Stochastic search for human upper body pose and appearance

B. Bascle
Orange Labs / France Telecom R&D
benedicte.bascle@orange-ftgroup.com

Résumé

This article presents an algorithm for automatic inference of human upper body pose and appearance from a monocular video. The approach consists in looking at pairs of randomly selected (and distant) images from the sequence. For each image pair $(I_t, I_{t'})$, a stochastic search is made to find pairs of likely upper body limb positions. This likelihood takes into account image-based probabilities (given by background subtraction and edge maps). It also measures the similarity of limb appearance (texture and color) between the two images. Because noise and outliers in distant images are unlikely to be coherent, false positives given by limb detection in each image do not usually have similar appearance to false positives in the other image in the pair. Thus they are discarded and only real positives have good likelihood. These give both upper body pose and limbs appearances models. A strength of the approach is that it can be applied to complex sequences where people appearance is unknown and where motion is at times too fast or unexpected for tracking approaches to perform well. It can be applied to people with different builds and types of clothing. Because it does not rely on pre-learning possible models of appearance (and associated edge maps for instance), it is less likely to fail when atypical appearances are seen.

Mots clefs

human motion capture, gesture analysis, stochastic search, MCMC, online learning of appearance model.

1 Introduction

Articulated pose estimation [1] is a problem of great interest to the computer vision and image processing communities, with a lot of interesting applications (surveillance, interaction, indexing, etc...). It is a difficult problem, because the state space is high-dimensional and the observation model is highly non linear. In addition, people can move fast and unpredictably, they can appear in a variety of poses and clothes, and are often surrounded by clutter. The problem is even more difficult if there is only one camera. To answer these challenges, a number of approaches to articulated pose estimation have been proposed in the literature [2] [3] [4] [5] [6].

For monocular video sequences, many works rely on manual initialization or learnt a prioris to track and detect articulated bodies [1]. However manual initialization (or at least an approximate initialization on the first image such as used by [6]) is not always available. Learnt a prioris (such as transition probabilities [7] and possible appearances and related statistics [5] [4] [8]) are dependable only if the database used for training is representative and exhaustive. Typically, types of motions and appearances not observed in the training data are difficult to deal with later on. In addition, motion models often break down for large unpredictable motions (such as can happen with hand motions for instance). To avoid this problem, we would like to make as few hypotheses about body appearance and motion as possible.

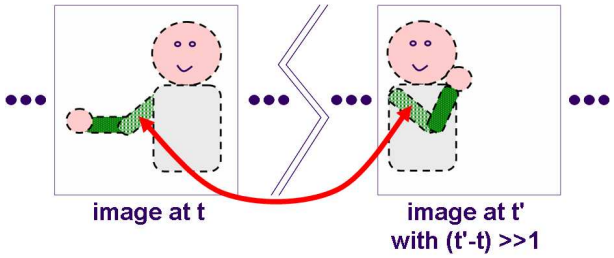

Figure 1 – **Motivation of the approach.** *Detection gives limb hypotheses in each image. However there are false positives due to noise, outliers, clutter, etc... To discard false positives, our approach looks at limb hypotheses from distant images. Real positives should have similar appearance (texture, color) between distant images, while noise, outliers and false positives are unlikely to be coherent.*

In fact, similarly to [9], we would like to build a model of appearance online from the images. [9] do this by either a bottom-up approach (detecting rectangular body parts and grouping them), or a top-down approach (detecting pre-determined typical poses). Detecting rectangular body parts is a hard problem if clothing is highly textured and the background is cluttered. That is why we adopt a top-down generative approach where body pose hypotheses are generated by MCMC-based sampling and then checked against image data. Likely poses give body appearance. We also adopt an opportunistic approach by detecting only certain poses, but these poses are not chosen a priori (like the mid-stance pose of lateral-walking of [9]). These poses are determined automatically by our algorithm, as they are the most likely and easiest to detect given image data. Another difference is that we do not look at single images. This

is because, when looking for good body hypotheses, and using no learnt a prioris about possible appearances and motions, looking at one frame only is not enough to solve the problem. And looking at successive images (e.g. tracking) does not give a lot of information if the positions are very similar. In addition, tracking can fail and thus information from one image to another can be lost. In this paper, we confront pairs of hypotheses for body poses and appearance from pairs of distant (and randomly picked) images in the sequence. A criterion of similarity of limb appearance (texture, color) between image frames is used to determine the most probable poses and appearance.

