior should be *pose-specific* as it depends on what pose the object (the human) is in. Kumar et al. (2005) in their work on 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 favored segmentations of a specific pose of the object. They called this model the Object Category Specific CRF, which had the following energy function:

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

with posterior $p(\mathbf{x},\boldsymbol{\Theta}|\mathbf{D}) = \frac{1}{Z_3}\exp(-\Psi_3(\mathbf{x},\boldsymbol{\Theta}))$. Here $\boldsymbol{\Theta} \in \mathbb{R}_p$ is used to denote the vector of the object pose parameters. The shape-prior term of the energy function for a particular pose of the human is shown in Fig. 4(e). This is a distance transform generated from the stick-man model silhouette using the fast implementation of Felzenszwalb and Huttenlocher (2004).

The function $\phi(x_i|\boldsymbol{\Theta})$ was chosen such that given an estimate of the location and shape of the object, pixels falling near to that shape were more likely to be labeled as 'foreground' and vice versa. It has the form: $\phi(x_i|\boldsymbol{\Theta}) = -\log p(x_i|\boldsymbol{\Theta})$. We follow the formulation of Kumar et al. (2005) and define $p(x_i|\boldsymbol{\Theta})$ as

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

where $d(i,\boldsymbol{\Theta})$ is the distance of a pixel i from the shape defined by $\boldsymbol{\Theta}$ (being negative if inside the shape). The parameter d_r decides how 'fat' the shape should be, while parameter μ 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 Inference in the CRF Using Graph Cuts

Energy functions like the one defined in (9) can be solved using graph cuts if they are *sub-modular* (Kolmogorov and Zabih 2002). A function $f : \{0,1\} \to \mathbb{R}$ is submodular if and only if all its projections on 2 variables ($f^p : \{0,1\}^2 \to \mathbb{R}$) satisfy:

$$f^p(0,0) + f^p(1,1) \leq f^p(0,1) + f^p(1,0). \tag{11}$$

For the pairwise potentials, this condition can be seen as implying 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 $\mathbf{x}^* = \min_{\mathbf{x}} \Psi_3(\mathbf{x},\boldsymbol{\Theta})$ using a single graph cut. The labels of the latent variable in this configuration give the segmentation solution.

