Monocular 3D Gait Tracking in Surveillance Scenes

Grégory Rogez, *Member, IEEE*, Jonathan Rihan, Jose J. Guerrero, *Member, IEEE*
and Carlos Orrite, *Member, IEEE*

***Abstract*—Gait recognition can potentially provide a noninvasive and effective biometric authentication from a distance. However, the performance of gait recognition systems will suffer in real surveillance scenarios with multiple interacting individuals and where the camera is usually placed at a significant angle and distance from the floor. We present a methodology for view-invariant monocular 3D human pose tracking in man-made environments in which we assume that observed people move on a known ground plane. First, we model 3D body poses and camera viewpoints with a low dimensional manifold and learn a generative model of the silhouette from this manifold to a reduced set of training views. During the online stage, 3D body poses are tracked using recursive Bayesian sampling conducted jointly over the scene's ground plane and the pose-viewpoint manifold. For each sample, the homography that relates the corresponding training plane to the image points is calculated using the dominant 3D directions of the scene, the sampled location on the ground plane and the sampled camera view. Each regressed silhouette shape is projected using this homographic transformation and matched in the image to estimate its likelihood. Our framework is able to track 3D human walking poses in a 3D environment exploring only a 4 dimensional state space with success. In our experimental evaluation, we demonstrate the significant improvements of the homographic alignment over a commonly used similarity transformation and provide quantitative pose tracking results for the monocular sequences with high perspective effect from the CAVIAR dataset.**

***Index Terms*—Monocular gait tracking, 3D body pose, video-surveillance, view-invariance, particle filtering.**

I. Introduction

VIDEO-based human gait analysis constitutes a very active field of research as shown by the recent special issue of the journal [1]. Gait recognition can potentially provide an noninvasive and remote biometric authentication. Most papers in the litterature consider isolated individuals walking in laboratory type environments. However, the performance of gait recognition systems will suffer in real surveillance scenarios with multiple interacting individuals and where the camera is usually placed at a significant angle and distance from the floor. While parameters such as variation of walking speeds [2], low resolution [3], severe occlusions [4], gait kinematics [5] and camera viewpoint [6] have been dealt with rather separately, we address all these problems together and propose a system which can track the 3D pose of multiple walking people in complex surveillance scenarios. Methods such as [7] could further be considered to recognize the identity based on the recovered 3D human joints.

G. Rogez, J.J. Guerrero and C. Orrite are with the Aragon Institute for Engineering Research (I3A) , Universidad de Zaragoza, 50017 SPAIN, e-mail: {grogez,josechu.guerrero, corrite}@unizar.es. J. Rihan was with the Dept of Computing, Oxford Brookes University, Oxford, UK. This work was supported by Spanish grants TIN2010-20177, DPI2012-31781, FEDER and by the regional government DGA-FSE

Example-based approaches have been very successful for human motion tracking but their accuracy depends on the similarity of the camera viewing angle in test and training images. The goal of this work is thus to track and estimate the 3D body pose of multiple walking people by means of view-based models independently of the point of view from which the scene is observed (see Fig. 1a), even in cases of high tilt angles and perspective distortion. The idea is to learn mappings between a body pose manifold and 2D silhouette features (shape) from as few training views as possible (Fig. 1b-c). The challenge is then to make use of those mappings successfully on any possible sequence taken from a single fixed camera with an arbitrary viewing angle. Supposing that the observed person walks on a planar ground in a calibrated environment, we use the homography relating the image points to the closest training plane on the viewing hemisphere. This projective transformation is applied when computing the observation likelihood of each sampled shape in a particle filtering framework. It compensates for the effect of both discretization along azimuth angle θ and variations along elevation angle φ, thus alleviating the effect of perspective distortion. Our main contributions are:

1) We combine the best components of state-of-the-art human pose trackers and exploit projective geometry in an efficient particle filtering framework for monocular 3D gait tracking in calibrated surveillance scenes: in our proposal only a 4-dimensional state space needs to be explored to track walking human poses in 3D world.
2) We replace the usual 2D similarity transformation relating image and model planes by a homography-based alignment. Our results demonstrate that the incorporation of this perspective correction in the tracking framework results in a higher tracking rate and allows for a better shape-based estimation of body poses under wide viewpoint variations using only 8 training views.
3) We propose an efficient likelihood computation whose only clues are edges and background subtraction resulting in a fast top-down shape matching. We also introduce a new state estimator.
4) We numerically demonstrate the efficiency of our algorithm by processing a set of challenging gait sequences[1]: our system successfully tracks the 3D pose of walking pedestrians in cases where a small number of people move together, have occlusion, and cast shadow.

[1]The lack of availability of standard test dataset, i.e. surveillance videos with perspective distortion and corresponding 2D/3D poses, encouraged us to build our own dataset (we labelled 2784 2D-poses on CAVIAR [10]) which will be made available to the scientific community for further research.

TABLE I
COMPARISON OF THE SETTINGS AND PERFORMANCES OF OUR ALGORITHM W.R.T. STATE-OF-THE-ART METHODS FOR MONOCULAR 3D POSE TRACKING. THESE APPROACHES SHARE OTHER SIMILARITIES WITH OUR WORK IN THAT THEY USE SILHOUETTE FEATURES AND LOW DIMENSIONAL POSE MANIFOLDS WITH PARTICLE FILTERING. THE MAIN DIFFERENCES ARE LISTED BELOW. NOTE THAT WE GIVE A RANGE OF VALUES FOR THE REQUIRED NUMBER OF PARTICLES PER INDIVIDUAL TRACKER AS THIS NUMBER VARIES DEPENDING ON SINGLE OR MULTIPLE TARGETS.

Authors	Settings			Performances				
	Scene Calib.	State Dim.	Training Views	Pose Evaluation	Localization	Multiple Targets	Perspective Videos	No. Particles
Jaeggli et al [8]	No	10	36	Qualitative	Yes	No	No	500
Elgammal et al [9]	No	2	12	Numerical	No	No	No	900
Our work	**Yes**	**4**	**8**	**Numerical**	Yes	**Yes**	**Yes**	**250-1000**

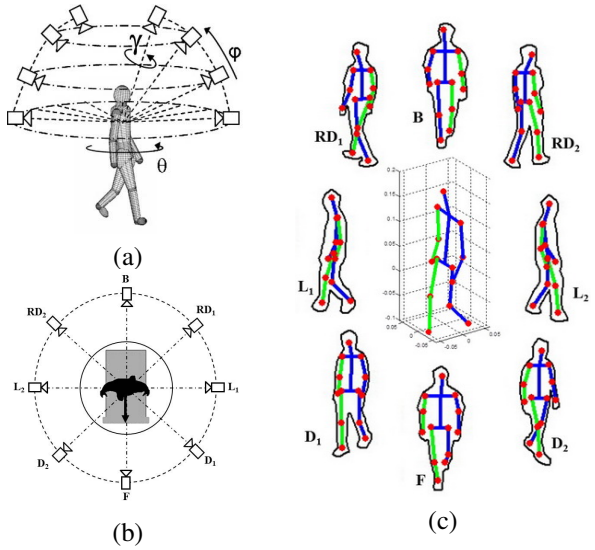


Fig. 1. (a) The position of the camera with respect to the observed subject, the view, can be parameterized as the combination of two angles: the *elevation* $\varphi \in \left[0, \frac{\pi}{2}\right]$ (also called latitude or tilt angle) and the *azimuth* $\theta \in [-\pi, \pi]$ (also called longitude) define the viewing hemisphere. A third angle $\gamma \in [-\pi, \pi]$ can be considered to parameterize the rotation around the viewing axis. (b) Training viewpoint discretization: in this work, we use the MoBo dataset [11] and discretize the viewing hemisphere into 8 locations where θ is uniformly distributed around the subject. (c) Training data used in this paper: example of a 3D pose and its 8 view-based 2D silhouettes and 2D poses.

A. Related Work

3D body pose tracking. Stochastic models have come to be the dominant way of approaching the problem of articulated 3D human body tracking: an approximate inference technique, usually a particle filtering, is used to tractably estimate the high-dimensional posture space [8], [9], [12]–[16]. Particle filtering allows modeling non-Gaussian multi-modal distributions and can maintain multiple hypotheses through time. However, the number of particles required to achieve an acceptable result considerably increases with the dimensionality of the search space. The number of degrees-of-freedom (generally more than 30) and the high dimensionality of the state space (i.e. valid poses) make the tracking problem computationally difficult. The search space gets even larger when the tracking algorithm also has to estimate the location, orientation and scale of the subject in the image or in the scene as in [8]. Some work has investigated the use of learnt models of human motion to constrain the search in state space by providing strong priors on motion [17], [18]. Others have focused their research on the problem of dimensionality reduction for pose tracking and proposed to use low dimensional embedding of human motion data: Gaussian process latent variable model (GPLVM) [19]–[21], Locally Linear Embedding (LLE) [8], supervised manifold learning [9], [16] or coordinated mixture of factor analysers [14] are some examples. More recently, Daubney et al. [22] introduced a method based on Pictorial Structures [23] to track articulated poses using the motion of a sparse set of moving features.

Most existing systems typically assume that the camera axis is parallel to the ground, i.e. $\varphi = 0$ in Fig. 1a, and that the observed people are vertically displayed ($\gamma = 0$). The viewpoint is discretized in a circle around the subjects, selecting a set of values for the azimuth θ: 36 orientations in [8], 16 in [24], [25], 12 in [9] and 8 in [21], [26]–[28]. In our approach, a general camera orientation ($\gamma, \varphi \neq 0$) is considered and a reduced 4-dimensional state is proposed for 3D pose tracking. In Tab. I, we compare the settings of our algorithm w.r.t. the most similar state-of-the-art approaches [8], [9]. Results have been presented using different testing datasets in laboratory environments like HumanEva [29] or challenging street views as in [8], [21], but generally training and test images are captured in similar environments or with a similar camera tilt angle. Very few present numerical evaluation of human pose tracking on surveillance scenarios with low resolution and perspective distortion, and few pose tracking algorithms exploit the key constraints provided by scene calibration which is available in most surveillance systems.

Viewpoint dependence is one of the bottlenecks of human motion analysis [30]. Considering a person walking along a straight line and making a constant angle with the image plane, a side-view can be synthesized using a homography as in [31]. A reconstruction method can be employed to rectify and normalize gait features recorded from different viewpoints into this side-view plane, exploiting such data for human recognition as in [32]. This rectification method is based on the anthropometric properties of human limbs and the characteristics of the walking action (see [33]). The concept of "virtual cameras" [34], allows for the reconstruction of synthetic 2D features from any camera location. The joint log-likelihood of body pose and camera parameters is maximized and results in an estimate of the 3D body pose. In [35], [36] we proposed to exploit projective geometry in training and test images to find viewpoint invariance. These papers are discussed in details in the next section. In the same spirit,

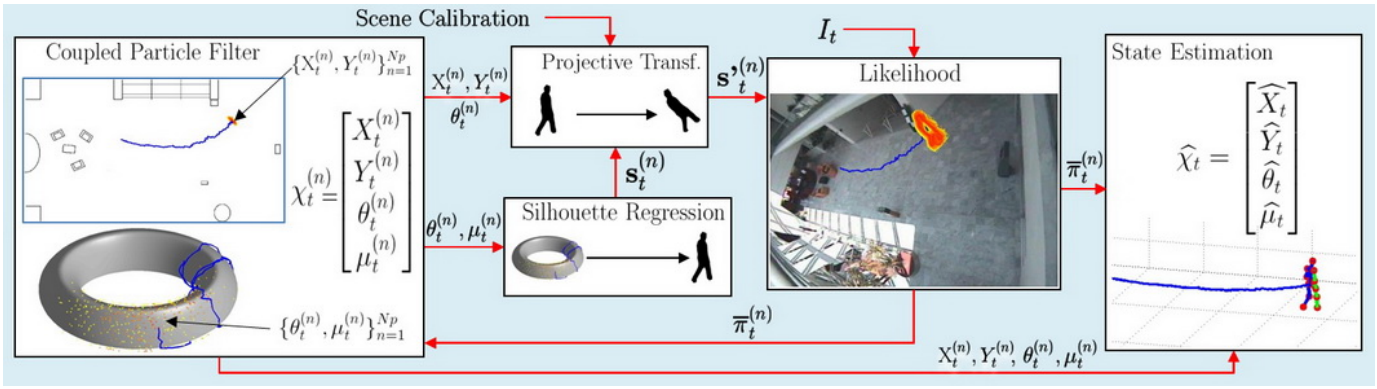


Fig. 3. System Flowchart: the 3D body poses are tracked using a recursive Bayesian sampling conducted jointly over the scene's ground plane (X, Y) and the pose-viewpoint (θ, μ) torus manifold ([9]). For each sample n, a projective transformation relating the corresponding training plane and the image points is calculated using the dominant 3D directions of the scene, the sampled location on the ground plane $(X_t^{(n)}, Y_t^{(n)})$ and the sampled camera view $\theta_t^{(n)}$. Each regressed silhouette shape $\mathbf{s}_t^{(n)}$ is projected using this homographic transformation obtaining $\mathbf{s}_t'^{(n)}$ which is later matched in the image to estimate its likelihood and consequently the importance weight. A state, i.e. an oriented 3D pose in 3D scene, is then estimated from the sample set.

a homographic transformation was later employed in [37] to improve human detection in images presenting perspective distortion. An improvement in detection rate from 38.3% to 87.2% was reported using 3D scene information instead of scanning over 2D on the same testing dataset [10] we consider.