Section 2 describes the outline of the approach. Sections 3 and 4 describe important details of the algorithm. And section 5 shows results and applications.

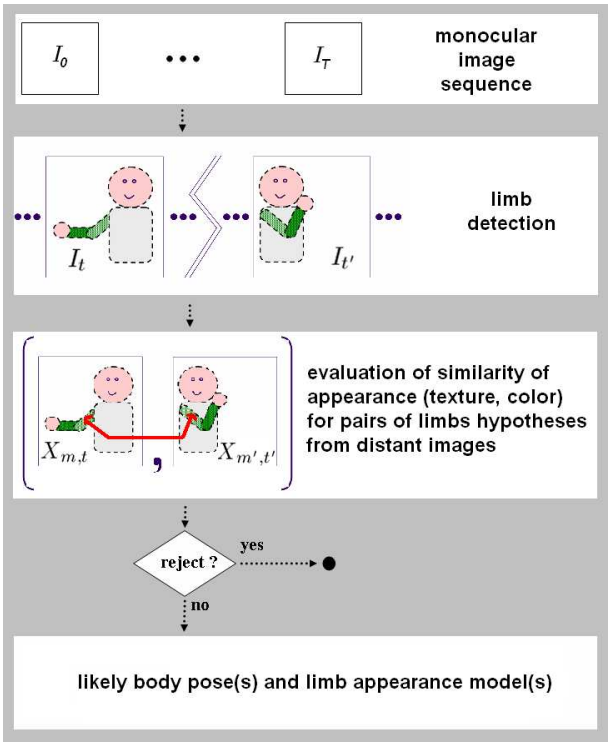

Figure 2 – **Principle of the stochastic search of solution space.** *The acceptance-rejection mechanism is that of MCMC. Multiple Markov chains are used in parallel.*

2 Principle of the approach

Let $(I_0, ..., I_T)$ be a monocular image sequence. Our approach to detecting body pose and modeling appearance is described by fig. 2 and is as follows.

As explained in the introduction, in the search for body pose and appearance, this article proposes to confront pairs of hypotheses for body poses and appearance from randomly picked pairs of images in the sequence. States $\mathbf{X}$ in solution space are pairs of pose hypotheses from two different images, e.g. $\mathbf{X} = (X_t, X_{t'})$ with X_t a body pose sample for image I_t. Note that for clarity's sake, we omit the sample index for $\mathbf{X}$ and X.

Given that a search of state space becomes exponentially more time-consuming as the number of state parameters increase, it is advisable to parameterize body X_t pose as compactly as possible. Because of this, the search and sampling in this paper are performed in 2D pose space, and the corresponding 3D poses are estimated as a post-processing step, as done in [10]. Thus X_t is a 2D body pose. In addition, when looking for upper body pose, both arms are sampled separately, since the search space for each arm will be much smaller than their combined search space. That is why X_t parameterises the 2D pose of one arm in image t. The state space $\mathbf{X} = (X_t, X_{t'})$ then represents the 2D poses of an arm in a pair of images I_t and $I_{t'}$. The approach is applied independently to both arms, for better processing speed.

Our stochastic search in solution space $\Omega_{\mathbf{X}}$ is based on a variant of Metropolis-Hastings Markov Chain Monte Carlo (MH-MCMC) called the Metropolized independent sampler (MIS). The general principle of MH-MCMC (and thus of MIS) is to perform a biased random walk in state space. The walk is guided by a Markov chain proposing the next step in state space, combined with an acceptance-rejection mechanism. The choice of the proposal distribution $\mathbf{Q}(\mathbf{X}', \mathbf{X})$ that proposes the next state $\mathbf{X}'$ in the walk is of course crucial to the success of the algorithm. In the general case of MH-MCMC, the proposal distribution proposes a new state $\mathbf{X}'$ dependent on the previous state $\mathbf{X}$ in the Markov Chain. $\mathbf{X}'$ is often a variation of $\mathbf{X}$, chosen randomly or inspired by a dynamics model. In the special case of MIS (the variant of MH-MCMC used in this paper), the proposal distribution is independent of the previous position, e.g. $\mathbf{Q}(\mathbf{X}', \mathbf{X}) = \mathbf{Q}(\mathbf{X}')$. It makes the sampler free to explore search space far from the previous sample. It also makes the sampler free of hypotheses about body dynamics, motion models and transition probabilities, which can be ad hoc, restrictive, dependent on a learning set, and/or not suitable for large motions.

