Tracking Human Position and Lower Body Parts Using Kalman and Particle Filters Constrained by Human Biomechanics

Jesus Martínez del Rincón, Dimitrios Makris, *Member, IEEE,* Carlos Orrite, *Member, IEEE* and Jean-Christophe Nebel, *Senior Member, IEEE,*

***Abstract*—In this paper, a novel framework for visual tracking of human body parts is introduced. The presented approach demonstrates the feasibility of recovering human poses with data from a single uncalibrated camera using a limb tracking system based on a 2D articulated model and a double tracking strategy. Its key contribution is that the 2D model is only constrained by biomechanical knowledge about human bipedal motion, instead of relying on constraints linked to a specific activity or camera view. These characteristics make our approach suitable for real visual surveillance applications. Experiments on a set of indoor and outdoor sequences demonstrate the effectiveness of our method on tracking human lower body parts. Moreover, a detail comparison with current tracking methods is presented.**

***Index Terms*—human pose, particle filter, biomechanics, 2D articulated model, bipedal motion, video surveillance.**

I. Introduction

HUMAN motion modelling is one of the most active areas in computer vision. It can be defined as the ability to estimate, at each frame of a video sequence, the position of each joint of a human figure which is represented by an articulated model. Because of the 3D nature of human motion, tracking methods based on 3D anthropomorphic articulated models have proved to be the most effective [14], [15], [16], [17], [18]. Their applications include analysis of human activity [55], entertainment, ambient intelligence and medical diagnosis to name a few. However, their main drawback is they generally rely on data capture synchronously by several cameras which have been accurately calibrated. Therefore, these techniques are unpractical for applications targeting unconstrained environments such as video surveillance [7], [6]. The alternative is usage of tracking methods based on 2D models which cannot deal by themselves with the intrinsic ambiguity of projected 3D postures, self occlusions and distortions introduced by camera perspective. Therefore, they are usually restricted to well defined motions and specific camera views; however these constraints reduce their value in many real applications.

We propose a "double tracking" strategy to accurately track simultaneously both the position of the body and its articulated motion. Position is tracked by a Kalman filter, while tracking of human body parts is achieved using a set of particle filters [19], [14], [57], which iteratively refine their solution. The key contribution of this method is that it relies on a generative approach based on a 2D model constrained only by human biomechanics. The inclusion of biomechanical knowledge about bipedal motion significantly reduces the complexity of the problem. This is achieved by the detection of the pivot foot - i.e. the foot which is static during a step - and its trajectory during a whole step.

In this work, we concentrate our effort on tracking the legs of a subject since the other body parts do not benefit from biomechanics constraints. Our results are evaluated against the HumanEva data set, which is becoming the standard for assessing human body tracking algorithms [5], and outdoor data from Sidenbladh [46]. After a brief description of the state of the art in human body part tracking, we present an overview of our methodology. Then we detail the key algorithms and the biomechanics constraints we use. Finally, after presentation and evaluation of our results, conclusions are drawn.

II. Related work

Tracking complexity increases exponentially with the number of targets when their motion is not independent from each other as it is the case when dealing with articulated objects. Articulated models have been shown to be essential tools to handle tracking and detection tasks by reinforcing motion constraints in either the 2D [43] or 3D space [13] so that motions of subparts are interrelated. Several approaches have been investigated to alleviate this challenge, such as dynamic programming [1], annealed sampling [20], partitioned sampling [17], eigenspace tracking [42], hybrid Monte Carlo filtering [21] and bottom-up [8] approaches.