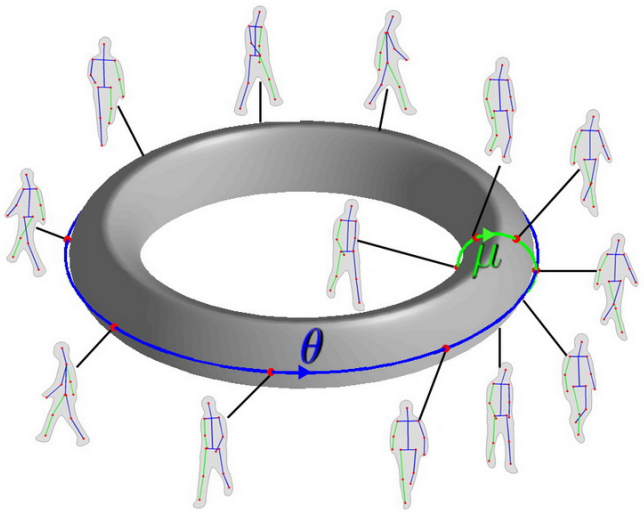


Fig. 2. Pose-viewpoint torus manifold (adapted from [9]) learned using the Mobo dataset (see Fig. 1): the 2 dimensions of the surface (θ, μ) represent camera viewpoint and gait cycle. We represent 8 different views of a same pose (blue circle), and 6 different poses from a same viewpoint (green circle).

B. Overview

We tackle the problem of view-invariant 3D gait tracking and explore the use of projective shape matching in a particle filtering framework which jointly explores a low dimensional pose-viewpoint manifold and the real world ground plane. Our approach is motivated by the encouraging preliminary results for view-invariant human motion analysis obtained in our earlier work [35], [36] and the recent advances in low-dimensional manifold learning for human pose tracking [8], [9].

The basic idea is that perspective distortion impairs a correct *top-down* silhouette matching or *bottom-up* pose inference and that projective geometry could be exploited when camera viewpoints in training and testing images are too different. In [35], we numerically demonstrated that the use of a projective transformation for shape registration, projecting both model and image in a canonical vertical view, improves silhouette-based pose estimation. In that case, the parameters required to estimate the homography, i.e. the subject's location on the ground plane (X, Y) and the viewpoint θ (Fig. 1a), were taken from ground truth. Later in [36], we proposed a framework where pedestrians were tracked using a Kalman filter to estimate X, Y and θ. For each frame, the input image was projected on the camera plane of the nearest training view and the pose was estimated in a *bottom-up* manner using the corresponding view-based model from [28]. Acceptable results were obtained for sequences with a single walking subject but we identified two main drawbacks: 1) the employed model did not handle pose ambiguities and did not recover from drift and, 2) more importantly, the result greatly depended on the accuracy achieved when estimating X , Y and θ.

In this paper, we propose a stochastic approach for estimating both location and viewpoint and improve the search of the optimum projective transformation for pose recognition by sampling multiple values for θ at multiple locations (X, Y). Applying a different projective transformation to the input image for each sampled triplet (X, Y, θ) and processing each resulting warped image as in [36] would be computationally inefficient. We instead propose a *top-down* approach where for each triplet, a silhouette is sampled (in training plane), transformed using the inverse projection and later matched in the original input image. Given its proven effectiveness, we choose to model 3D walking poses using a low dimensional torus manifold for viewpoint θ and embedded pose $\mu \in [0,\ 1)$ as in [9]. We map this manifold to our view-based silhouette manifolds using kernel-based regressors, which are learnt using a Relevance Vector Machine. Given a point on the surface of the torus, the resulting generative model can regress the corresponding pose and view-based silhouette (see Fig. 2).

During the online stage, 3D body poses are thus tracked using a recursive Bayesian sampling conducted jointly over the scene's ground plane and this pose-viewpoint torus manifold. For each sample, the homography that relates the corresponding training plane to the image points can be calculated using the dominant 3D directions of the scene, the sampled location on the ground plane and the sampled camera view. Each regressed silhouette shape is then projected using the associated homographic transformation and matched in the image to estimate its likelihood. Our tracking framework is depicted in Fig. 3. In our experimental evaluation, we demonstrate the significant improvements of the homographic matching over a commonly used similarity transformation and provide quantitative 3D pose tracking results for the challenging surveillance sequences from CAVIAR [10].

The rest of the paper is organized as follows. The geometrical considerations and the computation of the projective transformation are described in Section II while the torus manifold for pose and appearance modeling is presented in Sect. III, In Sect. IV, we detail our tracking framework. Experimentations with qualitative and quantitative evaluations are presented in Section V and some conclusions are finally drawn in Section VI.

II. Projective Transformation

In the presence of perspective distortion neither similarity nor affine model provide a good approximation for the transformation between a prior shape and a shape to segment. However, a planar projective transformation is a better approximation even though the object shape contour is roughly planar as demonstrated in [38]. We thus propose to find a projective transformation, i.e. a homography, between training and testing camera views to alleviate the effect of perspective distortion on silhouette-based human motion analysis and compensate for the effect of training viewpoint discretization.

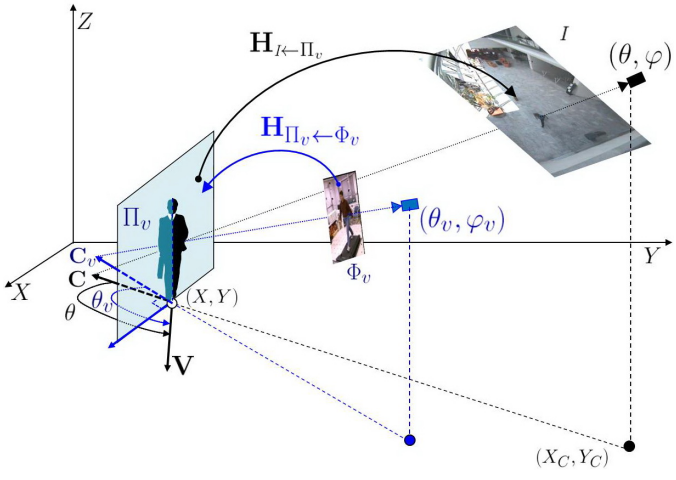


Fig. 4. Schematical representation of the transformation between 2 images through a vertical plane: the transformation $\mathbf{P}_{I\Pi_v\Phi_v}$ which relates training image plane Φ_v and testing image I through a vertical plane Π_v is obtained as the product of $\mathbf{H}_{I\leftarrow\Pi_v}$ and $\mathbf{H}_{\Pi_v\leftarrow\Phi_v}$. The facing direction $\mathbf{V}$ is the orientation vector relating the azimuth θ and the camera viewing direction $\mathbf{C}$. Note that in training view Φ_v, we have $\theta=\theta_v$ and $\mathbf{C}=\mathbf{C}_v$.

The 3×3 transformation $\mathbf{P}_{I\Pi_v\Phi_v}$ (illustrated in Fig. 4) between the model plane Φ_v and input image I through the vertical plane Π_v can potentially be obtained as the product of two homography transformations defined up to a rotational ambiguity:

$$\mathbf{P}_{I\Pi_v\Phi_v} = \mathbf{H}_{I\leftarrow\Pi_v}\cdot\mathbf{H}_{\Pi_v\leftarrow\Phi_v}, \tag{1}$$

where $\mathbf{H}_{\Pi_v\leftarrow\Phi_v}$ relates the points on the training plane Φ_v with the vertical plane Π_v and $\mathbf{H}_{I\leftarrow\Pi_v}$ relates this frontal plane with the input image I . These two homography transformations can be computed using the vanishing points of the dominant 3D directions of Π_v as explained in Appendix A. $\mathbf{H}_{\Pi_v\leftarrow\Phi_v}$ is evaluated off-line for each training view $\{\Phi_v\}_{v=1}^{N_v}$ while $\mathbf{H}_{I\leftarrow\Pi_v}$ is computed on-line using the vanishing points evaluated during the calibration stage. The ground plane position (X,Y) and the viewpoint angle θ, which are required to localize the vertical plane Π_v, are estimated by the tracker as explained later in Sect. IV. The readers are referred to [39] for more detailled information on the computation of the projective transformation.

III. Torus Manifold for Pose and Appearance Modeling

Full body pose configurations are necessarily high dimensional; in our case, we use 13 3D-joint locations in a human-centered coordinate system for our representation which results in a 39-dimensional pose configuration. To reduce the problem of high dimensionality in the learning stages, a dimensionality reduction step is needed to identify a low-dimensional embedding of the pose space. As we focus on the walking action which is cyclic, we consider a low-dimensional manifold embedding both camera viewing angle and body pose together and jointly model them by means of a torus manifold. Elgammal et al [9] numerically demonstrated with experimental evaluation that the supervised torus embedding shows much better performances than unsupervised manifold representations (LLE, Isomap, GPLVM).

If $\mu\in[0,\,1)$ is the body pose configuration on the torus and $\theta\in[-\pi,\pi]$ is the viewing angle[2], then the torus manifold illustrated in Fig. 2b can be defined parametrically in Euclidean space by:

$$\begin{aligned} x &= (R + r\cos 2\pi\mu)\cos\theta,\\ y &= (R + r\cos 2\pi\mu)\sin\theta,\\ z &= r\sin 2\pi\mu \end{aligned} \tag{2}$$

where R is the "major radius", i.e. the distance from the center of the tube to the center of the torus, and r is the "minor radius", i.e. the radius of the tube.

A. Body Pose Modeling

The training sequences are mapped onto the surface of the torus. We refer the reader to [9] for details. We learn the mapping back to the original data space from the torus manifold with a kernel regressor:

$$\mathbf{K} = f_p(\mu,\theta) = \mathbf{W}_p\mathbf{\Phi}_p(x,y,z), \tag{3}$$

[2]Note that for consistency with our viewing hemisphere parameterization, we keep the viewpoint parameter as an angle while Elgammal et al [9] define both viewpoint and action parameter in $[0,\,1)$ space.

where $\mathbf{K} \in \mathbb{R}^{3\times 13}$ is the orientated body pose configuration in the original 3D pose training space, $\mathbf{\Phi}_p$ is a vector of kernel functions and $\mathbf{W}_p$ is a matrix of weights[3]. The matrix $\mathbf{W}_p$ is learnt using a Relevance Vector Machine (RVM). We use radial basis functions as the kernel functions in $\mathbf{\Phi}_p$ computed at the training data locations. Any point $(\mu, \theta) \in [0,\ 1) \times [-\pi, \pi]$, on the surface of the torus, can be directly mapped to an oriented 3D pose.

B. Appearance Modeling

Different shape representations have been used for human silhouettes in recent literature, including parametric B-splines [40], shape context [41], [42], level-sets [43], pose-adaptive shape descriptors [44] and distance transform [8], [9]. We select the landmark parameterization [45]–[47], i.e. a set of N_l 2D-landmarks, to represent the silhouette:

$$\mathbf{s} = [\mathbf{x}_{s_1}, ..., \mathbf{x}_{s_{N_l}}] \in \mathbb{R}^{2\times N_l}. \tag{4}$$

Although non-linearity and normalization issues can appear during the training phase [28], landmark-based shape representations are lower dimensional and much simpler to manipulate and transform. They also facilitate a very quick matching with the image making them ideal in a top down particle filtering framework. We refer the reader to [28] for a description of the training shapes normalization and alignment.