A straightforward implementation of the MIS in our case would be to choose a set of random image pairs. For each image pair (I_t, I'_t), the algorithm would iterate to search the joint state space $\Omega_{\mathbf{X}}^{t,t'}$. Each step of the iteration would consist in : 1) proposing the next state sample $\mathbf{X}' = (X'_t, X'_{t'})$ using a proposal process $\mathbf{Q}(\mathbf{X}')$ 2) accepting or rejecting $\mathbf{X}'$ as the next state depending on the value of $\alpha(\mathbf{X}', \mathbf{X})$).

However, this implementation has been modified for two reasons. Firstly, since the state $\mathbf{X} = (X_t, X_{t'})$ is a combination of two 2D arm configurations X_t and $X_{t'}$, the proposal kernel $\mathbf{Q}(\mathbf{X}')$ is also set as a combination of (independent) proposal chains with $\mathbf{Q}(\mathbf{X}') = (Q(X'_t), Q(X'_{t'}))$ where $Q(X'_t)$ is the proposal distribution for a 2D arm configuration at time t'. Secondly, as shown by [11] with their image-derived "proposal maps", using a data-driven proposal mechanism is more efficient in finding promising modes, and we follow this idea by

making the single-image proposal distribution $Q(X'_t)$ dependent on the image information I_t and on associated measurements. Given these two considerations, and the fact that we do not want to sample body hypotheses on the same image twice (if that image belongs to two image pairs), we have made the single-image proposal process $Q(X'_t)$ a pre-processing step. This gives a modified implementation of the stochastic search algorithm (see fig. 2). First a two-stage pre-processing step is performed. It starts by choosing a random subset of images from the original image sequence. Then, for each image in this subset, it generates a set of M single-frame body hypotheses $(X'_{m,t})_{m=1..M}$ using the proposal process $Q(X'_t)$. After this pre-processing, the stochastic search of two-frame state space starts. Random image pairs are drawn from the already selected image subset. For each image pair (I_t, I'_t), the algorithm iterates its search step in state space. Each step consists in : 1) randomly picking a pair $(X'_{m,t}, X'_{m',t'})$ of arm samples from the pre-calculated sample sets for I_t and $I_{t'}$ 2) accepting or rejecting $(X'_{m,t}, X'_{m',t'})$ as the next state depending on the value of the acceptance ratio $\alpha((X'_{m,t}, X'_{m',t'}), (X_{m'',t}, X_{m''',t'}))$ where $(X_{m'',t}, X_{m''',t'})$ is the previously accepted sample. The detailed image-based proposal process is described in the following section. Then section 4 shows how the acceptance ratio is estimated and how it enforces the constraint that likely pairs of body pose hypotheses from two image frames show have similar appearance.

3 Single-frame stratified proposal process

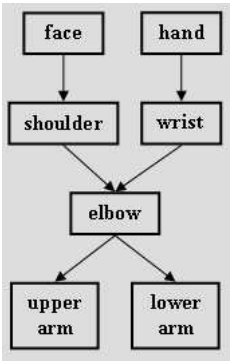

Figure 3 – **Graphical model used for sampling / detecting arm hypotheses on an image I_t. Both arms are detected independently to reduce the size of the state space.**

The single-frame proposal process generates 2D arm samples $X'_{m,t}$ from an image I_t by following the graphical model shown in figure 3. This graphical model does not follow the kinematic chain of the arm, as done by most authors. It is based on the idea that some arm components are easiest to recover and have the least degrees of freedom and so should be detected first. The face is detected first, because it is the easiest and most reliable part of the body to detect. Based on the face detection, likely torso positions are extracted using a combination of image clues (background subtraction and edge information). From these torso positions, possible locations of the shoulder joint in the image (with the associated uncertainties) are inferred. These are used to generate shoulder samples (x_s, y_s) in likely areas. In parallel to this, hand candidates are detected as fast-moving skin-toned blobs. These hands blobs should also be of a size compatible with that of the face in the image, given anthropomorphic relative limb sizes. From these hand hypotheses, approximate wrist positions (x_w, y_w) can be estimated. Once samples of possible shoulder and wrist positions have been gathered, elbow samples (x_e, y_e) are drawn from a search region based on anatomic constraints. These constraints correspond in 2D to the fact that in 3D, given elbow and hand positions, the elbow lies on a circle. At the end of the proposal process, multiple 2D arm hypotheses $X_t = \left(x_s\ x_e\ x_w\ y_s\ y_e\ y_w\right)^T$ are available. The upper and lower arm are modeled as 2D rectangles between the joints. This shape model is as good (or as bad) as any, given the shape variation of 2D limb contours with viewpoint, muscle activity and clothing movement. As can be seen from the sampling process, the rectangles' apparent lengths in the image is variable. The approximate range of variation of their apparent width is predicted given face size in the image and the effects of foreshortening.