Approaches to vision-based human motion analysis can broadly be divided into generative and discriminative. The first category explicitly uses a human body model [24], [25], [26], [27], [28], [29], [30], [20], [32] that describes both visual and kinematic properties of the human body. Discriminative approaches [34], [31], [33], [35], [36], [37], [38], [39], [40], [41], [42] learn the mapping from image space to pose space directly from carefully selected training data. Since discriminative approaches work in a learned pose space where the dimensionality has been reduced, they are computationally much less expensive, can potentially be applied in real-time and are more robust to noise or occlusions. Furthermore, discriminative approaches allow the recovery of poses with less information

J.C. Nebel, D. Makris and J. Martínez are with the Digital Imaging Research Centre, Kingston University, KT1 2EE, UK. email:{J.Nebel, D.Makris, Jesus.Martinezdelrincon,}@kingston.ac.uk .

C. Orrite is with the Aragon Institute of Engineering Research, University of Zaragoza, 50001, Spain. e-mail: corrite@unizar.es.

which make them more suitable for monocular application. However, they have a serious drawback: their accuracy relies on the similarity between the posture to recover and data used in the training data set. In addition, their performance tends to decrease when the variety of activities used in a training set increases [56].

Independently of the chosen modelling strategy, another key decision has to be taken regarding the dimension of the body model, i.e. 2D or 3D [29]. The first option involves working directly with the 2D features derived from the images. This has been successfully applied for constrained types of movements, such as walking parallel to the plane of the image and periodic motions. Nevertheless, their performance decreases significantly for unconstrained and complex human actions which include movements out of the camera plane (e.g. wandering, making gestures and turning) which produce frequent self-occlusions. Generally, 2D discriminative approaches are more robust when dealing with self-occlusions. However, prior knowledge about either the movement or the viewpoint is required to drive correctly their 2D models.

Many techniques based on 2D models have been proposed. In [1], an approach that analyses subparts locally is proposed for visual tracking of articulated models while reinforcing the structural constraints between different subparts. It combines a dynamic Markov network, which characterises the dynamics and the image observations of each individual subpart and motion constraints based on a Mean Field Monte Carlo (MFMC) in which a set of low dimensional particle filters interact with each other and solve the high dimensional problem collaboratively. Ju et al. [43] propose a cardboard model in which the human limbs are modelled by a set of connected planar patches. By constraining the parameterised motion of the patches in the image, the articulated motion is reinforced. Optical flow is used as feature to track the limbs as well as to estimate the viewpoint. Results confirm that 2D patches are able to track a limb not subject to occlusions if the viewpoint has been determined. Rehg et al. [44] describe a two-dimensional scaled prismatic model (SPM) for figure registration, which deals with variations in rotation and depth. SPM reduces significantly the number of singularities that appear due to the bidimensional projection of the 3D pose and does not require detailed knowledge of the 3D kinematics. Although they demonstrate the application of the model for motion capture from movies, only certain types of movements can be tracked and the system fails for fast movements. In [2], Random Sample Consensus (RANSAC) and Maximum Likelihood Estimation Sample Consensus (MLESAC) algorithms are incorporated in a planar patch tracker like feature weights to perform robust tracking. In [3], Noriega and Bernier propose a planar-patch articulated model, which is a loose limbed model, including attraction potentials between adjacent limbs and constraints to reject poses resulting in collisions. Compatibility between model and image is estimated using one particle filter per limb, while compatibility between limbs is represented by interaction potentials. The joint probability is obtained by belief propagation on a factor graph. The main drawback of all these 2D models is their usage is restricted to specific types of motions which are usually linear and seen from a specific viewpoint.

On the other hand, 3D methods [14], [15], [16], [17], [18] can be considered as more general purpose approaches since they provide a well-pose solution to tracking a 3D object. In particular, this enables taking advantage of a large amount of available prior knowledge about the kinematics, shape properties and biomechanics of human body and gait. This information makes the problem more tractable and permits to predict events such as self-occlusions. However, the fact that 3D models must be projected into the image plane has two consequences: first, in addition to the larger dimensionality of the model, projections make 3D tracking a computationally expensive methodology. Secondly, a constrained environment is required: cameras have to be calibrated and the transformation between the image plane and the 3D world has to be known. Consequently, they are not suitable for applications like video surveillance, where real time tracking is expected and camera calibration is not practical.