We now model the generative mapping from embedded pose μ to silhouette descriptors $\mathbf{s}$ that allows us to predict image appearance given an hypothesis for the pose μ and for the body orientation or camera viewpoint θ. In this work, the viewing hemisphere is discretized into a finite number N_v of training viewpoints $\{\theta_v\}_{v=1}^{N_v}$ varying the azimuth angle (see example in Fig. 2). For each training viewpoint a mapping is learnt from the torus manifolds to the corresponding view-based silhouette manifold, which are learnt using a Relevance Vector Machine (RVM):

$$\mathbf{s} = f_s(\mu, \theta),\ \forall \mu \in [0,\ 1)\,,\ \forall \theta \in \{\theta_v\}_{v=1}^{N_v}, \tag{5}$$

with

$$f_s(\mu, \theta) = \mathbf{W}_s \mathbf{\Phi}_s(x, y, z). \tag{6}$$

Once again, the mapping $f_s(\mu, \theta)$ is learnt using RVM with weights $\mathbf{W}_s$ and kernel functions $\mathbf{\Phi}_s(\mu, \theta)$. Given a point $(\mu, \theta) \in [0,\ 1) \times \{\theta_1, \cdots, \theta_{N_v}\}$ on the torus manifold, the resulting generative model can generate the corresponding view-based silhouette. Note that in this work, the shape descriptor $\mathbf{s}$ is augmented with the 13 2D-joints $\mathbf{k} = [\mathbf{x}_{k_1}, ..., \mathbf{x}_{k_{13}}] \in \mathbb{R}^{2\times 13}$ to facilitate the estimation of a 2D pose error in the experiment Section.

IV. Recursive Bayesian Sampling

A. Formulation.

At each time step, we simultaneously perform body pose estimation and image localisation since both processes can benefit from the coupling of the posture and image location as demonstrated in [8]. The advantage of assuming a calibrated environment and a planar ground plane is the considerable reduction of the search space as image location, scale and rotation can be recovered from the real world ground plane location using the homography relating the 2D points of the image and this plane. Thus, we define the state vector of the target as:

$$\chi_t = [X_t\ Y_t\ \theta_t\ \mu_t\ ], \tag{7}$$

consisting of the real-world ground plane location (X_t, Y_t) and the embedding coordinates on the torus surface $(\mu_t, \theta_t) \in [0,\ 1) \times [-\pi, \pi]$. The calibration of the scene and the torus embedding help us to face a much more tractable problem as the search has to be performed in a 4-dimensional state space while, for instance, Jaeggli et al [8] explore a 10-dimensional space.

We formulate the tracking problem as a Bayesian inference task, where the state of the tracked subject is recursively estimated at each time step given the evidence (image data) up to that moment. Formally, within the Bayesian filtering framework, we formulate the computation of the *posterior* distribution $p(\chi_t|\mathbf{I}_t)$ of our model parameters χ_t over time as follows:

$$p(\chi_t|\mathbf{I}_t) \propto p(I_t|\chi_t)\, p(\chi_t|\mathbf{I}_{t-1}), \tag{8}$$

where $\mathbf{I}_t$ is the image sequence up to time *t* and $p(I_t|\chi_t)$ is the *likelihood* of observing the image I_t given the parameterization χ_t of our model at time *t*, in other words the *observation density*. Finally $p(\chi_t|\mathbf{I}_{t-1})$ is the *a priori density*, which is the result of applying the *dynamic model* $p(\chi_t|\chi_{t-1})$ to the *a posteriori density* $p(\chi_{t-1}|\mathbf{I}_{t-1})$ of the previous time step:

$$p(\chi_t|\mathbf{I}_{t-1}) = \int p(\chi_t|\chi_{t-1})\, p(\chi_{t-1}|\mathbf{I}_{t-1})\, d\chi_{t-1}. \tag{9}$$

Unfortunately, when the involved distributions are non-Gaussian, Eq. (8) cannot be solved analytically. Instead, we use a particle filter [12], [40], [48] in order to approximate the true *posterior* pdf $p(\chi_t|\mathbf{I}_t)$ by means of a discrete weighted set of samples $\{\chi_t^{(n)}, \pi_t^{(n)}\}_{n=1}^{N}$:

$$p(\chi_t|\mathbf{I}_t) \approx \sum_{n=1}^{N} \pi_t^{(n)} \delta(\chi_t^{(n)}), \tag{10}$$

where for each particle, δ denotes the Dirac delta and $\pi_t^{(n)}$ is the normalized importance weight which is directly derived from measurement likelihood:

$$\pi_t^{(n)} = \frac{p(I_t|\chi_t^{(n)})}{\sum_{n'=1}^{N} p(I_t|\chi_t^{(n')})}, \tag{11}$$

as defined in [40].Hence, whilst the likelihood function decides which particles are worth propagating, the dynamic model is responsible for guiding the exploration through the state space.

B. Dynamic Model.

Since a static camera is being considered in this work, and assuming the people face along the direction of motion, we model the dependence of viewpoint on ground plane location while we assume statistical independence between the

[3]As done in [9], we map from the Euclidean space where the torus lives and not from the coordinate system (μ, θ) since this coordinate system is not continuous at the boundary.

remaining state variables[4]. The dynamic model $p(\chi_t|\chi_{t-1})$ is thus a product of four dynamic models, i.e. $p(\chi_t|\chi_{t-1}) = p(X_t|X_{t-1},\theta_{t-1})p(Y_t|Y_{t-1},\theta_{t-1})p(\theta_t|\theta_{t-1})p(\mu_t|\mu_{t-1})$.

Therefore, our state model has the following form on the torus manifold:

$$\theta_t = \theta_{t-1} + n_\theta, \tag{12}$$

$$\begin{bmatrix}\mu_t \\ \dot{\mu}_t\end{bmatrix} = \begin{bmatrix}1 & \delta t \\ 0 & 1\end{bmatrix}\begin{bmatrix}\mu_{t-1} \\ \dot{\mu}_{t-1}\end{bmatrix} + \begin{bmatrix}n_\mu \\ n_{\dot{\mu}}\end{bmatrix}, \tag{13}$$

where n_θ, n_μ and $n_{\dot{\mu}}$ are zero mean white Gaussian noises (whose variances are set to $\sigma_\theta = \frac{\pi}{10}$, $\sigma_\mu = 0.075$ and $\sigma_{\dot{\mu}} = 0.0125$ respectively) and δt the time interval between successive frames, and on the ground plane:

$$\begin{bmatrix}X_t \\ Y_t\end{bmatrix} = \begin{bmatrix}X_{t-1} \\ Y_{t-1}\end{bmatrix} + n_V\begin{bmatrix}V_X \\ V_Y\end{bmatrix} + n_{XY}\begin{bmatrix}1 \\ 1\end{bmatrix} \tag{14}$$

where n_{XY} and n_V are zero mean white Gaussian noise (with variance set to $\sigma_{XY} = 1$cm and $\sigma_V = 10$cm in our experiments) and $[V_X, V_Y]^T = \mathbf{V}/||\mathbf{V}||$ is the unit orientation vector relating θ and the camera viewing direction $\mathbf{C}$ (see Fig. 4). In this way, we model the fact that pedestrians are more likely to move in the facing direction and, after a stationary phase, we can predict in which way the subject is going to move based on his body orientation (i.e. viewpoint angle θ).

C. Image Measurements - Observation Model.

The *likelihood* function $p(I|\chi)$ computes how likely it is to observe the image I given the unknown state χ. Given the viewpoint θ and the location on the ground plane (X, Y), a training view Φ_v with angle θ_v is selected and the transformation $\mathbf{P}_{I\,\Pi_v\Phi_v}$ that relates the training plane Φ_v to the image points is calculated using the X, Y, θ and the dominant 3D directions of the scene following Sect. II. The regressed silhouette descriptor $\mathbf{s} = f_s(\mu, \theta_v)$ is then projected on the image I obtaining $\mathbf{s}' = [\mathbf{x}'_{s_1}, ..., \mathbf{x}'_{s_{N_l}}] \in \mathbb{R}^{2\times N_l}$. For each landmark $l \in \{1, N_l\}$, $\mathbf{x}'_{s_l} = (x'_{s_l},\ y'_{s_l})$ is obtained following:

$$[x'_{s_l},\ y'_{s_l},\ 1]^T = \mathbf{P}_{I\,\Pi_v\Phi_v} \cdot [x_{s_l},\ y_{s_l},\ 1]^T,\ \forall l \in \{1, N_l\}. \tag{15}$$

The *likelihood* $p(I|\chi)$ is now estimated using this projected silhouette $\mathbf{s}'$. To keep the required time for computing the likelihood of each sample as low as possible, we chose to employ only low-level processing tasks like background subtraction or edge detection algorithms. Thus, the observations are based on the edge map I_{edges} of the image, as well as the binary foreground detection mask I_{fgd}. The pixel color values are not considered in this work. The joint likelihood is approximated as:

$$p(I|\chi) = p(I_{edges}|\chi)p(I_{fgd}|\chi). \tag{16}$$

The projected silhouette shape $\mathbf{s}'$ is used to compute the first likelihood term $p(I_{edges}|\chi)$ using a Chamfer distance function as in [49]. Both silhouette $\mathbf{s}'$ and edge map I_{edges} are first decomposed into a number N_γ of separate orientation channels according to gradient orientation. The elements of $\mathbf{s}'$ are thus decomposed into N_γ lists of landmark indexes $\{\Gamma_\gamma\}_{\gamma=1}^{N_\gamma}$, i.e. $\forall l \in \{1, N_l\},\ \exists!\ \gamma \in \{1, N_\gamma\} : l \in \Gamma_\gamma$. A Distance Transform (DT) of the edge image I_{edges} is then computed separately for each channel obtaining $\{\mathcal{D}_\gamma\}_{\gamma=1}^{N_\gamma}$ and a Chamfer distance $d_{Ch} \in [0, 1]$ is computed [5] as:

$$d_{Ch}(\mathbf{s}', I_{edges}) = \frac{1}{\tau.N_l}\sum_{\gamma=1}^{N_\gamma}\sum_{l\in\Gamma_\gamma}\mathcal{D}_\gamma(x'_{s_l}, y'_{s_l}), \tag{17}$$

where τ, the upper bound on the distance to the edge, is used to threshold the DT image and increase robustness toward partial occlusion as indicated in [49]. We consider $\tau = 5$ pixels and $N_\gamma = 4$ orientation bins. Note that the elements in $\mathbf{s}'$ are rounded off before the computation of the Chamfer distance, i.e. $\mathbf{s}' \in \mathbb{N}^{2N_l}$. The edge based likelihood function is then defined as:

$$\begin{aligned} p(I_{edges}|\chi) &= p(I_{edges}|\mathbf{s}'), \\ &\propto \exp(-\lambda_e\ d_{Ch}(\mathbf{s}', I_{edges})), \end{aligned} \tag{18}$$

i.e. a Laplacian distribution over the distance d_{Ch} as in [49]. In this work, we select $\lambda_e = 4$ (see Fig. 5).

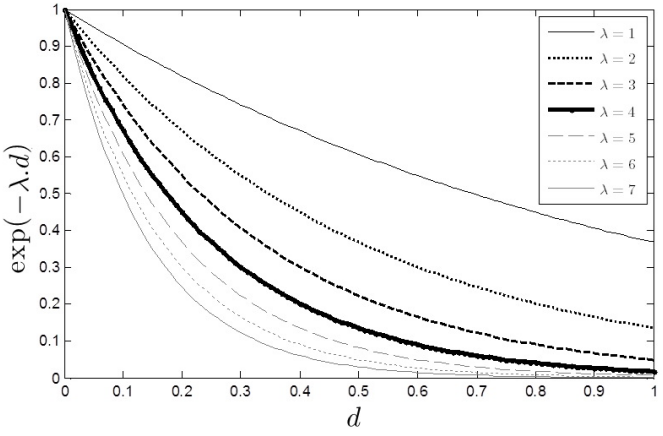
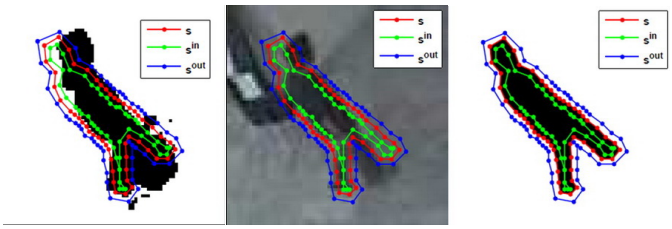


Fig. 5. Likelihood function as a Laplacian distribution $\exp(-\lambda\ d)$ over the distance $d \in [0, 1]$ for several values of the parameter λ. In this work, we consider an acceptable Likelihood function is obtained with $\lambda = 4$.



(a) (b) (c)

Fig. 6. Foreground Likelihood: (a) The shape $\mathbf{s}'$ is matched on an input foreground blob. The new shapes $\mathbf{s}^{in}$ and $\mathbf{s}^{out}$ (in green and blue respectively) are the inner and outer boundary points, 2 pixels away from the original landmarks in $\mathbf{s}'$. The resulting distance $d_{Fgd}(\mathbf{s}', I_{fgd}) = 0.44$. The same shapes are represented on top of the original input image in (b), while in (c) we visualize the same 3 shapes on top of a foreground blob returning $d_{Fgd}(\mathbf{s}', I_{fgd}) = 0$, i.e. a perfect likelihood $p(I_{fgd}|\chi) = 1$.