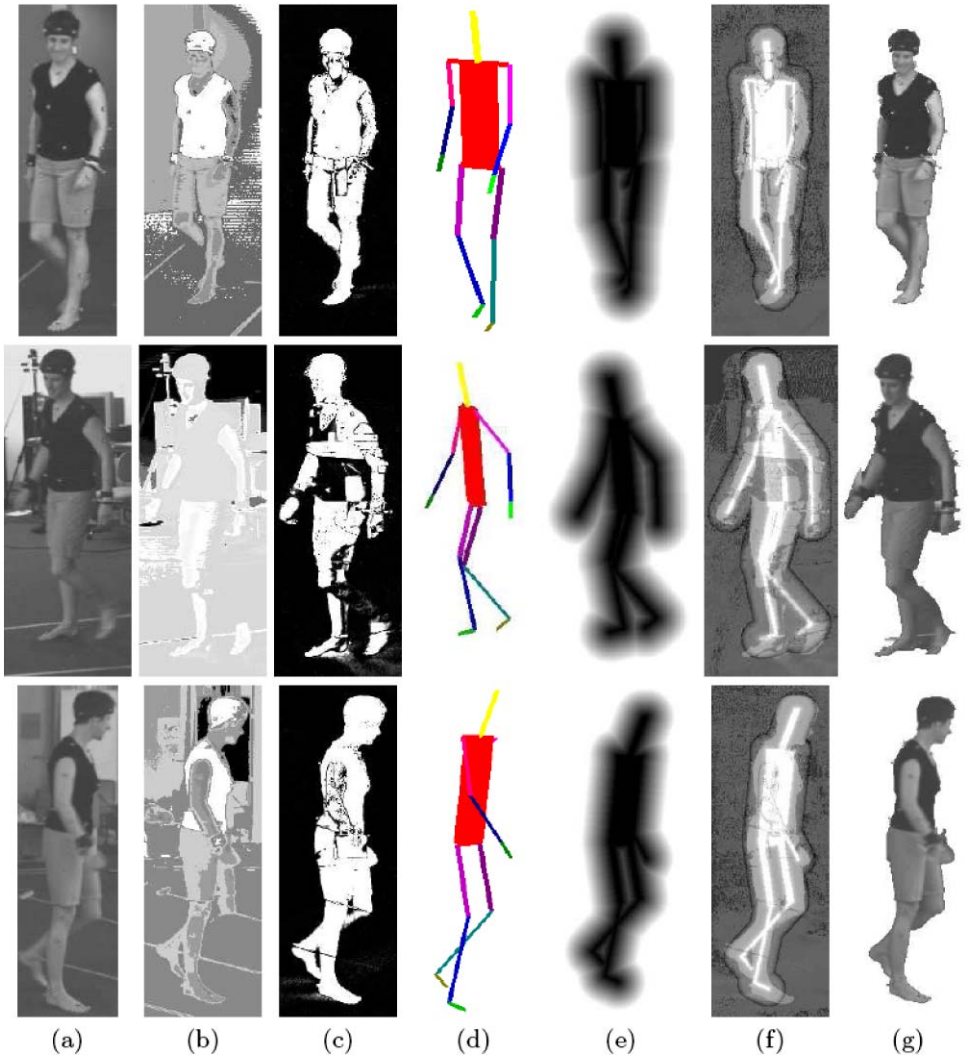

Fig. 4 Different terms of our pose specific CRF. (**a**) Original image. (**b**) The ratios of the likelihoods of pixels being labeled foreground/background ($\phi(\mathbf{D}|\mathbf{x}_i = \text{'fg'}) - \phi(\mathbf{D}|\mathbf{x}_i = \text{'bg'})$). These values are derived from the colour intensity histograms (see Sect. 2.2). (**c**) The segmentation results obtained by using the GMM models of pixel intensities. (**d**) The stickman in the optimal pose (see Sects. 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 labeled foreground/background using all the energy terms (colour histograms defining appearance models, GMMs for individual pixel intensities, and the pose-specific shape prior (see Sects. 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 Ψ_3 of the pose-specific CRF

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 (Stenger et al. 2003) 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 an optimization problem over the CRF energy given in (9). Specifically, we solve the problem:

$$\boldsymbol{\Theta}_{\mathbf{opt}} = \arg\min_{\boldsymbol{\Theta},\mathbf{x}} \Psi_3(\mathbf{x}, \boldsymbol{\Theta}). \tag{12}$$

The minimization problem defined above contains both discrete ($\mathbf{x} \in \{0,1\}^n$) and continuous ($\boldsymbol{\Theta} \in \mathbb{R}_P$) valued variables and thus is a mixed integer programming problem. The large number of variables involved in the energy function $\Psi_3(\mathbf{x}, \boldsymbol{\Theta})$ make it especially challenging to minimize. To solve the minimization problem (12), we decompose it as: $\boldsymbol{\Theta}_{\mathbf{opt}} = \arg\min_{\boldsymbol{\Theta}} \mathcal{F}(\boldsymbol{\Theta})$, where

$$\mathcal{F}(\boldsymbol{\Theta}) = \min_{\mathbf{x}} \Psi_3(\mathbf{x}, \boldsymbol{\Theta}). \tag{13}$$

For any value of $\boldsymbol{\Theta}$, the function $\Psi_3(\mathbf{x}, \boldsymbol{\Theta})$ is submodular in $\mathbf{x}$ and thus can be minimized in polynomial time by solving a single st-mincut problem to give the value of $\mathcal{F}(\boldsymbol{\Theta})$.

We will now explain how we minimize $\mathcal{F}(\boldsymbol{\Theta})$ to get the optimal value of the pose parameters. Figure 5 shows how the function $\mathcal{F}(\boldsymbol{\Theta})$ depends on parameters encoding the rotation and translation of our stickman model in the x-axes. It can be seen that the function surface is unimodal in a large neighborhood of the optimal solution. Hence, given a good initialization of the pose $\boldsymbol{\Theta}$, it can be reliably optimized us-