Kaadiaris and Metaxas [30] consider a multi camera system to cope with 3D model-based body part tracking. Kalman filter is applied to predict the location of each limb. The correspondence between the contour in the image and the projection of the 3D shape is used as likelihood function. Gavrila and Davis [29] extended this methodology to a 22 degree-of-freedom model. Hunter et al. [45] build a model composed of 5 ellipsoids with 14 degrees of freedom where a particle filter was successfully combined with a 3D articulated model. Probably one of the most important papers in this field is the one presented by Deutscher, Blake and Reid [20]. It is not only the most important generative approach but also the method of reference used to benchmark new algorithms. They propose a modified version of particle filter to estimate efficiently the multi-modal distribution of the human body articulated model in a huge dimensional space. The main drawback is the prohibitive computational cost associated with the processing of each frame. Sidenbladh, Black and Fleet [46] present another relevant probabilistic method for tracking 3D articulated human figures in monocular sequences. It is based on a generative model of appearance, a robust likelihood function which works out gray level differences, and a prior probability distribution which introduces knowledge about human gait and joint angles. Moreover, valid 3D human motions are constrained by prior probability distribution over the dynamics of the human body.

Recently, discriminative approaches based on latent spaces and manifolds have achieved a high popularity [53], [54], [47], [41]. This is mainly because they reduce the computational cost by constraining the space of possible poses with prior information. Elgammal [31] proposes a manifold to relate silhouettes with 3D poses. A different 1D manifold is learned per view and activity. In [40], two different regression algorithms are used for the forward mapping (dimensionality reduction) and inverse mapping. The representatives used in the regression are chosen in a heuristic manner. In [39], GPLVM and a second order Markov model are used for tracking applications. The learned GPLVM model is used to provide model prior. Tracking is then done by minimising a cost of 2D image matching, with the negative log-likelihood of the model prior

as the regularisation term. Both [40] and [39] advocate the use of gradient descent optimisation techniques; hence, the low-dimensional space learned has to be smooth and accurate initialisation is required for the success of such techniques. An alternative approach [47] employs the GPLVM in a modified particle filtering algorithm where samples are drawn from the low-dimensional latent space. The smoothness enforced in the low-dimensional space by the learning algorithms in these three papers works well for tracking small limb movements, but may fail when large movements occur over time.

This overview of human pose recovery methodology informs us about the design of a solution on the basis of the requirements and characteristics of the particular problem that we want to solve. Since our objective consists in recovering the human pose in unconstrained environments, where the subject can perform any kind of movement and where initialisation should eventually be automated, we are constrained to choose a generative approach based on a 2D model. However, unlike the previous work within this framework which was limited to specific types of motions, we propose an approach able to deal with variations in rotation and depth so that it can be applied to real life data. This is achieved by constraining the 2D model, which is designed to tackle 3D motion patterns such as changes in the pose of the object with respect to the camera, by using specific knowledge about human biomechanics and gait analysis.

III. Double tracking strategy

A. General Principle

In previous work [48], we proposed a methodology where the global location of the person, as well as the relative pose of the limbs, were tracked simultaneously. Although this integrated strategy was elegant, it showed some inefficiency since an error of global location affects directly the process of limb pose recovery.

To deal with this problem, we propose a double tracking strategy (see Figure 1). The estimation of the pose of the limbs X_{leg} is calculated using the combination of two trackers: one tracks the global location of the person X^{ext}, and the second one recovers the relative pose of the limbs X'_{leg}.

$$X_{leg} = X^{ext} + X'_{leg} \tag{1}$$

As first tracker, we use a Kalman filter which has been shown as a very efficient paradigm to track pedestrians in visual surveillance applications [52]. For body part tracking, we use a set of particle filters.