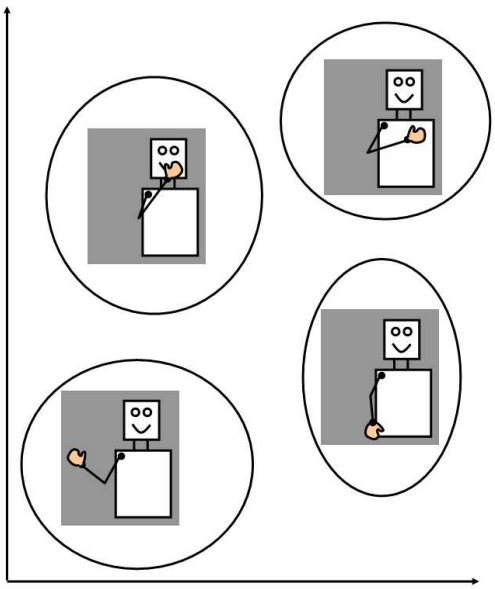

Figure 4 – **Stratified sampling in 2D arm space.**

The proposal process we just described has one drawback. If classical sampling methods (such as Poisson, jittered, etc...) are used at each step, clumps of samples might occur and the search space might be badly covered, unless a huge number of samples is used. To avoid clumps, and to optimise the coverage of search space given a maxi-

mum number of samples, two strategies are used. Firstly, the search space is stratified. This means that the search space is split into several subspaces and that samples are drawn from each subspace separately. This ensures that none of the subspaces will be ignored by the sampling process. Here we have defined subspaces corresponding to relative hand positions of the hand and the rest of the body (see fig. 4). Hence positions where the hand touches the face or is in front of the torso are systematically investigated. The second strategy used to improve the proposal process is to sample joint positions using "blue" sampling, also known as Poisson disc sampling. Poisson disk sampling ensures minimum distance spacing between samples. This gives good search space coverage and avoids clumps. At the end of the improved proposal process, a number of arm hypotheses are available for a subset of the original images.

4 Joint state space acceptance ratio

As explained in section 2, arm hypotheses obtained independently from different images by the proposal process (described in section 3) need to be confronted to find arm hypotheses whose appearance (texture, color) is consistent with time. This is done by randomly drawing image pairs $(I_t, I_{t'})$. For each image pair, the posterior distribution of arm pose pairs is built by a random walk. Each step of the walk draws an arm hypothesis for each image and pairs them. The pair of hypotheses $(X'_{m,t}, X'_{m',t'})$ is selected to be part of the posterior distribution depending on a Metropolis-style acceptance ratio α. α depends on the previously accepted sample $(X_{m'',t}, X_{m''',t'})$ and is a function of the relative posterior probabilities of the current sample and the previously accepted sample. These probabilities combine the image-based likelihoods $p(I_t|X'_{m,t})$ and $p(I_{t'}|X'_{m',t'})$ of the two arm hypotheses in the two images, and a criterion $\psi(X'_{m,t}, X'_{m',t'}, I_t, I_{t'})$ estimating the similarity (color, texture) of arm appearance between the two hypotheses and images. This gives :

$$\alpha = \frac{p(I_t|X'_{m,t}) * p(I_{t'}|X'_{m',t'}) * \psi(X'_{m,t}, X'_{m',t'}; I_t, I_{t'})}{p(I_t|X_{m'',t}) * p(I_{t'}|X_{m''',t'}) * \psi(X_{m'',t}, X_{m''',t'}; I_t, I_{t'})}$$