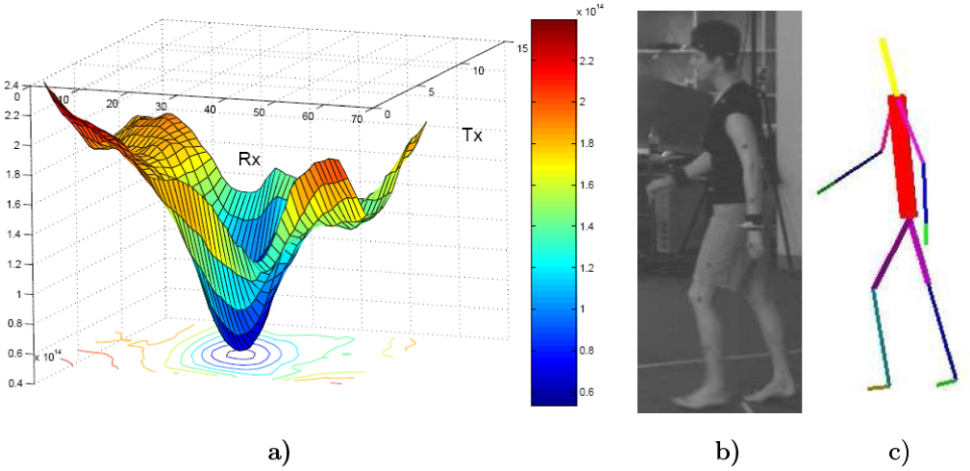

Fig. 5 Inferring the optimal pose. (**a**) The values of $\min_{\bar{x}} \Psi_3(\mathbf{x}, \boldsymbol{\Theta})$ obtained by varying the global translation and rotation of the shape prior in the x-axis. (**b**) Original image. (**c**) The pose obtained corresponding to the global minimum of the energy

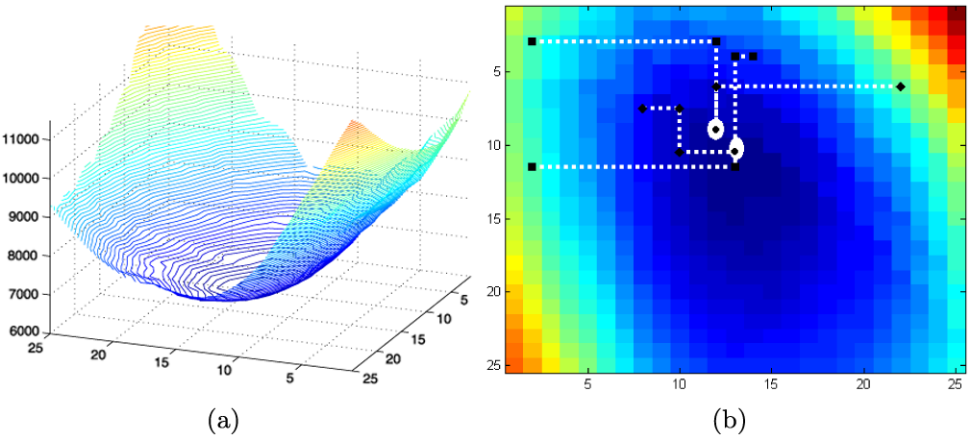

Fig. 6 Optimizing the pose parameters. (**a**) The values of $\min_{\bar{x}} \Psi_3(\mathbf{x}, \boldsymbol{\Theta})$ obtained by varying the rotation and length parameters of the neck. (**b**) The image shows five runs of the Powell minimization algorithm which are started from different initial solutions. The runs converge on two solutions which are very close and have almost the same energy

ing any standard optimization algorithm like gradient descent. In our experiments, we used the Powell minimization (Press et al. 1988) algorithm for optimization.

Figure 6(a) shows how the function $\mathcal{F}(\boldsymbol{\Theta})$ changes with changes to the neck angle and length parameters of the upper body model shown in Fig. 3. As in the case of the 3D stickman model the energy surface is well behaved near the optimal pose parameters. Our experiments showed that the Powell minimization algorithm is able to converge to almost the same point for different initializations (see Fig. 6(b)).

Failure Modes It can be seen that the function $\mathcal{F}(\boldsymbol{\Theta})$ is not unimodal over the whole domain and contains local minima. This multi-modality of $\mathcal{F}(\boldsymbol{\Theta})$ can cause a gradient descent algorithm to get trapped and converge to a local minimum. In our experiments we observed that these spurious minima lie quite far from the globally optimal solution. We also observed that the pose of the human subject generally does not change substantially from one frame to the next. This lead us to use the pose estimate from the previous frame as an initialization for the current frame. This good initialization for the pose estimate made sure that spurious minima do not effect our method.

The failure rate of our method can be further improved by using object detection systems which provide a better initialization of the pose of the object. Scenarios where the method still converges to a local minima can be detected and dealt with using the strategy discussed in Sect. 5 which was used in our recent work on object detection and segmentation (Rihan et al. 2006).

