Multicues 3D Monocular Upper Body Tracking Using Constrained Belief Propagation

Philippe Noriega, Olivier Bernier
France Telecom Research and Development
France Telecom R&D
2 Av. Pierre Marzin 22307 Lannion Cedex France
{philippe.noriega, olivier.bernier}@orange-ftgroup.com

Abstract

This paper describes a method for articulated 3D upper body tracking in monocular scenes using a graphical model to represent an articulated body structure. Belief propagation on factor graphs is used to compute the marginal probabilities of limbs. The body model is a loose-limbed model including attraction factors between adjacent limbs and constraints to reject poses resulting in collisions. To solve ambiguities resulting from monocular view, robust contour and colour based cues are extracted from the images. Moreover, a set of constraints on the model articulations is implemented according to human pose capabilities. Quantitative and qualitative results illustrate the efficiency of the proposed algorithm.

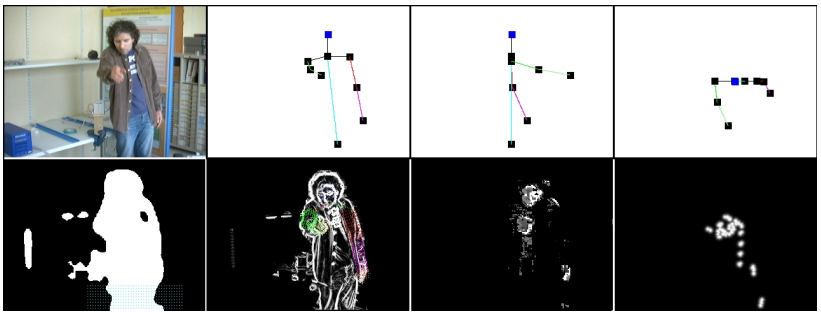

Figure 1: Upper body tracking. First row: original image, front, right side and top views of the obtained limbs positions with a single camera. Second row: background subtraction, contours, face colour map and energy motion distance map.

1 Introduction

Algorithms for body tracking must cope with a high dimensional space in which the joint probability function is highly multimodal and sharp. In this context, deterministic

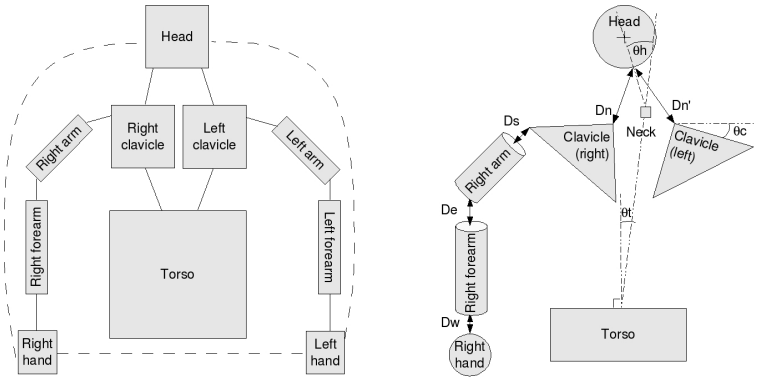

Figure 2: Limbs interactions (Left): nodes correspond to limbs, articulation constraints are represented by solid lines and dashed lines are additional non-collision constraints between head and hands. Upper body model (Right): arms and forearms are modeled by cylinders and the head by a sphere. Other limbs (hands torso and clavicles) are represented by 2D patches. Limb interaction factors are computed with the distances (Dn, Ds, De, Dw) between them. Other joints constraints are determined by the angles θh, θc and θt. The neck is located at equal distance from both clavicles.

methods can track in real time with stereo cameras [5], but may fail for monocular view because of the many local optimums owing to ambiguities in monocular scenes [13].