[4]We choose not to model the dependencies between the gait parameter μ and the ground plane location because stride length depends on the subject morphology and walking style.

[5]The DT image takes the set of feature points as input and assigns each location the distance to its nearest feature.

The second likelihood term $p(I_{fgd}|\chi)$ aims at comparing two binary silhouettes: $\mathbf{s}'$ and the detection blob from I_{fgd}, obtained by state-of-the-art background subtraction. In surveillance videos, the problem is not as straightforward as it appears: occlusions, shadows, cluttered background, motion blur and low image resolution lead to low quality foreground silhouettes. Some approaches thus consider the foreground image as a binary mask to select foreground edges [50]. To cope with low quality foreground silhouettes, we introduce a new way to measure the fitness of a shape in a binary image. From $\mathbf{s}'$, we define two sets of 2D-landmarks constituting two new shapes $\mathbf{s}^{in}, \mathbf{s}^{out} \in \mathbb{N}^{2N_l}$, the inner boundary points and outer boundary points so that $\forall l \in \{1, N_l\}$:

$$\begin{bmatrix} x^{in}_{s_l} \\ y^{in}_{s_l} \end{bmatrix} \triangleq \begin{bmatrix} x'_{s_l} \\ y'_{s_l} \end{bmatrix} - \delta_s \mathbf{u}^{\perp}_{s_l} \text{ and } \begin{bmatrix} x^{out}_{s_l} \\ y^{out}_{s_l} \end{bmatrix} \triangleq \begin{bmatrix} x'_{s_l} \\ y'_{s_l} \end{bmatrix} + \delta_s \mathbf{u}^{\perp}_{s_l}, \tag{19}$$

where $\mathbf{u}^{\perp}_{s_l}$ is a unit vector passing through the landmark l and perpendicular to the shape, pointing outside of the shape. In this work, we consider $\delta_s = 2$ pixels (see examples in Fig. 6). We define $\mathcal{S}_{in}, \mathcal{S}_{out} \in [0, 1]$:

$$\begin{aligned} \mathcal{S}_{in} &\triangleq \tfrac{1}{N_l} \textstyle\sum_{l=1}^{N_l} I_{fgd}(x^{in}_{s_l}, y^{in}_{s_l}), \\ \mathcal{S}_{out} &\triangleq 1 - \tfrac{1}{N_l} \textstyle\sum_{l=1}^{N_l} I_{fgd}(x^{out}_{s_l}, y^{out}_{s_l}) \end{aligned} \tag{20}$$

as the shape-to-foreground and shape-to-background similarities respectively. $\mathcal{S}_{in}$ indicates the amount of foreground pixels inside the shape $\mathbf{s}'$ while $\mathcal{S}_{out}$ informs on the quantity of background outside $\mathbf{s}'$. Finally the resulting distance is defined as:

$$d_{Fgd}(\mathbf{s}', I_{fgd}) \triangleq 1 - \frac{\mathcal{S}_{in} + \mathcal{S}_{out}}{2}. \tag{21}$$

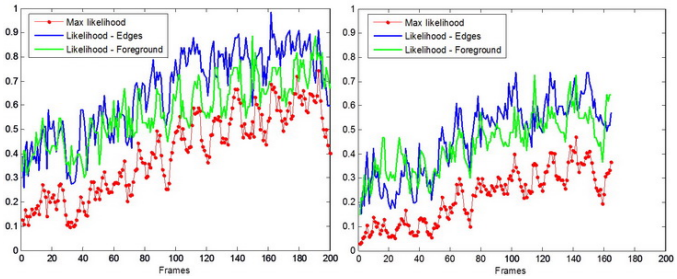
The rationale behind this definition of d_{Fgd} is that the distance metric should be high with noisy segmentation (occlusions and shadows) and zero with perfect matches. The likelihood is modeled, again, as a Laplacian distribution over this new distance measure:

$$p(I_{fgd}|\chi) \propto \exp(-\lambda_f \, d_{Fgd}(\mathbf{s}', I_{fgd})), \tag{22}$$

where λ_f is chosen so that the two likelihood terms have the same importance in Eq. 16, i.e. $\lambda_f = \lambda_e = 4$. See examples in Fig. 7. On a state-of-the-art laptop with an Intel Core @ 1.73GHz, the average computation time of the likelihood is about 2 ms per sample (0.4 ms for $p(I_{edges}|\chi)$ and 1.6 ms for $p(I_{fgd}|\chi)$ in unoptimized Matlab code and considering $N_l = 50$ landmark points to parameterize the shapes. In Fig. 8, the likelihood of the entire sample set can be visualized for four frames of the Walk2 sequence.

D. Tracking Multiple Pedestrians.

There is an extensive literature on particle filtering for tracking multiple interacting targets with a single calibrated camera [51]–[54]. Visual interactions among targets can be exploited when defining the likelihood term of a multiple-object filter as done in the BraMBLe system [53]. Dealing with occlusion is simplified when using the 3D positions of the multiple targets and even more when a roof-top surveillance camera is used. With such cameras, the heads are almost always visible and



(a) (b)

Fig. 7. Likelihood Examples: for Walk1 (a) and Walk2 (b) sequences. For each frame, we give the largest likelihood value (*red*) over a sample set of 500 particles, and the corresponding foreground $p(I_{fgd}|\chi)$ (*green*) and edges $p(I_{edges}|\chi)$ (*blue*) likelihood terms. In both sequences, the subject follows a similar path but, in Walk1 (a) the subject wears dark clothes while in Walk2 (b) the subject wears pale clothing against a pale background, explaining the highest likelihood values in (a) compared to (b). See dataset details in Tab. II.

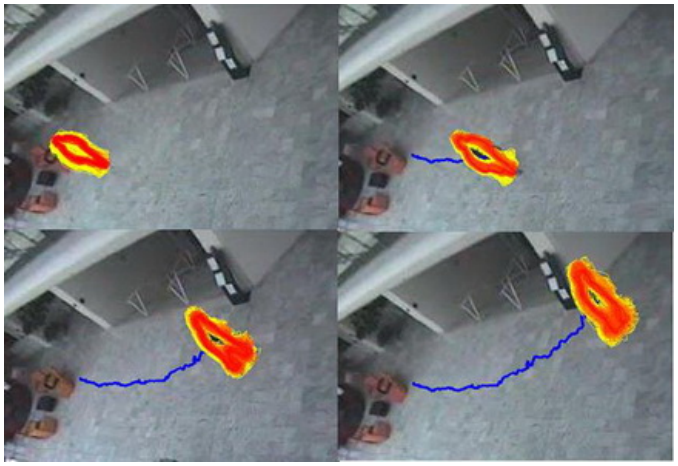


Fig. 8. Visualization of the likelihood for the entire sample set in several frames of the Walk2 sequence. For each frame, we plot the set of sampled and aligned shapes: darker colors indicate a higher likelihood.

complete occlusions rarely occur. We thus choose to simply instantiate several independent 1-subject trackers and follow a simple but effective approach to avoid the coalescence of the trackers onto the best-fitting target (e.g. subject wearing dark clothes with a pale background): we model each subject's 3D occupancy on the ground floor with a Gaussian probability function centered on the subject's estimated location which is then employed to downweight the particles from the other targets. Considering N_s subjects/targets with N_s individual trackers, the samples of each tracker are reweighted accordingly and the normalized importance weight of the n^{th} particle of subject s at time t thus becomes:

$$\pi^{(n)}_{t,s} \propto p(I_t|\chi^{(n)}_{t,s}) \prod_{\substack{s'=1 \\ s' \neq s}}^{N_s} (1 - \lambda_o \, \exp(-\frac{||\widehat{\chi}_{t-1,s'} - \chi^{(n)}_{t,s}||^2_{gd}}{\sigma^2_o}))^{\eta_o}, \tag{23}$$

with $\sum_{n=1}^{N} \pi^{(n)}_{t,s} = 1$ and where $||.||_{gd}$ is the Euclidean distance on the ground floor and σ_o, λ_o and η_o are defined empirically. Results provided in the next section are obtained with $\sigma_o = 50$ cm, $\lambda_o = 1$ and $\eta_o = 3$.

The approach we follow to deal with multiple targets may

Algorithm 1: Particle Filter Algorithm

Initialize a sample set $\{\chi_{0,s}^{(n)}, \frac{1}{N}\}_{n=1}^{N}$ for each subject s according to prior distribution $p(\chi_0)$;

for *each time step t* **do**
  for *each subject s* **do**
    for *each particle n* **do**
      Resample $\{\chi_{t-1,s}^{(n)}, \pi_{t-1,s}^{(n)}\}_{n=1}^{N}$ to obtain a new sample $\{\chi_{t-1,s}^{\prime(n)}, 1\}$;
      Propagate $\chi_{t-1,s}^{\prime(n)}$ using the dynamic model $p(\chi_t|\chi_{t-1})$ to obtain $\chi_{t,s}^{(n)}$;
      Compute likelihood $p(I_t|\chi_{t,s}^{(n)})$ from Eq. 16;
      Update weight $\pi_{t,s}^{(n)}$ using Eq. 23;
    Normalize N weights $\pi_{t,s}^{(n')} = \pi_{t,s}^{(n)} / \sum_{n=1}^{N} \pi_{t,s}^{(n)}$;
    Estimate the state $\widehat{\chi}_{t,s}$;

seem simplistic at first sight. However, it performs sufficiently well for the cases we have considered in this work, i.e. no severe occlusions and basic interactions where a few people meet, chat and walk together. The problem of multiple target tracking in more complex situations is out-of-scope for this paper and we leave for future work the use of a multiple-object filter or a more adequate modeling of the interactions. The complete tracking algorithm is summarized in Alg. 1.

The state $\widehat{\chi}_t$ at a particular time step is usually estimated using a Monte Carlo approximation of the expectation of the posterior pdf, i.e. a weighted sum over the set of samples: $\widehat{\chi}_t^{MC} = \mathcal{E}[\chi_t] = \sum_{n=1}^{N} \pi_t^n \chi_t^n$.

An estimation of the state can also be made by selecting one of the particles. For instance, the maximum a posteriori (MAP) estimate $\widehat{\chi}_t^{MAP}$ given by the particle with the largest normalized weight has been broadly considered, especially for human pose estimation problems [9]. The Viterbi path finding algorithm can also be considered to choose one of the samples at each time t and form a trajectory through time and state space that best satisfies both observation likelihood and temporal prior as in [8]. Viterbi estimate $\widehat{\chi}_t^{Vit}$ takes into account temporal consistency and can solve possible ambiguities and multi-modal distributions, which often happen in articulated body tracking. The particle filter will usually be able to concentrate particles in the main mode of the likelihood function. However, multiple modes of similar size in the likelihood function might bias MC estimation $\widehat{\chi}_t^{MC}$. MAP and Viterbi based methods require a large number of particles to reach the optimal position precisely leading to high computational cost. Moreover it is not guaranteed that the optimal position is necessarily sampled, even when a large number of particles are employed.

In this work, we propose a new hybrid way of estimating the state at each time step which is derived based on the discrete approximation of the posterior but also takes advantage of the temporal consistency of a Viterbi based estimate. We first define $\mathcal{N}_t^{Vit}$ a neighborhood around $\widehat{\chi}_t^{Vit}$, the sample selected by Viterbi, and consider a local weighted sum of the particles belonging to that neighborhood:

$$\widehat{\chi}_t = \frac{\sum_{n \in \mathcal{N}_t^{Vit}} \pi_t^{(n)} \chi_t^{(n)}}{\sum_{n \in \mathcal{N}_t^{Vit}} \pi_t^{(n)}}, \tag{24}$$

where the neighborhood $\mathcal{N}_t^{Vit}$ is defined in the coupled ground floor-torus state space with a circular region around the Viterbi estimate $\widehat{\chi}_t^{Vit}$ on both ground floor and torus surface:

$$\mathcal{N}_t^{Vit} \triangleq \left\{ n \ : \ ||\widehat{\chi}_t^{Vit} - \chi_t^{(n)}||_{gd} \leq \rho_{gd} \ \wedge \ ||\widehat{\chi}_t^{Vit} - \chi_t^{(n)}||_{tor} \leq \rho_{tor} \right\}, \tag{25}$$

where $||.||_{gd}$ and $||.||_{tor}$ are the Euclidean distance on ground floor and torus manifold respectively while ρ_{gd} and ρ_{tor} are the two radii defining the neighborhood on the ground floor and torus manifold respectively. In our experiments, we set $\rho_{gd} = 10$ cm and $\rho_{tor} = 0.1$.