Resolving Ambiguity Using Multiple Views The problem of estimating the 3D pose of the human from monocular images suffers from ambiguity. This arises from the one-many nature of the mapping that relates a human shape and the corresponding 3D human pose. In other words, many possible 3D poses can explain the same human shape, and thus will have the same energy. This multi-modality of the energy function may result in our algorithm producing a wrong pose estimate.

The ambiguity in 3D pose can be resolved by using multiple views of the object ('human'). Our method has the advantage that information from multiple views can be integrated very easily into a single optimization framework.

Specifically, when dealing with multiple views, we solve the problem:

$$\boldsymbol{\Theta}_{\mathbf{opt}} = \arg\min_{\boldsymbol{\Theta}} \left(\sum_{\text{Views}} \min_{\mathbf{x}} (\Psi_3(\mathbf{x}, \boldsymbol{\Theta})) \right). \tag{14}$$

The framework is illustrated in Fig. 7. An alternative approach to deal with the pose ambiguity problem is to use a dynamic model for the pose (Agarwal and Triggs 2006). Such a model, given the correct pose in a particular image frame, can produce a small set of plausible poses for the subsequent image frame. However, this method is only applicable when we are dealing with videos.

Dynamic Energy Minimization Using Graph Cuts As explained earlier global minima of energies like the one defined in (9) can be found by graph cuts (Kolmogorov and Zabih 2002). The time taken for computing a graph cut for a reasonably sized CRF 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 $\boldsymbol{\Theta}$.

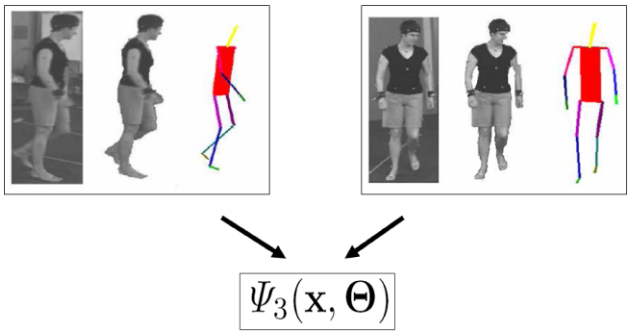

Fig. 7 Resolving ambiguity using multiple views. The figure shows how information from different views of the human can be integrated in a single energy function, which can be used to find the true pose of the human subject

The graph cut computation can be made significantly faster by using the dynamic graph cut algorithm proposed recently in Kohli and Torr (2005). 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 our method. We provide the segmentation and pose estimation results individually.

4.1 Segmentation Results

As expected, the experimental results show that the segmentation results improve considerably as we increase the amount of information in our CRF framework. Figure 8 shows how integrating more information in the CRF improves the segmentation results. Quantitative results for the segmentation problem are shown in Table 1.

In order to demonstrate the performance of our method, we compare our segmentation results with those obtained using the method proposed in Stauffer and Grimson (1999). It can be seen from the results in Fig. 10 that the segmentations obtained using the method of Stauffer and Grimson (1999) 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 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.

Fig. 8 Results showing the effect of incorporating a shape prior on the segmentation results. The first image is the original image to be segmented. The second, third and fourth images shows the segmentation results obtained using colour, colour + smoothness prior and colour + smoothness + shape information respectively

Table 1 Quantitative segmentation results. The table shows the effect of adding more information in the Bayesian framework on the quantitative segmentation accuracy. The accuracy was computed over all the pixels in the image. The ground truth for the data used in this experiment was generated by hand labeling the foreground and background regions in the images

Information used	Correct object pixels	All correct pixels
Colour	45.73%	95.2%
Colour + GMM	82.48%	96.9%
Colour + GMM + Shape	97.43%	99.4%