Due to articulation constraints, consistent poses are bounded in a smaller subspace making learning based tracking methods efficient if their learning set sufficiently covers this subspace. Various regressions methods, aiming at deducing a pose directly from an image, have been tested on walking sequences with constrained environments [2]. Non negative matrix factorisation [1] can enhance such methods by rejecting non discriminative data. Other methods like GPDM [15] introduce probabilities in the computation of a latent space to smooth the resulting pose, but test scenes are restricted to cyclic motions. Other methods that perform a comparison between an image and a learning base require a huge database even when robust locally-weighted regression between candidates poses is used [10]. Increasing the data base may slow down drastically the comparison process and, to speed up the selection of a subset of nearest neighbours, the comparison process can use locally sensitive hashing and Hamming distance [14]. The likelihood of a body pose is computed with this previous method using a Bayesian framework but some poses that are dissimilar to the learned ones are not correctly estimated and generally, the huge pose space and the variability in external parameters such as clothing or hairstyle is the major cause of failure in learning based methods.

Stochastic algorithms are useful in monocular vision to resolve ambiguities resulting from 2D to 3D pose inference, particularly when a multi-hypothesis algorithm, such as particle filtering [4], is used. The main drawback with such methods is the high dimensional pose space. A way to avoid this problem consists in using a loose-limbed body model [11] where the likelihood of each limb is evaluated independently. In this manner, a particle filter can be associated with each limb reducing the search space dimension to the number of *dof* of a limb [3]. Influence between limbs is taken into account by propagating limb beliefs through a factor graph using belief propagation [8]. A similar

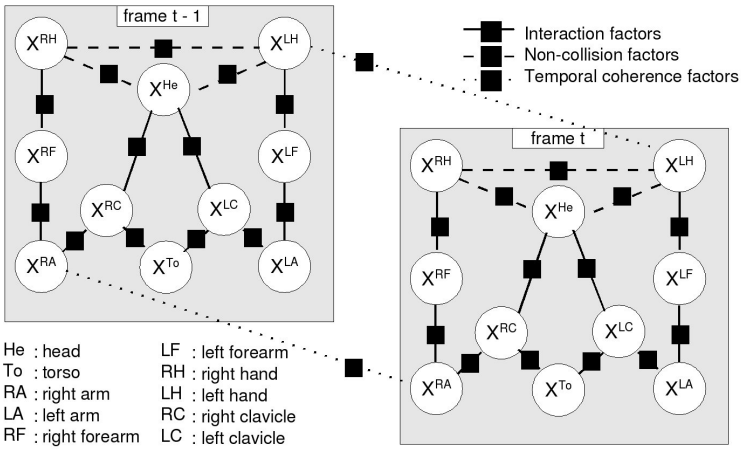

Figure 3: Factor graph. Circles corresponds to variable nodes (limb states) and black squares to factor nodes (temporal coherence T^{μ} and interaction or non-collision factors $\psi^{\mu\nu}$). For clarity, only two consecutive frames with two temporal factor links are shown and the factor nodes corresponding to the observations Y^{μ} are omitted.

technique is used in monocular scenes [7] with only motion energy as cue to detect arms and forearms position.

In this paper, the number of cues is increased to enhance the robustness of the tracking. Moreover, the use of interacting particle filters with belief propagation [3] simplify our algorithm by computing recursively an estimation in a discrete space instead of using, for example, a Gibbs sampler in a continuous one [11]. More general articulation constraints rules are built in the compatibility factors computation instead of learning them from specific walking sequences with a mixture of Gaussians [11]. The proposed algorithm performs at six *fps* using a standard webcam.

2 Recursive Bayesian tracking

The upper body is modeled as a graph including M limbs represented by nodes and links corresponding to articulations or non collision constraints between limbs (figure 2). Basically, a Markov network can be used to represent this structure but the non-collision constraints between the head and the hands generate a three nodes clique. A factor graph is constructed to simplify the model by using only pairwise factors [3]. The joint probability can be decomposed as a products of these factors. The complete graph includes the previous states to take into account the temporal coherence (figure 3). Given a limb μ, its state X_t^{μ} at time t and the image observations Y_t^{μ}, the model parameters are the observations compatibility factors $\phi^{\mu}(X^{\mu}, Y^{\mu})$, the time interaction factors $T^{\mu}(X_t^{\mu}, X_{t-1}^{\mu})$, and the interaction factor for the link between limbs μ and ν: $\psi^{\mu\nu}(X^{\mu}, X^{\nu})$. Adopting these notations, the joint probability knowing all observations from time 0 to T is:

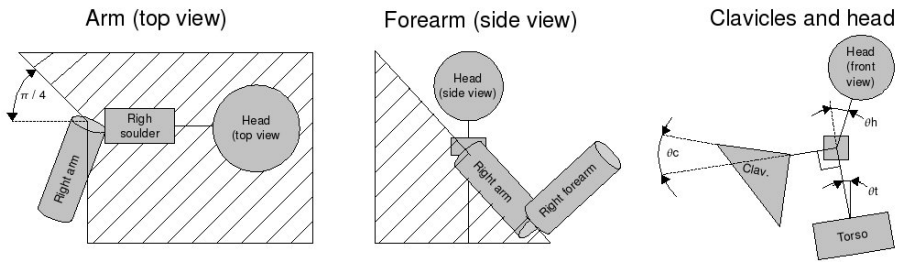

Figure 4: Articulations constraints. Arm and forearm: dashed lines show limb forbidden areas. The angular constraints are $|\theta_c| \leq 15^\circ$ for clavicles and $|\theta_h| \leq 25^\circ$ for head.

$$P(X_{0:T}|Y_{0:T}) = \prod_{t=0}^{T} \Phi(X_t, Y_t)\Psi(X_t) \prod_{t=1}^{T} T(X_t, X_{t-1}) \;, \tag{1}$$

with:

- $\Phi(X_t, Y_t) = \prod_{\mu=1}^{M} \phi^\mu(X_t^\mu, Y_t^\mu)$,
- $\Psi(X_t) = \prod_{(\mu,\nu)\in\Gamma} \psi^{\mu\nu}(X_t^\mu, X_t^\nu)$, where Γ is the set of links,
- $T(X_t, X_{t-1}) = \prod_{\mu=1}^{M} T^\mu(X_t^\mu, X_{t-1}^\mu)$.

The marginal probabilities of the limbs' state are obtained using the belief propagation algorithm on a factor graph [3]. As the graph includes cycles, the obtained marginal is an approximation of the true one. This approximation further depends on the messages update order. To simplify the algorithm, the messages are propagated to all nodes within the current frame for a fixed number of iterations (10 in our case) and then propagated only once from a frame to the following one. Therefore, the estimation of a marginal at any time t does not depend on the observations after time t, and the estimation of the marginals can be computed recursively.

The messages are represented by sets of weighted samples. From one frame to the next, they are calculated using a particle filter scheme consisting in a re-sampling step followed by a prediction step based on the time coherence factors [4]. The loopy belief propagation algorithm is then reduced, for the current frame, to a loopy propagation algorithm for discrete state spaces, the space state for each limb being restricted to its samples. Moreover the marginal probability is then simply represented as a weighted sum of the same samples. In this manner, a full recursive estimation is obtained. The algorithm is equivalent to a set of interacting particle filters, where the sample weights are re-evaluated at each frame through belief propagation to take into account the links between limbs. This algorithm is relatively fast because for a frame t, as opposed to [11], the image based compatibility factors $\phi^\mu(X_t^\mu, Y_t^\mu)$ have to be evaluated only once for each sample, and the link interaction factors only once for each pair of samples for all connected limbs.

3 Application to monocular upper body tracking

The model is applied to articulated upper-body tracking using monocular colour images from a webcam. Head and hands are tracked using image colour information and grey levels are used to compute cues: background subtraction, motion energy and orientation contour map (figure 1).

3.1 Initialisation

An accurate face detector [6] is used to detect the face in the colour image. Once detected, a starting pose corresponding to the arms along the body with the torso vertical and facing the camera is supposed. The tracker can easily recover the real pose as long as it is not too far from this hypothesis. The detected face is also used to initialise a face colour histogram.