V. Experimental Results

The comparison with state-of-the-art work is not straightforward for several reasons. First, standard testing data sets for pose estimation (e.g. HumanEva [29]) do not consider perspective distortion and multiple interacting subjects, and can not be used in this paper to offer a quantitative comparison. We will instead employ the Caviar dataset [10] that presents very challenging sequences with perspective distortion but, as far as we know, no pose estimation results (apart from our work) have been published on this dataset. The ground truth labelled for this paper on the Caviar dataset will be made publicly available to the scientific community for further research and comparison. We present in Tab. II the selected sequences where one or several subjects walk and interact.

Many papers consider 3D body pose estimation or localization in real-world images separately. Few papers [8], [21], [55] tackle both problems simultaneously as we do, but they pay no attention to the problem of perspective distortion (they consider a camera elevation angle $\varphi = 0$) and do not include scene knowledge in their frameworks. Since our system is more complete than state-of-the-art methods and takes into account a calibration of the camera w.r.t the scene, running these algorithms on the proposed testing dataset and making a comparison with our results would be unfair.

Nevertheless, we will compare the performances of our complete framework based on projective geometry with a simpler solution considering a shape alignment based on similarity and keeping the rest of the framework unchanged. The similarity is defined as:

$$\mathbf{T} \cdot \mathbf{x} = \mathbf{u} + s\mathbf{R}(\gamma) \cdot \mathbf{x}, \ \forall \mathbf{x} \ \in \mathbb{R}^2, \tag{26}$$

in which $(\mathbf{u}, \gamma, s)$ are offset, rotation angle and scaling factor respectively. These parameters are readily calculated using head center $\mathbf{x}_H$ and "feet"/ location on the ground floor $\mathbf{x}_F$ in image and model views. This points are recovered in the test image from the real world ground plane coordinate (X, Y) using two 3×3 homography matrices which are calculated during the off-line calibration stage: $\mathbf{H}_{\mathrm{g}}$ which characterizes the mapping between the ground plane in the image and the real world ground plane Π_{gd} and $\mathbf{H}_{\mathrm{h}}$ relating the head plane in the image with Π_{gd}.

TABLE II

TESTING DATASET CONSIDERED IN THIS PAPER. THE SELECTED SEQUENCES BELONG TO THE CAVIAR DATASET [10]. ELEVEN TRACKS OF WALKING PEOPLE HAVE BEEN CONSIDERED FOR MANUAL GROUND TRUTH ANNOTATION (I.E. MANUAL LOCALIZATION OF THE 13 2D-JOINTS DEFINING A 2D POSE). FOR EACH TRACK, WE INDICATE THE SELECTED FRAMES, THE SUBJECT ID, THE NUMBER OF AVAILABLE GROUND TRUTH 2D POSES AND A SHORT DESCRIPTION WITH POSSIBLE DIFFICULTIES.

Sequence	**Track**	**Subject ID**	**Frames**	**No. of Poses**	**No. of Frames**	**Description - Difficulty**
Walk1	1	1	260-459	200	200	Dark clothing - good segmentation
Walk2	2	2	0304-0468	165	289	Pale clothing - bad segmentation
	3		0931-1054	124		
Walk3	4	3	0500-0649	150	310	Dark clothing - good segmentation
	5		1200-1359	160		
LeftBag_PickedUp	6	4	0314-0413	100	100	Subject carries a bag
Meet_WalkTogether2	7	2	190-509	320	320	Varied clothing color - Occlusions
	8	1	209-506	298		
Meet_Split_3rdGuy	9	5	077-742	666	666	Varied clothing color - Occlusions
	10	3	189-506	318		
	11	6	332-614	283		
Total		**6**		**2784**	**1885**	

TABLE III

PERCENTAGE OF TRACKING FAILURE. FOR EACH OF THE 11 SELECTED TRACKS (SEE DETAILS IN TAB. II), WE PRESENT THE AVERAGE PERFORMANCE OVER 20 RUNS OF THE TRACKING ALGORITHM FOR DIFFERENT NUMBER OF SAMPLES: A TRACK IS CONSIDERED LOST WHEN TRACKING HAS FAILED DURING 20 FRAMES OR MORE (THE DISTANCE BETWEEN THE NEAREST PARTICLE AND GROUND TRUTH LOCATION IS OVER 1 METER) AND IT HAS NOT RECOVERED BY THE END OF THE SEQUENCE. RESULTS ARE GIVEN FOR THE TWO TYPES OF IMAGE ALIGNMENT: SIMILARITY TRANSFORMATION AND THE PROPOSED HOMOGRAPHIC PROJECTION.

Alignment		**Similarity**							**Homography**						
No. Particles		20	50	100	250	500	1000	2000	20	50	100	250	500	1000	2000
Track	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0
(no	2	95	85	60	5	0	0	0	95	60	20	0	0	0	0
occlusion)	3	45	0	0	0	0	0	0	40	5	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	10	0	0	0	0	0	0	0	0	0	0	0	0	0
Track	7	65	40	25	40	30	25	20	50	45	10	0	0	0	0
(with	8	35	5	10	5	0	0	0	40	0	0	0	0	0	0
occlusions)	9	100	70	45	30	10	5	10	95	65	45	20	15	0	0
	10	35	30	15	30	10	5	10	10	15	10	15	15	0	0
	11	20	0	0	0	0	0	0	40	0	0	0	0	0	0
Average		**37.27**	**21.36**	**14.09**	**9.55**	**4.55**	**3.18**	**3.64**	**33.64**	**16.82**	**7.73**	**3.18**	**2.73**	**0**	**0**

Existing methods implicitly [8], [21], [55] or explicitly [56] apply a similarity transformation between their models and the processed images, most of the time with only scale and translation elements but without any rotation in the image plane (i.e. $\gamma = 0$ in Eq. 26)[6]. Thus, a comparison of the performances of our framework replacing the projective transformation in Eq. 15 by a similarity transformation will provide a quantitative evaluation of the improvement achieved by our proposal w.r.t. state-of-the-art.

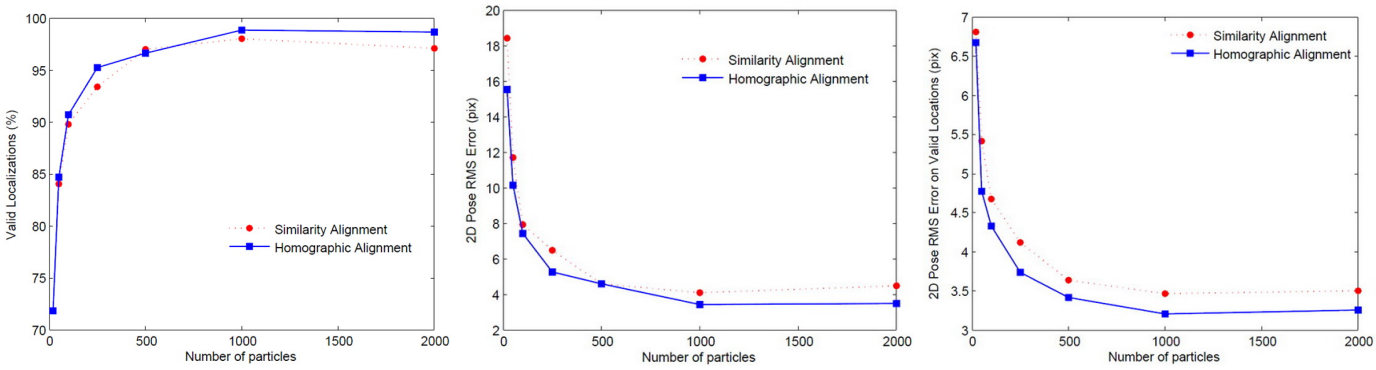
A. *Settings and Parameters.*

We use training silhouettes and 2D/3D poses extracted from the MoBo dataset [11] illustrated in Fig. 2a: for each one of the 8 training views, 15 walking cycles corresponding to 15 different subjects are temporally aligned, subsampled and averaged to compute a mean walking cycle made of 100 silhouettes and 2D poses. Thus, 800 silhouettes and associated 2D poses are used to learn the mapping between the torus manifold and the original data space. The same operation is performed with the 3D poses which are rotated around the Z-axis (azimuth θ) to cover the entire torus manifold. For each training view, we localize the horizontal and vertical vanishing points and compute the 8 homographies $\{\mathbf{H}_{\Phi_v \leftarrow \Pi_v}\}_{v=1}^{8}$.

We remove lens distortion from Caviar testing images and calibrate the camera w.r.t. the scene by localizing the vertical vanishing point and the horizontal vanishing line, and compute the homographies $\mathbf{H}_{\mathrm{g}}$ and $\mathbf{H}_{\mathrm{h}}$ from manual annotations.

In this work, we do not address the detection problem and take the ground plane location in the first frame from ground truth data, but the detector from [37] would perfectly suit our framework as it can deal with perspective distortion and has shown significantly improved detection performance on the Caviar dataset. When a subject appears in the scene, we initialize a tracker by sampling in the entire space of possible poses and probable viewpoints. Supposing that the subject is facing in the direction of motion, the most probable viewpoints can defined based on the location in the first frame and the

[6]Most of these techniques are based on a scheme where the images are scanned with a sliding window at different scales.



(a) (b) (c)

Fig. 9. Tracking Results. Average Performance over 20 runs of the tracking algorithm on the 11 tracks for similarity and homographic alignment vs number of particles. (a) Percentage of valid localizations (valid if the distance between nearest particle and ground truth location is below 1 meter), (b) 2D pose error over all the poses and (c) 2D pose error computed using only valid poses from (a). The pose error in (b) and (c) is computed as the MRSE between nearest particle and ground truth using the 13 2D-joints in pixels.

scene knowledge: e.g. if the subject is entering the scene by the right side, the viewpoint is most likely to be L_1, D_1 or RD_1 and viewpoint should be sampled in the corresponding part of the torus.

Note that, during tracking, the particles which fall in non-valid areas of the ground plane (such like walls, plants, etc) are assigned a 0 likelihood. When a subject is not moving, the likelihood is computed using only the shape landmarks corresponding to the upper part of the body. Since we model walking poses, our framework is not supposed to recognise standing poses. When motion is detected after a stationary phase, the tracker is reset by sampling in the entire space of possible poses.

B. Experiments.

We ran a series of tests on the selected Caviar sequences varying the number of particles in the filter. Since randomness is involved in the re-sampling of the particles, to gain better statistical significance, we perform the same experiments 20 times and from now on we compute numerical result as the average over these 20 runs. We repeat the same operation using a similarity transformation instead of our homographic transformation. First we propose to evaluate the performance of the tracker, independently of the state estimator. We thus consider that a target has been lost and the localization is not valid if the minimum distance (in the particle set) to ground truth location exceeds a certain value. We believe that a pose estimation does not make sense if the nearest particle is 1 meter away from the target true location, i.e. $\min_{n\in\{1,N\}}\left(||\widehat{\chi}_t^{Gt}-\chi_t^{(n)}||_{gd}\right) \geq 100$ cm. A track is then considered lost when then the target has been lost during 20 frames or more and has not been recovered in the last frame of the sequence.

Results show that, in average, the proposed homographic alignment reduces the average percentage of lost tracks as can be observed in Tab. III. The percentage of lost tracks decreases with the number of particles employed in the filter for both methods, but we reach 0% of lost tracks with 1000 particles and over while 5% of the tracks are still lost when considering 2000 particles and a similarity transformation. The perspective correction allows for better shape matching and consequently a more efficient shape-based tracking. If we look at the detailed results, we can observe how, on average, the homographic alignment outperforms the similarity alignment for 7 of the 11 tested tracks, i.e. it reaches 0% tracking failure with a smaller particle set (tracks 1, 2, 6, 7, 8, 9 and 10), while performing similarly for 3 other tracks (4, 5 and 11). Good tracking performance requires larger particle sets for the sequences presenting occlusions and multiple interacting people. We can also see that much fewer particles are required when a good foreground detection is available: a perfect result is obtained with our projective alignment method and only 20 particles in sequences where people wear dark clothes.

If we compute the average number of valid localizations, i.e. the cases where the distance between the nearest particle and ground truth location is below 1 meter, the tracking usually looses less targets when a homographic alignment is used rather than a similarity alignment, (see Fig. 9a). We even reach an average of 99% of valid localizations above 1000 particles.

We now evaluate the pose estimation performance and compute a RMS error between the 2D pose of the best particle and the ground truth 2D pose, thus evaluating the best pose that could be estimated from the posterior independently of the employed state estimator. In Fig. 9b, we see how the 2D pose error (RMSE between nearest particle and ground truth) decreases with the number of particles and how the framework, again, performs better when a projective transformation is used and allows for a more accurate pose estimation. If we compute the same error using only the valid localizations (from Fig. 9a,) we reach lower 2D pose errors, especially for small particle sets and for the similarity based approach (see Fig. 9c). This makes sense because of the larger amount of failed localizations which return a bad pose estimation and influence the average pose error.