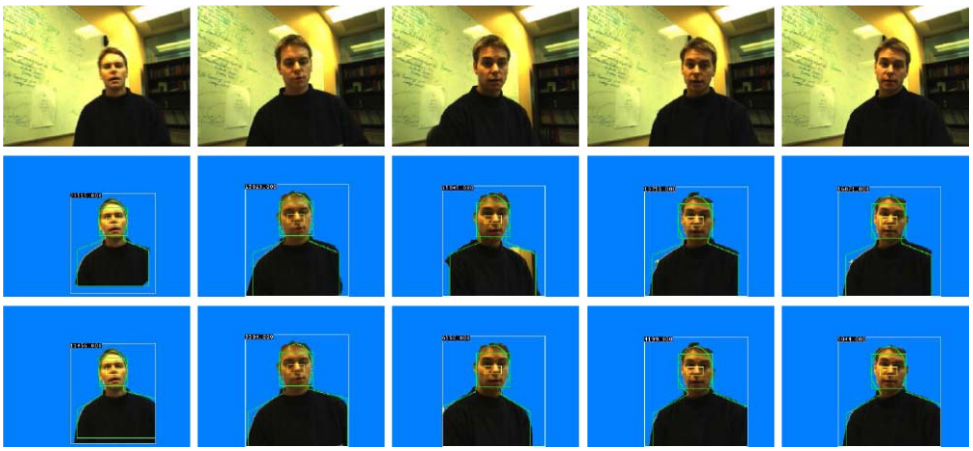

Fig. 9 Segmentation results using the 2D upper body model. The *first row* shows some frames from the video sequence. The *second row* shows the initial values of the pose parameters of the model and the resulting segmentations. The *last row* shows the final pose estimate and segmentation obtained using our method

Our experiments on segmenting humans using the 2D upper body model (Fig. 3) also produced good results. For these experiments, video sequences from the Microsoft Research bilayer video segmentation data-set (Kolmogorov et al. 2005) were used. The results of our method are shown in Fig. 9.

4.2 Segmentation and Pose Estimation

Figures 11 and 12 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 an Intel Pentium 2 GHz machine, when dealing with 644×484 images, is about 50 seconds per view.[5] As shown in these figures, the pose estimates match the original images accurately. In Figs. 11 and 12, it should be noted that the appearance models of the foreground and background are quite similar: for instance, in Fig. 12, 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 Fig. 11 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]However, this could be speed up by computing the parameters of the CRF in an FPGA (Field-programmable gate array).

5 Discussion

Localizing the object in the image and inferring its pose is a computationally expensive task. Once a rough estimate of the object pose is obtained, the segmentation can be computed extremely efficiently using graph cuts (Bray et al. 2006). In our work on real time face detection and segmentation (Rihan et al. 2006), we showed how an off the shelf face-detector such as the one described in Viola and Jones (2004) can be coupled with a CRF to get accurate segmentation and improved face detection results in real time.

The object (face) localization estimate (obtained from any generic face detector) was incorporated in a discriminative CRF framework to obtain robust and accurate face segmentation results as shown in Fig. 13. The energy $E(\mathbf{x}^*)$ of any segmentation solution $\mathbf{x}^*$ is the negative log of the probability, and can be viewed as a measure of how uncertain that solution is. The higher the energy of a segmentation, the lower the probability that it is a good segmentation. Intuitively, if the face detection is correct, the resulting segmentation obtained from our method should have high probability and hence have low energy compared to that of false detections. This characteristic of the energy of the segmentation solution can be used to prune out false face detections