3.2 Body model and link interaction factors

Figure 2 shows the body model. 3D limbs are represented by a sphere for the head and cylinders for arms and forearms. Hands, clavicles and torso are represented by 2D patches using respectively circles, triangles, and a rectangle. Limbs are discretized using a grid of regularly distributed points around them. A Gaussian of the distance between two link points is used to compute the link interaction factors (see figure 2 for distances Dn, Ds, De and Dw). This Gaussian is zero centred for the shoulder-arm and arm-forearm joints, and on a reference distance for the head-neck and forearm-hand joints. Other constraints are added giving zero factor for angles θh and θc above a fixed threshold (figure 4). Three additional links are defined, which simply give a zero probability to solutions where hands and head intersect (non collision constraints).

3.3 Time coherence factor

The time coherence factors $T^{\mu}(X_t^{\mu}, X_{t-1}^{\mu})$ are simple Gaussian, independent for each parameter, centred on the value in the previous frame. For hands, which can move fast and rapidly change speed, the time coherence factors is a mixture of two similar Gaussian, one centred on the previous parameter and the other centred on the prediction of the current parameter using previous hand speed. The standard deviation is chosen to be 10 cm for hands positions, and 5 cm for other limb positions. For angles, the standard deviation is set to $\pi/8$.

4 Image features

The image compatibility factors $\phi^{\mu}(X_t^{\mu}, Y_t^{\mu})$ are computed from scores S_f^{μ} representing the compatibility between a limb hypothesis μ and cue f extracted from the image. Contrary to stereo [3], monocular images needs more cues to reach a sufficient level of robustness. Thus, multicues image based compatibility terms are fused to provide an overall score: $S^{\mu} = \prod_f S_f^{\mu}$. To avoid taking into account background distractors, a robust background subtraction [9] is used.

Figure 5: Finding the torso. The bottom grid points (black pixels) representing the pelvis moves horizontally in order to maximise the correspondence between the points and the positive background subtraction pixels (white pixels). The maximum energy is reached when the grid is centred on the bottom positive background subtraction zone. The top of the torso is located at equal distance between the two clavicles.

4.1 Face and hands tracking

Considering the head position detected during initialisation step (§ 3.1), a colour model is provided by computing a normalised colour histogram of the head. The pixels p corresponding to the projected points belonging to the head or the hands are compared with this model by computing the colour score:

$$S_c^{\mu} = \sum_p H(p) \tag{2}$$

The function $H(p)$ returns the histogram bin value corresponding to the pixel p colour.

4.2 Torso tracking

The torso is hard to detect because of clothes deformations or occlusions produced when a person moves. Suposing that the pelvis is located at the bottom of images, its position can be found using a rectangular grid of weighted points p interacting with a background subtraction to slide on the bottom of the image (figure 5). The torso score is:

$$S^t = \sum_{p \in t} W(p) Bg(p) \tag{3}$$

Where $W(p)$ is the weight of p corresponding to the Gaussian distance between p and the grid center. $Bg(p)$ returns the probability that pixel p belongs to the foreground according to a background subtraction [9]. The upper torso point corresponds to the neck located at half distance of the two clavicles.

4.3 Arms, forearms and clavicles tracking

Arms tends to move rapidly and are subject to many partial occlusions. Thus, to reach a sufficient level of robustness, a fusion of a contour based cue and motion energy is implemented. An accurate contour based score can be estimated by not only considering the contours magnitude but also their orientations. Given $M(\|\overrightarrow{p}\|) = \frac{1}{\lambda}\|\overrightarrow{p}\| tanh(\frac{\lambda}{\|\overrightarrow{p}\|})$, a function that penalise low and high magnitude contour points $\|\overrightarrow{p}\|$ with λ a tuning parameter, a score S_{or}^{μ} for a limb hypothesis μ is computed by considering the Gaussian