To aid the comparison of pose estimation performance and focus on the pose estimation when localisation is satisfactory, from now on, we exclude the non-valid poses and the different errors are computed over poses from valid localizations only. More frames are then considered for our homography based

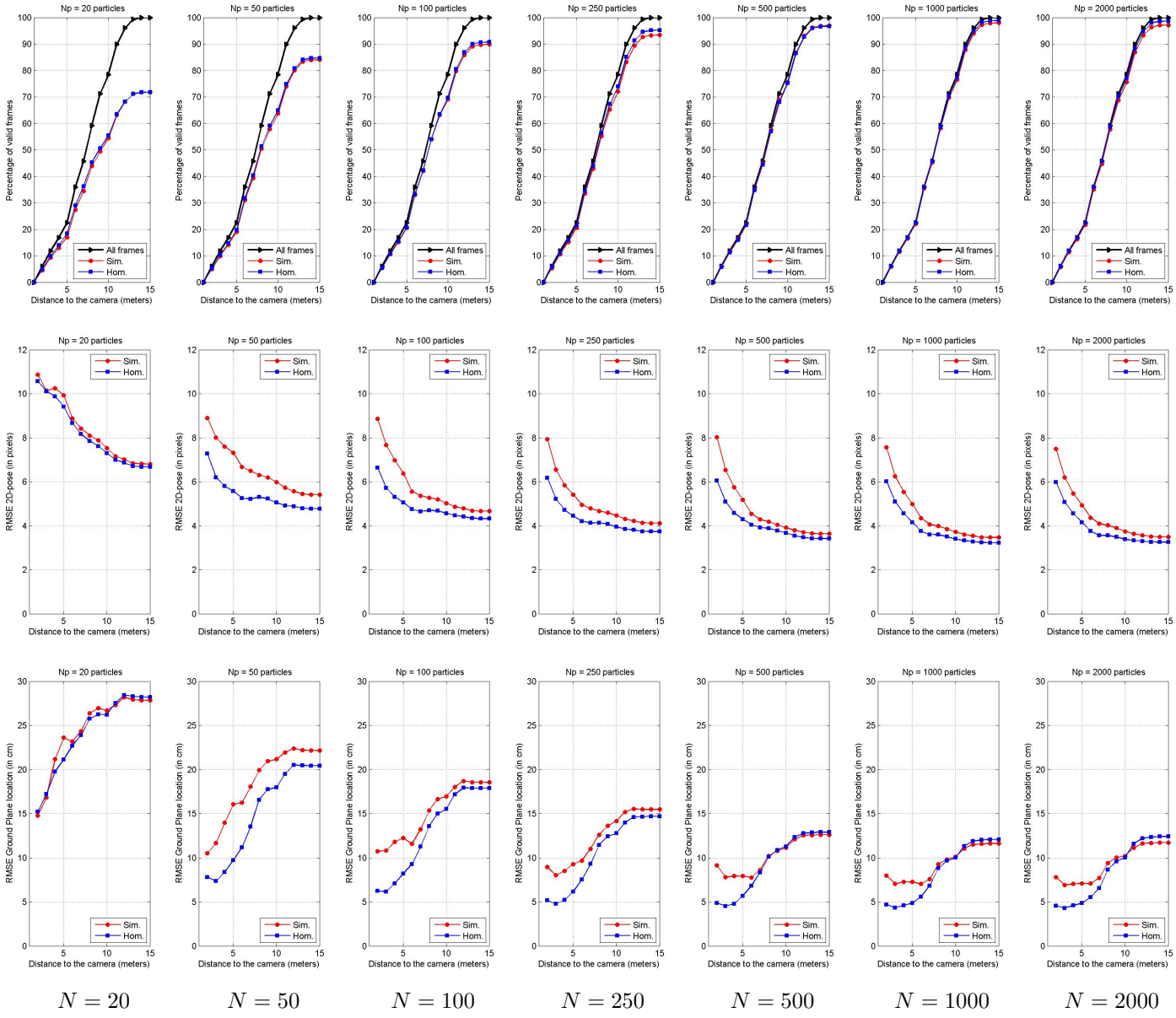


Fig. 10. Detailed performances w.r.t. the distance to the camera: percentage of valid localizations (top row), 2D pose error (middle row) and (X, Y) ground plane localization error (bottom row) are represented (from left to right) for 20, 50, 100, 250, 500, 1000 and 2000 particles. Note that the different values are computed using the poses from 0 meter up to the given distance to the camera. Only valid localizations from the top row have been used to compute the performances in middle and bottom rows. Again 2D pose error and (X, Y) ground plane localization error are computed as the MRSE between nearest particle and ground truth using the 13 2D-joints locations in pixels and the 2D location in cm respectively.

alignment because of its lower failure rate. We can see in Fig. 9c that the average 2D pose error obtained using our projective transformation is not too far from the result returned with a similarity transformation despite a qualitative improvement observed when watching the estimated poses and silhouettes.

This last observation inspired us to carry out a deeper analysis of the different results, in particular visualize the different rates in function of the distance between the subject and the camera. In Fig. 10, we present the average percentage of valid localizations, the average 2D pose error and the average ground plane location error varying the maximum distance to the camera. Results are presented for different sized particle sets. In the middle row, we can observe that the average pose error globally decreases as we augment the maximum distance to the camera and add new poses further away. The opposite happens with the ground plane location error. This is expected because when people move away from the camera their size in the image gets smaller. Thus, the 2D pose gets smaller when moving away from the camera leading to a consecutive lower 2D pose error while an accurate localization on the ground plane becomes more difficult with the distance. We should also point out that the different errors are computed using a ground truth data obtained from manual labelling and the accuracy and reliability of this labelling also decrease with the distance to the camera.

From Fig. 10, we can observe that the improvement in terms of pose and ground plane localization is globally obtained

TABLE IV
AVERAGE 2D POSE ERROR: PERFORMANCE OVER 20 RUNS OF THE TRACKING ALGORITHM ON THE 11 TRACKS FOR HOMOGRAPHIC AND SIMILARITY ALIGNMENTS. WE PRESENT THE RMS 2D POSE ERROR (IN PIXELS) WHEN ESTIMATING THE STATE AT EACH TIME STEP USING THE MONTE CARLO APPROXIMATION $\widehat{\chi_t}^{MC}$, THE MAP CRITERIA $\widehat{\chi_t}^{MAP}$, VITERBI PATH FINDING ALGORITHM $\widehat{\chi_t}^{Vit}$ AND THE WEIGHTED SUM AROUND THE VITERBI ESTIMATE $\widehat{\chi_t}^{Vit+WS}$.

Alignment		**Similarity**							**Homography**						
No. Particles		20	50	100	250	500	1000	2000	20	50	100	250	500	1000	2000
State	$\widehat{\chi_t}^{MC}$	**7.59**	**6.53**	**5.97**	**5.50**	5.13	4.98	5.14	**7.52**	**5.71**	**5.37**	**4.96**	4.75	4.61	4.72
Estimator	$\widehat{\chi_t}^{MAP}$	7.66	6.62	6.05	5.63	5.30	5.21	5.38	7.65	5.85	5.55	5.17	4.97	4.90	5.05
	$\widehat{\chi_t}^{Vit}$	7.77	6.74	6.18	5.63	5.17	4.98	5.07	7.78	6.00	5.66	5.16	4.84	4.67	4.73
	$\widehat{\chi_t}^{Vit+WS}$	7.72	6.65	6.08	5.53	**5.08**	**4.89**	**5.00**	7.70	5.88	5.52	5.02	**4.71**	**4.54**	**4.61**

when the subjects are close to the camera. This makes perfect sense since the viewpoint changes when a subject goes far away from the camera and tends to a tilt angle $\varphi = 0$ which is similar to the training viewpoint employed in this paper. It seems that when the subject moves far away from the camera, a projective transformation is not required and a similarity transformation could be enough.

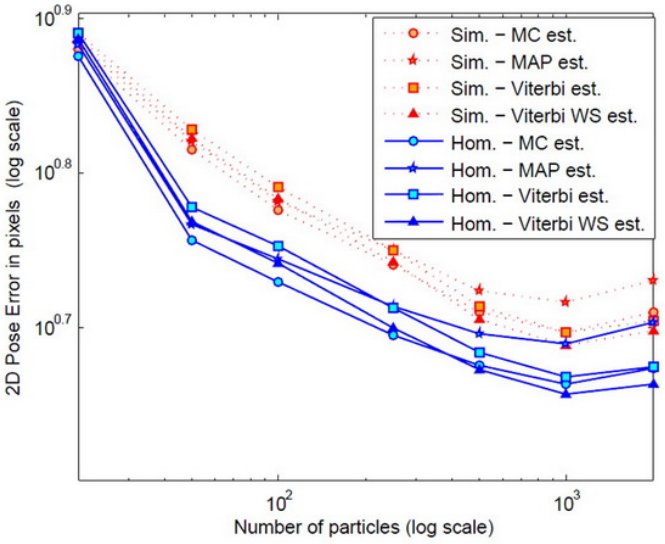


Fig. 11. Average 2D pose error varying the state estimator: performance over 20 runs of the tracking algorithm on the 11 tracks for homographic and similarity alignments. We present the RMS 2D pose error when estimating the state at each time step using the Monte Carlo approximation (MC), the MAP criteria, Viterbi path finding algorithm (Viterbi) and the weighted sum around the Viterbi estimate (Viterbi WS).

In Fig. 11, we present the average 2D pose error obtained when estimating the state at each time step using the Monte Carlo approximation (MC), the Maximum A Posteriori (MAP) criteria, the Viterbi path finding algorithm (Viterbi) and the proposed weighted sum around the Viterbi estimate (Viterbi WS)[7]. We present the results for the two different types of alignment. Corresponding numerical evaluation is given in Tab. IV. The first observation is that our homographic alignment clearly outperforms the similarity transformation independently of the employed state estimation technique. We can also observe that our proposed approach for state estimation outperforms all the other techniques for $N \geq 500$ particles while MC estimate is better for smaller particle sets.

In Fig. 12, we present qualitative results for tracks 1, 4 and 5 from Tab. II using our projective method for view-invariant pose tracking and 500 particles. For each sequence, we can observe the tracked silhouettes for a few frames and the trajectory of the subject in the image as well as the trajectory on the torus manifold and the estimated 3D poses which have been successfully tracked. If we look at the temporal evolution of the viewpoint, we can see that using the Viterbi algorithm and our approach, we achieve a smooth continuous estimation of the viewpoint angle θ while using a model constructed from a discrete set of training views. As in [9], we recover the typical sawtooth curve of the walking cycle but in our case with challenging perspective videos.

We present more qualitative results for 2 sequences with multiple interacting subjects in Fig. 13 (tracks 7 and 8) and Fig. 14 (tracks 9, 10 and 11) using our proposed method and 1000 particles. For each sequence, we show the result for 4 frames: the tracked silhouettes and the trajectories in the image, the trajectories on the torus manifold and the estimated 3D poses which have been successfully tracked despite the occlusions and the perspective effect.

Our method has its limitations: in the current settings where only one value is considered for $\varphi \approx 0$, the pose error increases considerably when the subjects get very close to the camera and φ is extreme. In these cases of top views, i.e. $\varphi \approx \frac{\pi}{2}$, the distance on the viewing hemisphere between training and testing images is too large to allow a correct shape matching as observed in Fig. 13 (blue subject in frame 415). This issue could be addressed by considering more training views with additional values for φ which could then be modelled as an additional third dimension in the torus manifold. A trade-off would then have to be made between the number of training views and the desired accuracy.

VI. CONCLUSIONS

We have presented a complete framework for view invariant 3D gait tracking in man-made environments from monocular surveillance videos with high perspective effect. We have assumed that the camera is calibrated w.r.t. the scene and that observed people move on a known ground plane, which are realistic assumptions in surveillance scenarios.

[7]Note that the state estimate used to model each subject's 3D occupancy on the ground floor in Eq. 23 is always computed using the Monte Carlo approximation as we want to compare the different state estimators from the same clouds of samples.