Each image-based arm likelihood (for instance $p(I_t|X'_{m,t})$) combines the likelihoods of the upper and lower arm. The limb likelihood (for either the lower or upper arm) is the product of evidence from background subtraction and edges. This gives $p(I_t|X'_{m,t}) = p_{\text{bckg}}^{\text{upper arm}}\ p_{\text{edg}}^{\text{upper arm}}\ p_{\text{bckg}}^{\text{lower arm}}\ p_{\text{edg}}^{\text{lower arm}}$. The local limb evidence (as an example $p_{\text{edg}}^{\text{upper arm}}$) is given by the sum S of an edge measure over all the pixels of the contour of the limb, raised to the power γ. γ measures the capacity of edge information to discriminate between the contour of the limb and the rest of the pixels in the neighborhood. It is given by $\gamma = \frac{\overline{S}+\epsilon}{S+1}$, with $\overline{S}$ the sum of the edge measure over all the pixels of the neighborhood not on the contour of the limb. Similar calculations are made for all limbs and information from background subtraction (in this case, S is summed over the interior region, and $\overline{S}$ over the exterior). The criterion $\psi(X'_{m,t}, X'_{m',t'}, I_t, I_{t'})$ estimating the similarity of arm appearance between the two hypotheses combines similarity information from color and gradient for both limbs. In each case, the similarity is estimated by the (Battacharyah) distance between histograms (of color and gradient orientation respectively). As above, the distances are raised to powers corresponding to the capacity of histogram information (respectively from color or gradient) to discriminate between the interior and exterior region of a limb. For reasons of space, the formulae for ψ are not given here, but should be straightforward to extrapolate from above.

5 Results

Results (see fig. 5) show that the approach performs well on sequences of images showing different people, with different clothing (textured or not, baggy or not), different (unknown a priori) body sizes, different backgrounds and lighting conditions, different types of motion, and complex poses. For instance, the last 2 rows of fig. 5 show correct detection for baggy clothing, in very noisy images (due to low light conditions, the images being taken at night).

6 Conclusion

Our approach performs probabilistic inference of body pose and appearance by looking at body hypotheses in randomly sampled image pairs in a video sequence, and finding those with similar body appearances (texture, color) in both images. No model or constraint on limb motion is used. Hypothesis sampling is done using a MCMC mechanism with a (quasi-)independent proposal mechanism designed to maximise state space coverage.
The approach can be used on sequences where tracking fails at times (when motion models not do apply for instance, or are fooled by unexpected motions). It does not rely on a database of possible appearances (or edge maps) which are by necessity not completely exhaustive and can fail on unusual appearances. Our method takes advantage of the fact that body pose detection is easier in some images than others (due to body configuration, self-occlusion, clutter, illumination, etc...). It can deal with a variety of body sizes, appearances and clothing (cf bagginess, texture).

Références

[1] Moeslund, T.B., Hilton, A., Kruger, V. : A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2) (2006) 90–126

[2] Deutscher, J., Davison, A., Reid, I. : Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. IEEE Int. Conference on Computer Vision and Pattern Recognition (CVPR'01) (2001)

[3] MacCormick, J., Isard, M. : Partitioned sampling, articulated objects, and interface-quality hand tracking. In : ECCV. (2000)