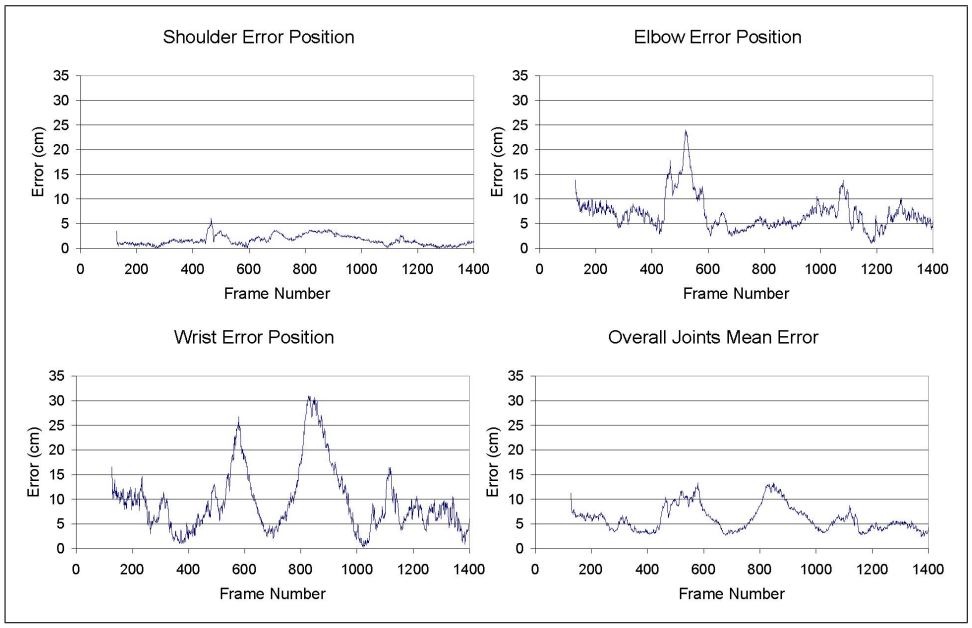

Figure 6: Quantitative results. For each joint, the error corresponds to the distance between estimated and true joint positions. As [12, 14], the mean error made on estimating the three joints is computed to provide the overall joint mean error.

difference $G_\theta(.)$ between the limb orientation θ_{limb} and each pixel contour orientation θ_p that corresponds to projected limb points p onto the image plane:

$$S_{or}^{\mu} = \sum_{p \in \mu} M(\|\overrightarrow{p}\|) G_\theta[\theta_{limb} - \theta_p] \tag{4}$$

The motion energy score is computed considering the Gaussian distance $G(d(p))$ between each projected limb point p and the nearest pixel where a motion has been detected: $S_m^{\mu} = 1 + \sum_{(p \in \mu)} G(d(p))$. Motion detection is provided by adjacent frame difference. This formula ensures that the motion score is at least 1 even if no motion is detected. Only the contour score is used for clavicles because they are strongly constrained by head position during belief propagation.

5 Experimental results

The system was tested on sequences grabbed with a standard webcam. Quantitative results were obtained comparing the estimated pose with a ground truth provided by a magnetic motion sensor. The true joint positions are measured for the right arm joints (shoulder, elbow and hand). The test sequence includes full 3D movements with limb occlusions and cluttered background (figure 9). Instead of only computing the overall limb mean error [12, 14], our results are complemented by the estimation error for each limb (figure 6). Qualitative results are shown on figure 7 where various user on different backgrounds and clothes are successfully tested.