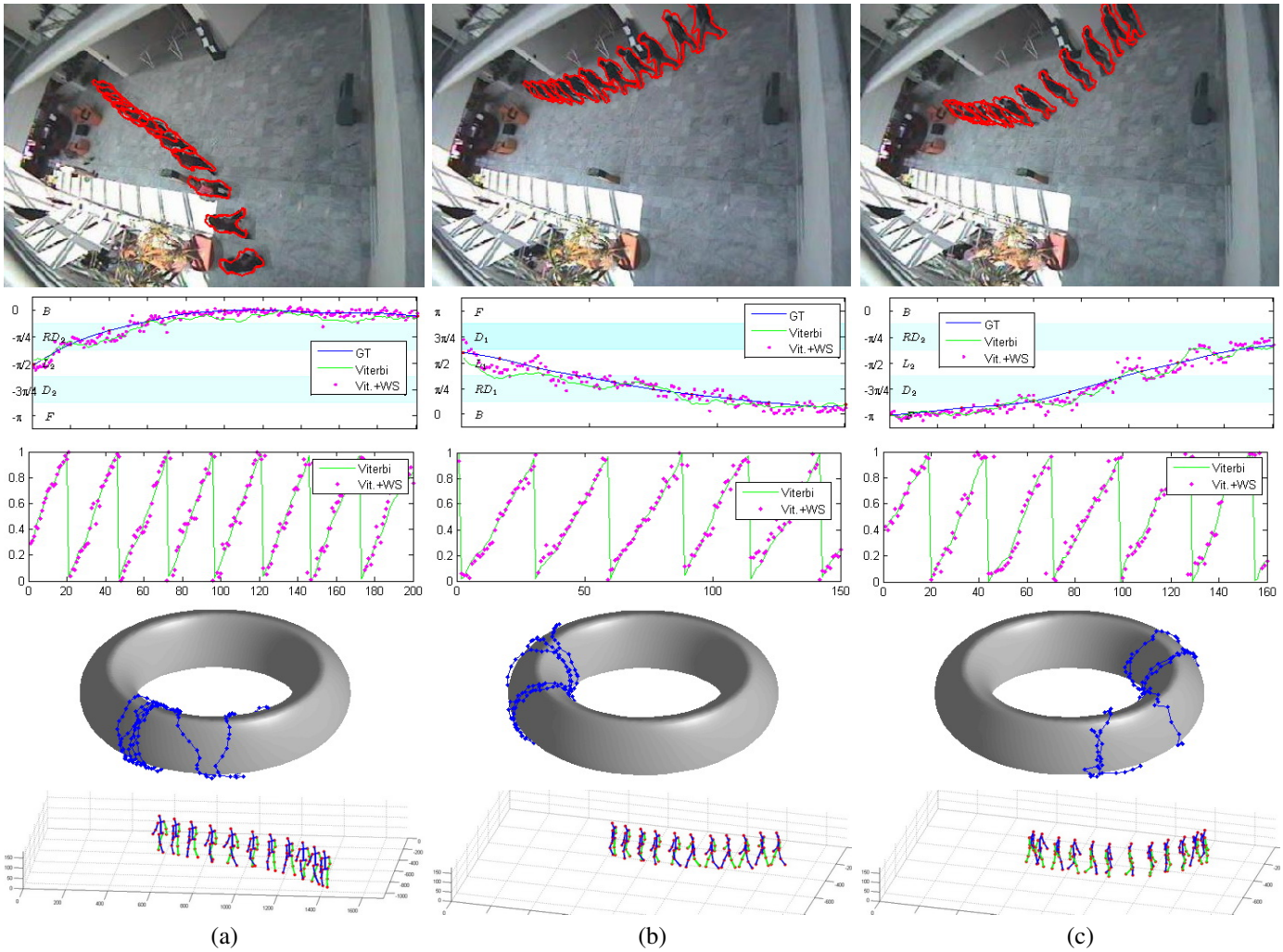


Fig. 12. Qualitative 3D pose tracking results for tracks 1 (a), 4 (b) and 5 (c) in Tab. II using our view-invariant pose tracking framework and 500 particles. For each sequence, we show from top to bottom: the tracked silhouettes for a few selected frames, the estimated viewpoint θ (vs ground truth), the estimated gait parameter μ, the trajectory of the subject on the torus manifold (μ and θ together) and the estimated 3D poses corresponding to the silhouette in 1^{st} row.

At the high level, the proposed approach follows a standard Bayesian tracking framework implemented as a particle filter. In order to reduce the search space, we rely on a supervised dimensionality reduction, mapping the space of human walking poses seen from various viewpoints onto a 2D torus manifold. The tracking is then performed in 4 dimensional space corresponding to position on the manifold and position of the person on the ground plane. The dynamical model is a first order model with constant velocity and Gaussian noise. It couples the position and orientation of the person, making use of the fact that people are more likely to move in the forward direction. The likelihood model is based on the silhouette-based representation. However, we focus on the particularly difficult case when the observed viewpoints during testing and training are significantly different. In order to address this, we have proposed to estimate the homography transformation between the training and test views and use this projective transformation to compensate for changes in silhouettes due to the novel viewpoint.

We have conducted a series of experiments to quantitatively and qualitatively evaluate our framework for a wide variety of viewing angles and a variety of sequences from the CAVIAR dataset, some with multiple interacting subjects and occlusions. We have demonstrated that exploiting projective geometry alleviates the problems caused by roof-top cameras with high tilt angles, and have shown that using a mapping from a low dimensional pose manifold to 8 training views was enough to produce good results when using a projective alignment for silhouette matching. Our approach is shown to be effective in tracking multiple walking people in the 3D environment and estimating their 3D body poses.

In future work, multi-stage particle filters, gradient ascent techniques or a more complex study of the posterior could be employed for searching for the optimal solution of the estimate at each time step. To deal with more subtle details such as variation of step length or motion style, an extra dimension could be considered in the framework to represent the walking style as in [9]. An extra dimension could also be used to model the elevation angle. Finally, adapting our projective view-invariant method for uncalibrated scenes and unconstrained environments offers another intriguing line for future research.

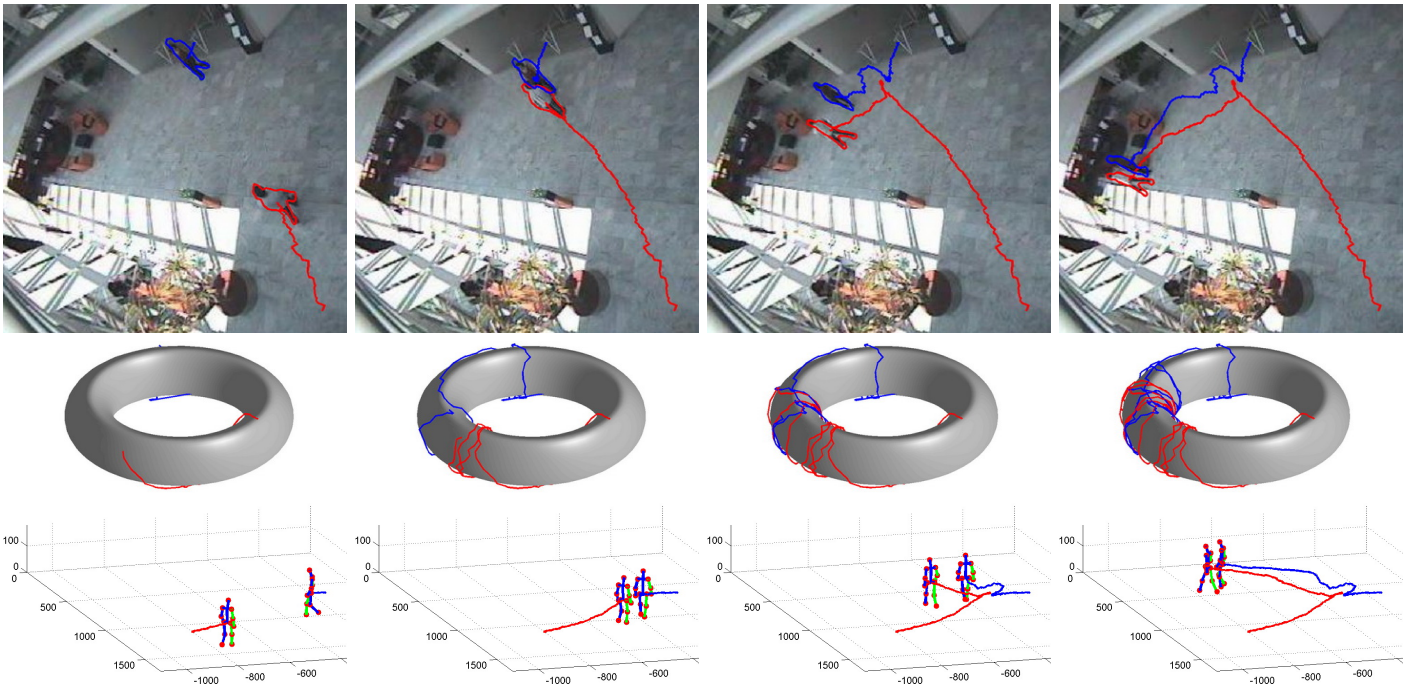


Fig. 13. Qualitative results for the *Meet_WalkTogether2* sequence (2 interacting subjects) using our view-invariant pose tracking framework and 1000 particles per target. For each presented frame (40, 120, 230 and 310), we show from top to bottom: tracked silhouettes and trajectories in the image, trajectories on the torus manifold and estimated 3D poses. Results for the entire sequence are presented in the attached video *Meet_WalkTogether2_processed.avi*.

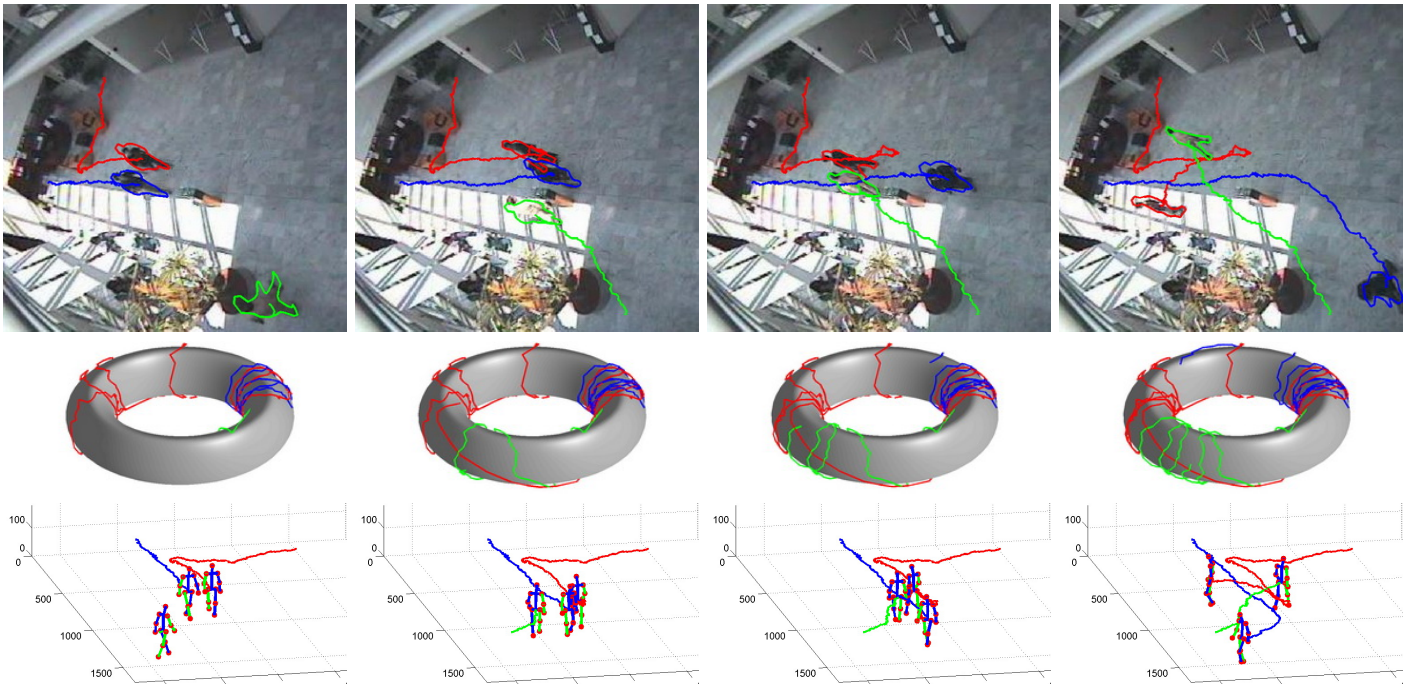


Fig. 14. Qualitative results for the *Meet_Split_3rdGuy* sequence (3 interacting subjects) using our view-invariant pose tracking framework and 1000 particles per target. For each presented frame (260, 330, 365 and 415), we show from top to bottom: tracked silhouettes and trajectories in the image, trajectories on the torus manifold and estimated 3D poses. Results for the entire sequence are presented in the attached video *Meet_Split_3rdGuy_processed.avi*.

APPENDIX

Following the classical notation of 3D projective geometry [57], a 3D point $[X, Y, Z, 1]^T$ is related to its 2D image projection $[u, v, 1]^T$ via a 3×4 projection matrix M:

$$[u, v, 1]^T = \mathbf{M} \cdot [X, Y, Z, 1]^T, \tag{27}$$

where points in the projective space $\mathbb{P}^2$ are expressed in homogeneous coordinates and " $=''$ means equality up to scale. The projective transformation matrix $\mathbf{M}$ can be determined with a series of intrinsic and extrinsic parameters or, as shown in [58], it can be defined as a function of the vanishing points of the dominant 3D directions.