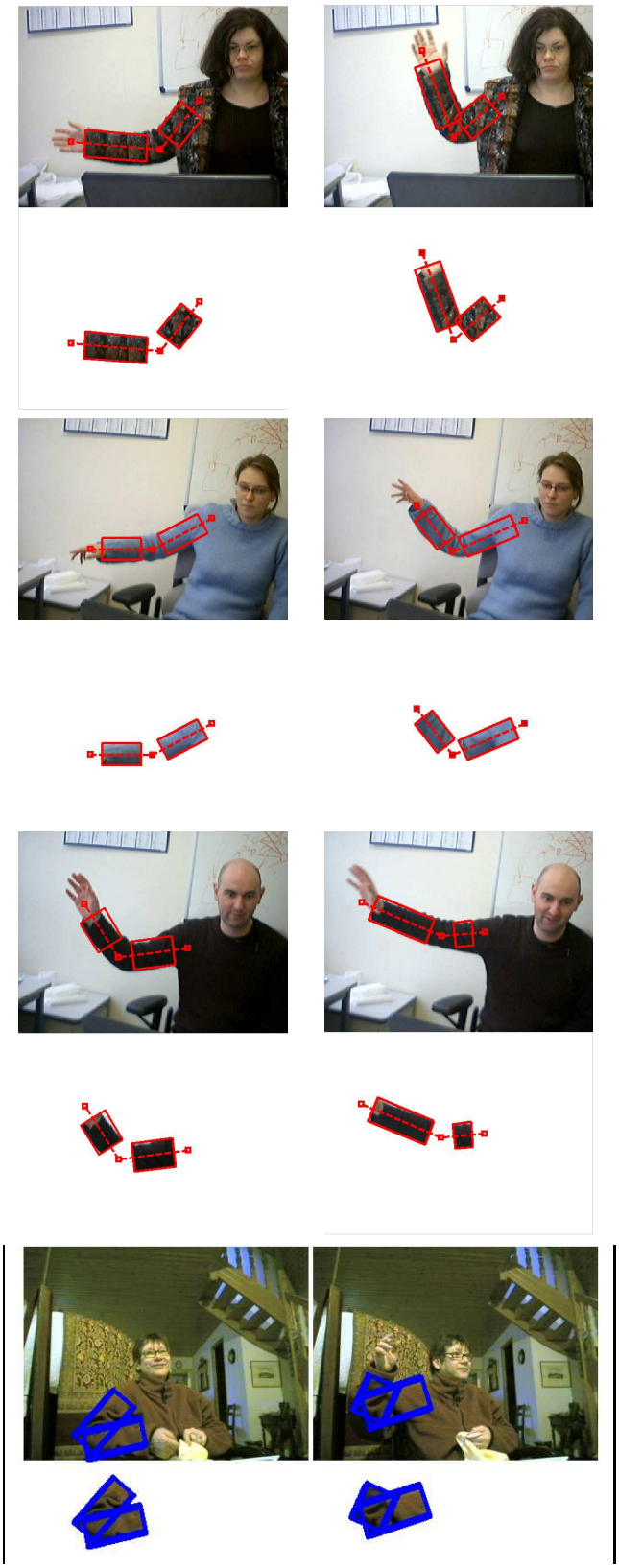

Figure 5 – **Most likely poses and appearances from the posterior distribution estimated by Metropolised independence sampling on pairs of images, and refined locally by a deformable model. For each example, the top row shows a pair of images of the same subject. The bottom row shows the most likely corresponding arm poses and appearances.**

[4] Taycher, L., Demirdjian, D., Darrell, T., Shakhnarovich, G. : Conditional random people : Tracking humans with crfs and grid fi lters. In : CVPR. (2006)

[5] Agarwal, A., Triggs, B. : Recovering 3d human pose from monocular images. IEEE Transactions on Pattern Analysis & Machine Intelligence **28**(1) (jan 2006)

[6] Bray, M., Kohli, P., Torr, P.H.S. : Posecut : Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In : ECCV (2). (2006) 642–655

[7] Sigal, L., Sidharth, B., Roth, S., Black, M., Isard, M. : Tracking loose-limbed people. In : CVPR. (2004)

[8] Agarwal, A., Triggs, B. : A local basis representation for estimating human pose from cluttered images. In : Asian Conference on Computer Vision. (2006)

[9] Ramanan, D., Forsyth, D.A., Zisserman, A. : Strike a pose : Tracking people by fi nding stylized poses. Technical Report UCB/CSD-04-1362, EECS Department, University of California, Berkeley (2004)

[10] Taylor, C.J. : Reconstruction of articulated objects from point correspondences in a single uncalibrated image. Computer Vision and Image Understanding : CVIU **80**(3) (2000) 349–363

[11] Lee, M.W., Cohen, I. : Proposal maps driven mcmc for estimating human body pose in static images. In : IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04). (2004)

[12] Liu, J. : Metropolized independent sampling with comparisons to rejection sampling and importance sampling. In : Statist. Comput. 6, 113–119. (1996)

[13] Gao, J., Shi, J. : Multiple frame motion inference using belief propagation. In : 6th IEEE Conference on Automatic Face and Gesture Recognition. (2004)

[14] Urtasun, R., Fleet, D., Fua, P. : 3d people tracking with gaussian process dynamical models. (2006) I : 238–245

[15] Ju, S.X., Black, M.J., Yacoob, Y. : Cardboard people : A parameterized model of articulated motion. In : International Conference on Automatic Face and Gesture Recognition. (1996) 38–44

[16] Khan, Z., Balch, T., Dellaert, F. : Mcmc-based particle fi ltering for tracking a variable number of interacting targets. IEEE Trans. Pattern Anal. Mach. Intell. **27**(11) (2005) 1805–1918

[17] Fox, D., Thrun, S., Burgard, W., Dellaert, F. : Particle fi lters for mobile robot localization. In Doucet, A., de Freitas, N., Gordon, N., eds. : Sequential Monte Carlo Methods in Practice, Springer (2001)

[18] Sidenbladh, H., Black, M.J., Fleet, D.J. : Stochastic tracking of 3d human fi gures using 2d image motion. In : ECCV (2). (2000) 702–718

[19] Balan, A.O., Black, M.J. : An adaptive appearance model approach for model-based articulated object tracking. In : CVPR '06 : Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2006) 758–765