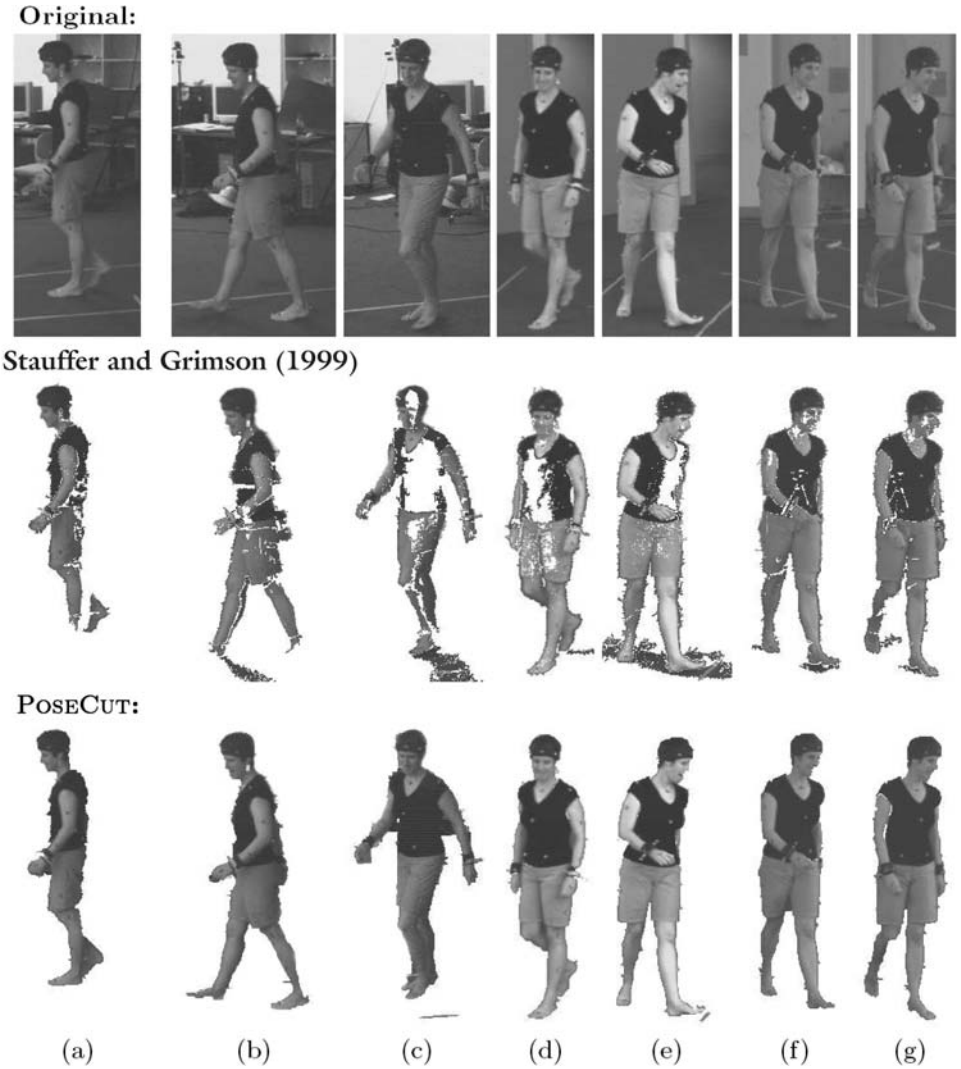

Fig. 10 Segmentation results obtained by Stauffer and Grimson (1999) and PoseCut

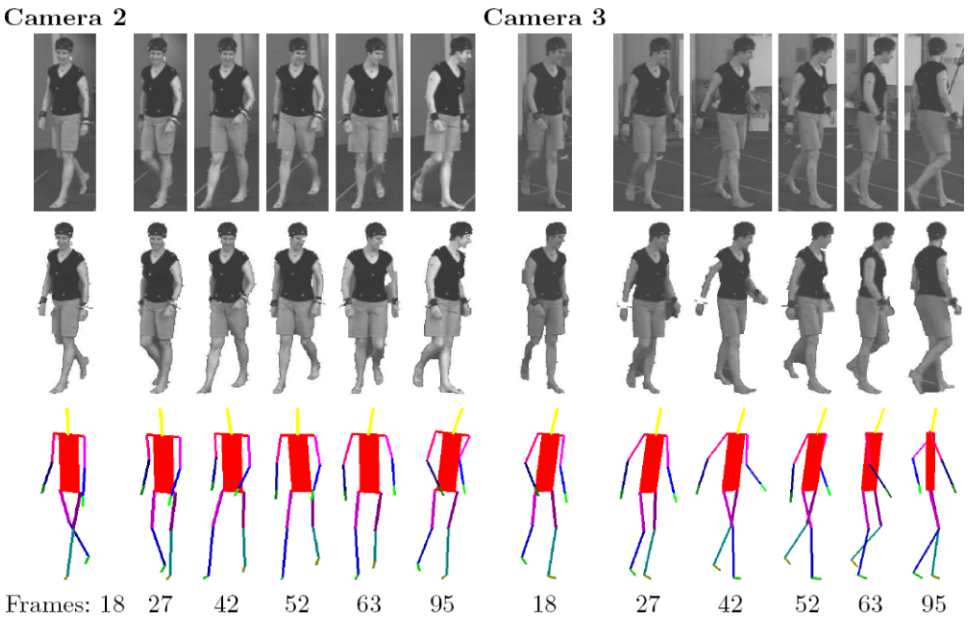

Fig. 11 Segmentation (*middle*) and pose estimation (*bottom*) results from PoseCut