Suppose we want to relate the image I with a vertical plane Π ($\Pi \perp \Pi_{\text{gd}}$), whose intersection with the ground plane Π_{gd} is $\mathbf{G}$. The plane Π is thus spanned by the vertical Z-axis and horizontal G-axis. In that sense, (27) becomes:

$$[u, v, 1]^T = \mathbf{H}_{I \leftarrow \Pi} \cdot [G, Z, 1]^T, \tag{28}$$

with G a coordinate on the G-axis and $\mathbf{H}_{I \leftarrow \Pi}$ a homography matrix that can be computed from the vanishing points of the dominant 3D directions of Π :

$$\mathbf{H}_{I \leftarrow \Pi} = [\mathbf{v}_G \quad \alpha \mathbf{v}_Z \quad \mathbf{o}]. \tag{29}$$

where $\mathbf{v}_Z$ is the vertical vanishing point, $\mathbf{o}$ is the origin of the world coordinate system and α is a scale factor. $\mathbf{v}_G$ is the horizontal vanishing point of plane Π in I i.e. the vanishing point along the horizontal direction $\mathbf{G}$ in image I. This vanishing point $\mathbf{v}_G$ can be localized as the intersection of line $\mathbf{g}$, the projection of $\mathbf{G}$ in the image I and $\mathbf{l}$, the horizontal vanishing line in I:

$$\mathbf{v}_G = \mathbf{l} \times \mathbf{g}, \tag{30}$$

where $\times$ represents the vector product, and $\mathbf{l}$ is the vanishing line of the ground plane (see [57] for details). Two examples of horizontal vanishing point localizations are given in Fig. 15.

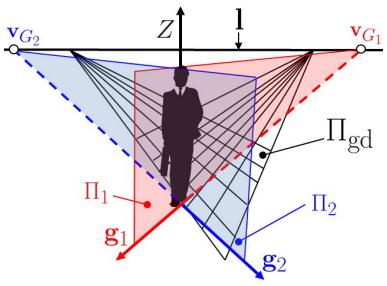


Fig. 15. Horizontal vanishing point localization for homography to vertical plane centered on the human body: 2 examples are given for 2 different directions $\mathbf{g}_1$ and $\mathbf{g}_2$ on the ground plane Π_{gd}. Π_1 is the vertical plane parallel to the real-world direction G_1 and Π_2 the one parallel to G_2. The vanishing points $\mathbf{v}_{G_1}$ and $\mathbf{v}_{G_2}$ are the intersection points of $\mathbf{g}_1$ and $\mathbf{g}_2$ with the horizon line $\mathbf{l}$, i.e. the vanishing line of the ground plane.

REFERENCES

[1] L. Wang, G. Zhao, N. Rajpoot, and M. S. Nixon, "Special issue on new advances in video-based gait analysis and applications: Challenges and solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 982–985, 2010.

[2] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition across various walking speeds using higher order shape configuration based on a differential composition model," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 6, pp. 1654–1668, 2012.

[3] J. Zhang, J. Pu, C. Chen, and R. Fleischer, "Low-resolution gait recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 986–996, 2010.

[4] Y. Ran, Q. Zheng, R. Chellappa, and T. M. Strat, "Applications of a simple characterization of human gait in surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 1009–1020, 2010.

[5] X. Zhang and G. Fan, "Dual gait generative models for human motion estimation from a single camera," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 1034–1049, 2010.

[6] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 997–1008, 2010.

[7] J. Gu, X. Ding, S. Wang, and Y. Wu, "Action and gait recognition from recovered 3-d human joints," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 1021–1033, 2010.

[8] T. Jaeggli, E. Koller-Meier, and L. J. V. Gool, "Learning generative models for multi-activity body pose estimation," *International Journal of Computer Vision*, vol. 83, no. 2, pp. 121–134, 2009.

[9] A. M. Elgammal and C.-S. Lee, "Tracking people on a torus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 520–538, 2009.

[10] Caviar dataset, "EC funded CAVIAR project IST 2001 37540," http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. 2004.

[11] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," 2001.

[12] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, 2005.

[13] C. Canton-Ferrer, J. R. Casas, and M. Pardás, "Human motion capture using scalable body models," *Computer Vision and Image Understanding*, vol. 115, no. 10, pp. 1363–1374, 2011.

[14] R. Li, T.-P. Tian, S. Sclaroff, and M.-H. Yang, "3d human motion tracking with a coordinated mixture of factor analyzers," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 170–190, 2010.

[15] I.-C. Chang and S.-Y. Lin, "3d human motion tracking based on a progressive particle filter," *Pattern Recognition*, vol. 43, no. 10, pp. 3621–3635, 2010.

[16] C.-S. Lee and A. M. Elgammal, "Coupled visual and kinematic manifold models for tracking," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 118–139, 2010.

[17] H. Ning, T. Tan, L. Wang, and H. W, "People tracking based on motion model and motion constraints with automatic initialization," *Pattern Recognition*, vol. 37, no. 7, pp. 1423–1440, 2004.

[18] R. Urtasun, D. Fleet, and P. Fua, "Temporal motion models for monocular and multiview 3d human body tracking," *Computer Vision and Image Understanding*, vol. 103, pp. 157–177, November 2006.

[19] R. Urtasun, D. J. Fleet, and P. Fua, "3d people tracking with gaussian process dynamical models," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 238–245.

[20] C. H. Ek, J. Rihan, P. H. S. Torr, G. Rogez, and N. D. Lawrence, "Ambiguity modeling in latent spaces," in *Proc. of the Intern. Workshop on Machine Learning for Multimodal Interaction*, 2008, pp. 62–73.

[21] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 623–630.

[22] B. Daubney, D. P. Gibson, and N. W. Campbell, "Estimating pose of articulated objects using low-level motion," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 330–346, 2012.

[23] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[24] R. Rosales and S. Sclaroff, "Combining generative and discriminative models in a framework for articulated pose estimation," *International Journal of Computer Vision*, vol. 67, no. 3, pp. 251–276, 2006.

[25] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr, "Randomized trees for human pose detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] J. Zhang, R. T. Collins, and Y. Liu, "Bayesian body localization using mixture of nonlinear shape models." in *Proc. of the IEEE International Conference on Computer Vision*, 2005, pp. 725–732.

[27] X. Lan and D. P. Huttenlocher, "A unified spatio-temporal articulated model for tracking." in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 722–729.

[28] G. Rogez, C. Orrite, and J. Martínez, "A spatio-temporal 2d-models framework for human pose recovery in monocular sequences," *Pattern Recognition*, vol. 41, no. 9, pp. 2926–2944, 2008.

[29] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

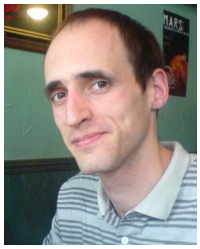
[30] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 1, pp. 13–24, 2010.

[31] A. Kale, A. K. R. Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," in *IEEE Conference on Advanced Video and Signal based Surveillance*, 2003, pp. 143–150.

[32] I. Bouchrika, M. Goffredo, J. N. Carter, and M. S. Nixon, "Covariate analysis for view-point independent gait recognition," in *Proc. of the*

International Conference on Advances in Biometrics, 2009, pp. 990–999.
[33] M. Goffredo, R. D. Seely, J. N. Carter, and M. S. Nixon, "Markerless view independent gait analysis with self-camera calibration," in *Proc. of the IEEE Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1–6.
[34] R. Rosales, M. Siddiqui, J. Alon, and S. Sclaroff, "Estimating 3d body pose using uncalibrated cameras," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 821–827, 2001.
[35] G. Rogez, J. J. Guerrero, J. Martínez, and C. Orrite, "Viewpoint independent human motion analysis in man-made environments," in *Proc. of the British Machine Vision Conference*, vol. 2, 2006, pp. 659–668.
[36] G. Rogez, J. J. Guerrero, and C. Orrite, "View-invariant human feature extraction for video-surveillance applications," in *Proc. of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 324–329.
[37] Y. Li, B. Wu, and R. Nevatia, "Human detection by searching in 3d space using camera and scene knowledge," in *Proc. of the International Conference on Pattern Recognition*, 2008, pp. 1–5.
[38] T. Riklin-Raviv, N. Kiryati, and N. A. Sochen, "Prior-based segmentation and shape registration in the presence of perspective distortion," *International Journal of Computer Vision*, vol. 72, no. 3, pp. 309–328, 2007.
[39] G. Rogez, "Advances in monocular exemplar-based human body pose analysis: Modeling, detection and tracking," Ph.D. dissertation, Universidad de Zaragoza, june 2012.
[40] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.
[41] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.
[42] G. Mori and J. Malik, "Recovering 3d human body configurations using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1052–1062, 2006.
[43] D. Cremers, "Dynamical statistical shape priors for level set-based tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1262–1273, 2006.
[44] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 604–618, 2010.
[45] A. Baumberg and D. Hogg, "Learning flexible models from image sequences." in *Proc. of the European Conference on Computer Vision*, 1994, pp. 299–308.
[46] N. T. Siebel and S. J. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," in *Proc. of the European Conference on Computer Vision (4)*, 2002, pp. 373–387.
[47] J. Giebel, D. Gavrila, and C. Schnörr, "A bayesian framework for multi-cue 3d object tracking," in *Proc. of the European Conference on Computer Vision (4)*, 2004, pp. 241–252.
[48] H. Sidenbladh, M. J. Black, and L. Sigal, "Implicit probabilistic models of human motion for synthesis and tracking." in *Proc. of the European Conference on Computer Vision (1)*, 2002, pp. 784–800.
[49] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1372–1384, 2006.
[50] M. Hofmann and D. Gavrila, "Multi-view 3d human pose estimation in complex environment," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 103–124, 2012.
[51] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects." *International Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.
[52] K. Smith, D. Gatica-Perez, and J.-M. Odobez, "Using particles to track varying numbers of interacting people," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 962–969.
[53] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *Proc. of the IEEE Internation Conference on Computer Vision*, 2001, pp. 34–41.
[54] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1208–1221, Sept. 2004.
[55] R. Okada and S. Soatto, "Relevant feature selection for human pose estimation and localization in cluttered images," in *Proc. of the European Conference on Computer Vision*, 2008, pp. 434–445.
[56] K. Toyama and A. Blake, "Probabilistic tracking with exemplars in a metric space," *International Journal of Computer Vision*, vol. 48, no. 1, pp. 9–19, 2002.
[57] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.
[58] A. Criminisi, I. D. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

Grégory Rogez is a Marie Curie fellow and a visiting scientist at the University of California, Irvine. He graduated from the Ecole Nationale Superieure de Physique de Marseille with a M.Eng. in physics in 2002 and received a M.Sc. in biomedical engineering and a Ph.D. from the University of Zaragoza in 2005 and 2012, respectively. His work on monocular human body pose analysis received the best Ph.D. thesis award 2011-2012 from the Spanish Association for Pattern Recognition and Image Analysis (AERFAI). Since 2002, he has been affiliated with the Aragon Institute of Engineering Research. His current research interests include computer vision and machine learning, with special focus on human activity analysis, environment understanding and egocentric vision.



Jonathan Rihan received the M.Sc. and Ph.D. degrees in computer science from Oxford Brookes University, Oxford, UK, in 2005 and 2011, respectively. In 2011, he was a Post-doctoral Research Assistant in applied computing at Oxford Brookes University. Since February 2012, he is a Senior Software Developer in Computer Graphics at ANSYS UK, Ltd. He is interested in applied computer vision and machine learning research, motion capture, computer graphics, and computer game development.

J.J. Guerrero received the M.S. degree in electrical engineering and the Ph.D. degree from the Universidad de Zaragoza, Zaragoza, Spain, in 1989 and 1996, respectively. He is currently an Associate Professor and the Deputy Director of the Department of Informatics and Systems Engineering, Universidad de Zaragoza. His research interests are in the area of computer vision, particularly in 3-D visual perception, photogrammetry, visual control, robotics, and vision-based navigation.

Carlos Orrite received the M.Eng. degree in industrial engineering, the M.Sc. degree in biomedical engineering and the Ph.D. degree in computer vision from University of Zaragoza, Zaragoza, Spain, in 1989, 1994, and 1997, respectively. He is currently an Associate Professor with the Department of Electronics and Communications Engineering, University of Zaragoza, Spain and Associate Director of the Aragon Institute of Engineering Research (I3A), coordinator of the Information & Communications Technologies Division. He serves as secretary of the Spanish Association for Pattern Recognition and Image Analysis (ARFAI). His research interests include computer vision and human-machine interface.