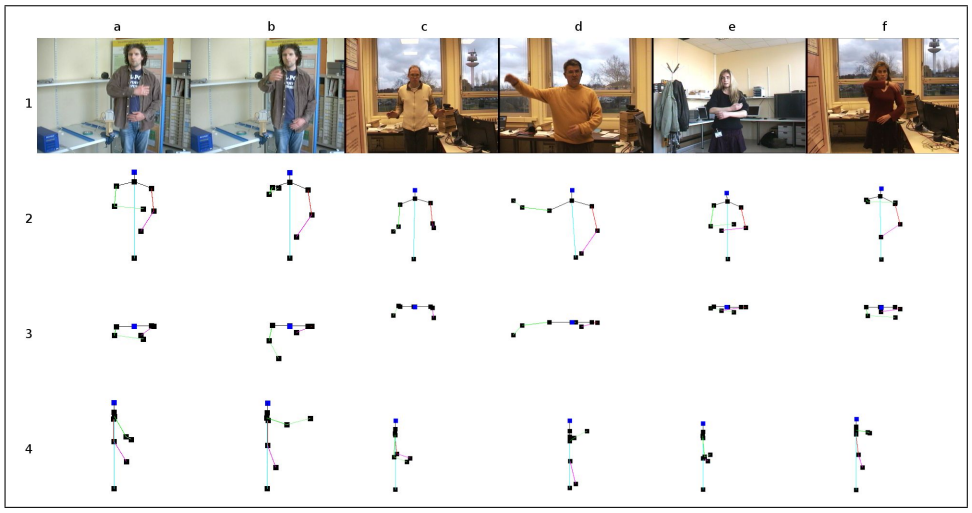

Figure 7: Monocular 3D tracking. Challenging poses are shown including occlusions, cluttered background and unconstrained environment (lighting and clothes).

Error (cm)	Shoulder	Elbow	Wrist	Overall Mean Error
Mean	1.7	7.1	9.7	6.1
Max	6.1	24.1	31.0	13.4
Std. Dev.	1.0	3.5	6.6	2.6
Average Speed ($cm.s^{-1}$)	2.83	4.28	8.5	

Table 1: Mean, maximum and standard deviation of the estimated position error for shoulder, elbow and wrist. Overall mean error is the mean error made on estimating the pose of theses three joints. Average speed is computed for the whole test sequence on each joint.

In monocular tracking, significant errors are usually made on depth estimation. It is the case in the test sequence around frame 500 owing to a wrong estimated elbow position that constrains the wrist in an exaggerated forward position. A similar problem occurs around frame 850 where forearm bends perpendicularly to the image plane and wrist depth is wrongly estimated by our algorithm (figure 8). Anyway, the maximal estimated pose error stays below 31 cm and below 15 cm considering the measure protocol used in [12, 14] (table 1). The comparison with other tracking algorithms is a difficult task owing to the disparity between used test sequences. However, the obtained results outperform or are as accurate than those computed with existing algorithms [12, 14].

6 Conclusion

We have presented an algorithm for monocular upper body tracking performing at 6 *fps* using a standard webcam with unconstrained environments (lighting and clothes). The used cues based on contours provide sufficient robustness to succeed on unconstrained environments. Belief propagation provides a judicious solution in order to reduce the

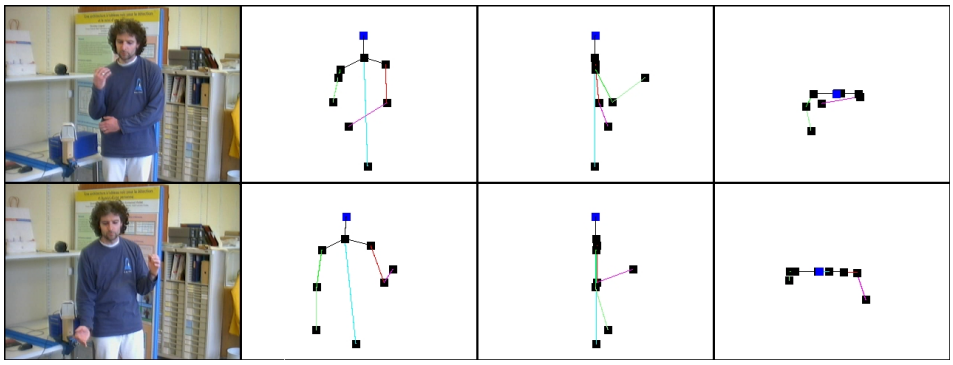

Figure 8: Examples of wrong depth estimation on frames 581 (first row) and 850 (second row). In both cases, right forearm is not bended sufficiently involving errors larger than 25 cm on wrist pose estimation.

space dimension of the generated hypothesises making particle filtering framework suitable. Articulation constraints are easily integrated into factors computation to provide consistent resulting poses. Future work will include a learning based image compatibility term to handle occlusions and more accurate depth estimation.