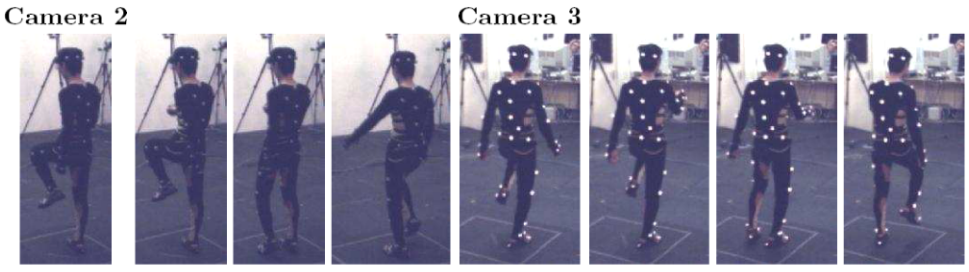

Fig. 12 Segmentation (*middle row*) and pose estimation (*bottom row*) results obtained by POSECUT. Observe that although the foreground and background appearances are similar, our algorithm is able to obtain good segmentations

Fig. 13 Real Time Face Segmentation using a face detections. The *first image* on the first row shows the original image. The *second image* shows the face detection results. The image on the second row shows the segmentation obtained by using shape priors generated using the detection and localization results

thus improving the face detection accuracy. The procedure is illustrated in Fig. 14. A similar strategy was recently used in Ramanan (2007).

6 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 ability 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. This method was recently also used for the problem of reconstructing objects from multiple views (Sun et al. 2006).

Searching over Pose Manifolds It is well known that the set of all human poses constitutes a low-dimensional manifold in the complete pose space (Ek et al. 2007; Urtasun et al. 2005; Sidenbladh et al. 2000a, 2000b). Most work in exploiting this fact for human pose inference has been limited to finding linear manifolds in pose spaces. The last few years have seen the emergence of non-linear dimensionality reduction techniques for solving the pose inference problem (Sminchisescu and Jepson 2004). Recently, Urtasun et al. (2005) showed how Scaled Gaussian Process Latent Variable Models (SGPLVM) can be used to learn prior models of human pose for 3D people tracking. They showed impressive pose inference results using monocular data. 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 algo-

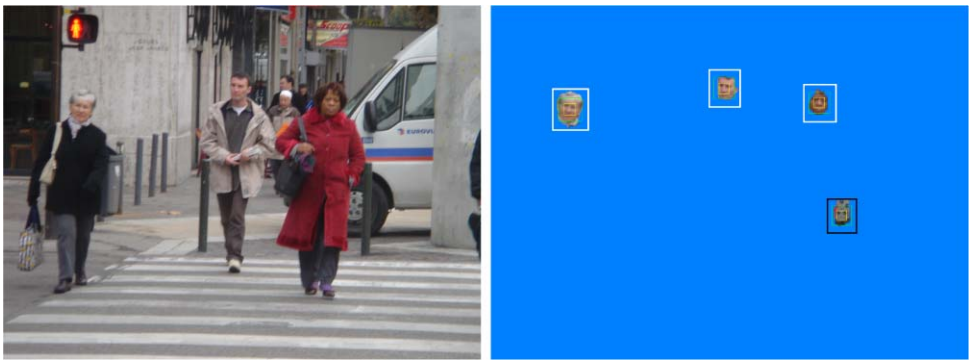

Fig. 14 The figure shows an image from the INRIA pedestrian data set. After running our algorithm, we obtain four face segmentations, one of which (the one bounded by a black square) is a false detection. The energy-per-pixel values obtained for the true detections were 74, 82 and 83 while that for the false detection was 87. As you can see the energy of false detection is significantly higher than that of the true detections, and can be used to detect and remove it

rithm. 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.

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

References

Agarwal, A., & Triggs, B. (2004). 3D human pose from silhouettes by relevance vector regression. In: *CVPR* (Vol. II, pp. 882–888).

Agarwal, A., & Triggs, B. (2006). Recovering 3D human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell., 28*.

Blake, A., Rother, C., Brown, M., Pérez, P., & Torr, P. (2004). Interactive image segmentation using an adaptive gmmrf model. In: *ECCV* (Vol. I, pp. 428–441).

Boykov, Y., & Jolly, M. (2001). Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: *ICCV* (Vol. I, pp. 105–112).

Bray, M., Kohli, P., & Torr, P. H. S. (2006). Posecut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In: *ECCV* (Vol. 2, pp. 642–655).

Cremers, D., Osher, S., & Soatto, S. (2006). Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *International Journal of Computer Vision*, *69*, 335–351.

Deutscher, J., Davison, A., & Reid, I. (2001). Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In: *CVPR* (Vol. 2, pp. 669–676).

Ek, C., Laurence, N., & Torr, P. (2007). Gaussian process latent variable models for human pose estimation. In *4th joint workshop on multimodal interaction and related machine learning algorithms*.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2000). Efficient matching of pictorial structures. In: *CVPR*.

Felzenszwalb, P., & Huttenlocher, D. (2004). *Distance transforms of sampled functions* (Technical Report TR2004-1963). Cornell University.

Freedman, D., & Zhang, T. (2005). Interactive graph cut based segmentation with shape priors. In: *CVPR* (Vol. I, pp. 755–762).

Gavrila, D., & Davis, L. (1996). 3D model-based tracking of humans in action: a multi-view approach. In: *CVPR* (pp. 73–80).

Huang, R., Pavlovic, V., & Metaxas, D. (2004). A graphical model framework for coupling mrfs and deformable models. In: *CVPR* (Vol. II, pp. 739–746).

Kehl, R., Bray, M., & Van Gool, L. (2005). Full body tracking from multiple views using stochastic sampling. In: *CVPR* (Vol. II, pp. 129–136).

Kohli, P., & Torr, P. (2005). Efficiently solving dynamic Markov random fields using graph cuts. In: *ICCV*.

Kolmogorov, V., & Zabih, R. (2002). What energy functions can be minimized via graph cuts? In: *ECCV* (Vol. III).

Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., & Rother, C. (2005). Bi-layer segmentation of binocular stereo video. In: *CVPR* (Vol. 2, pp. 407–414).

Kumar, M., Torr, P., & Zisserman, A. (2005). Obj cut. In: *CVPR* (Vol. I, pp. 18–25).

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *ICML* (pp. 282–289).

Lan, X., & Huttenlocher, D. P. (2005). Beyond trees: common-factor models for 2D human pose recovery. In: *ICCV* (pp. 470–477).

Leventon, M. E., Grimson, W. E. L., & Faugeras, O. D. (2000). Statistical shape influence in geodesic active contours. In: *CVPR* (pp. 1316–1323).

Mori, G., Ren, X., Efros, A. A., & Malik, J. (2004). Recovering human body configurations: Combining segmentation and recognition. In: *CVPR* (Vol. 2, pp. 326–333).

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). *Numerical recipes in C*. Cambridge: Cambridge University Press.

Ramanan, D. (2007). Using segmentation to verify object hypotheses. In: *CVPR*.

Ramanan, D., & Forsyth, D. A. (2003). Finding and tracking people from the bottom up. In: *CVPR* (Vol. 2, pp. 467–474).

Rihan, J., Kohli, P., & Torr, P. H. S. (2006). Objcut for face detection. In: *ICVGIP* (pp. 576–584).

Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. In: *ICCV* (pp. 750–757).

Sidenbladh, H., Black, M. J., & Fleet, D. J. (2000a). Stochastic tracking of 3D human figures using 2D image motion. In: *ECCV* (Vol. 2, pp. 702–718).

Sidenbladh, H., Black, M. J., & Fleet, D. J. (2000b). Stochastic tracking of 3D human figures using 2D image motion. In: *ECCV* (pp. 702–718).

Sminchisescu, C., & Jepson, A. D. (2004). Generative modeling for continuous non-linearly embedded visual inference. In: *ICML*.

Sminchisescu, C., & Triggs, B. (2001). Covariance scaled sampling for monocular 3D body tracking. In: *CVPR* (pp. 447–454).
Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In: *CVPR* (pp. 246–252).
Stenger, B., Thayananthan, A., Torr, P., & Cipolla, R. (2003). Filtering using a tree-based estimator. In: *ICCV* (pp. 1063–1070).
Sun, Y., Kohli, P., Bray, M., & Torr, P. H. S. (2006). Using strong shape priors for stereo. In: *ICVGIP* (pp. 882–893).
Urtasun, R., Fleet, D. J., Hertzmann, A., & Fua, P. (2005). Priors for people tracking from small training sets. In: *ICCV* (pp. 403–410).
Viola, P. A., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*, 137–154.
Zhao, L., & Davis, L. S. (2005). Closely coupled object detection and segmentation. In: *ICCV* (pp. 454–461).