References

[1] Ankur Agarwal and Bill Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV (1)*, pages 50–59, 2006.

[2] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28(1), jan 2006.

[3] Olivier Bernier and Pascal Cheung-Mon-Chang. Real-time 3d articulated pose tracking using particle filtering and belief propagation on factor graphs. In *British Machine Vision Conference*, volume 01, pages 005–008, 2006.

[4] Andrew Blake and Michael Isard. The condensation algorithm - conditional density propagation and applications to visual tracking. In *NIPS*, pages 361–367, 1996.

[5] David Demirdjian, T. Ko, and Trevor Darrell. Constraining human body tracking. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1071. IEEE Computer Society, 2003.

[6] Raphaël Féraud, Olivier Bernier, Jean Emmanuel Viallet, and Michel Collobert. A fast and accurate face detector based on neural networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 23(1):42–53, 2001.

[7] Jiang Gao and Jianbo Shi. Multiple frame motion inference using belief propagation. In *FGR*, pages 875–882, 2004.

[8] Kschischang, Frey, and Loeliger. Factor graphs and the sum-product algorithm. *IEEETIT: IEEE Transactions on Information Theory*, 47, 2001.

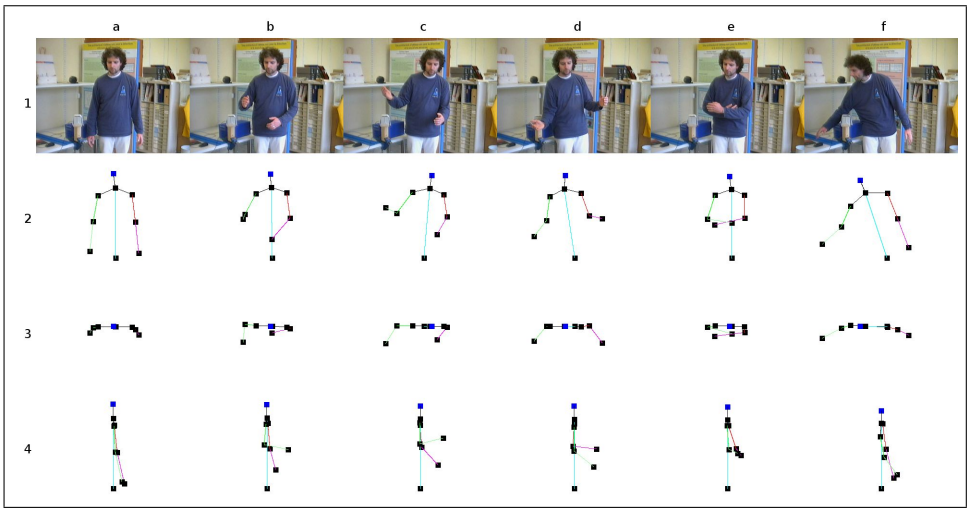

Figure 9: Test sequence from which the ground truth were measured. From column (a) to (f): frames 246 (initialisation), 409, 709, 813, 1184 and 1437. The face, top and right side estimated poses are shown on row 2, 3 and 4.

[9] Philippe Noriega and Olivier Bernier. Real time illumination invariant background subtraction using local kernel histograms. In *Proceedings of the British Machine Vision Conference*, pages 979–988, 2006.

[10] Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 750. IEEE Computer Society, 2003.

[11] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. Tracking loose-limbed people. In *CVPR (1)*, pages 421–428, 2004.

[12] Leonid Sigal and Michael J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2041–2048, Washington, DC, USA, 2006. IEEE Computer Society.

[13] Cristian Sminchisescu and Alexandru Telea. Human pose estimation from silhouettes - a consistent approach using distance level sets. In *WSCG*, pages 413–420, 2002.

[14] Leonid Taycher, David Demirdjian, Trevor Darrell, and Gregory Shakhnarovich. Conditional random people: Tracking humans with crfs and grid filters. In *CVPR (1)*, pages 222–229, 2006.

[15] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 238–245, Washington, DC, USA, 2006. IEEE Computer Society.