Human articulated motion is highly multi-modal and this non-Gaussian characteristic is amplified in the image plane by the camera perspective. Therefore, a tracking framework capable of working with non-linear distribution is required. Since particle filters have been successfully applied for this purpose [19], this algorithm is at the core of our tracking framework. A detailed explanation about particle filtering is shown in [19]

Once the first tracker has obtained an estimation of the location of the person using a motion blob, this information is introduced as prior knowledge in the proposal distribution

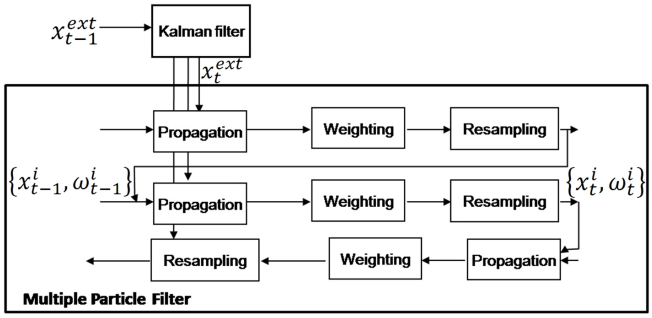


Fig. 1. Principle of double tracking strategy

of the particle filter. Thus, the particle distribution in the next prediction step is guided by the global location in the (x,y)-coordinates $[x^{ext}, y^{ext}]$. Moreover, limb sizes of the new hypotheses are estimated by taking into account blob height changes between two frames l^{ext}. In this manner, tracking can recover from incorrect estimations from the particle filters without being limited to the result of the first tracker. Indeed the dynamic model of the new hypotheses gives the ability to correct the first tracker estimation, which is only used as a guide or "soft" constraint that helps to put the hypotheses near to the global optimum.

The limb model employed is identical to the one proposed in [48], but here the spatial coordinates have been normalised with respect to the central point of the line which links both hip points, and the size parameters have been normalised regarding the human height of the blob.

Limb tracking is based on a set of particle filters to fit a 2D articulated model on each frame of a video sequence. In addition, we take advantage of a biomechanics constraint inherent in human bipedal motion: during a 'step', one leg pivots around a single point. This allows us dealing with many more motions than other techniques which rely on training on a specific activity. Since we are able to detect the position of this point, this constraint is integrated in an asymmetrical 2D model where the two legs are treated differently. Finally, model fitting is performed after different trackers have been applied successively.

Initially, a 'standard' particle filter process operates to track lower limb locations until the end of the 'step'. Due to the high dimensionality of the problem and the ill-conditioned model, it may not be able to produce satisfactorily tracking. In order to refine the tracking of the articulated model, two assistant particle filters are then launched in parallel using information intrinsic to the 'step' of interest. The main reason for using two trackers instead of one is to handle the degradation and potential divergence of tracking over time.

To take advantage of the 'pivot' point constraint and trajectory information, we propose to rely on data captured during a full 'step' before completing the tracking task. While a short delay is introduced - typically around 10 frames (i.e. 0.5s) - in a real time system, this allows processing a wide range of human activities without loss of accuracy. Moreover, since this delay does not increase, if suitable processing power is available, the whole system can operate on-line.

Although some actions, such as running or jumping, break the 'pivot' constraint during short periods of time and the 'pivot' point can be momentarily occluded, this can be detected and handled without affecting significantly the proposed tracking framework, since the 'standard' particle filter is still able to estimate the poses without those constraints.

B. Biomechanics constraints for human motion tracking

Most human motion tracking methods rely on constraints such as specific activity, constant velocity, linear or periodic motion which critically impact on their accuracy and/or their genericity. Study of human biomechanics, however, reveals that human motion itself provides some explicit constraints. In this section, we show they can be utilised to simplify the task of tracking human body parts. Walking is a very common human activity whose many other motions, such as loitering, balancing and dancing, can be seen as derivatives and where the underlying mechanics of walking can be applied. All these bipedal motions are based on a series of 'steps' defined as one leg 'swinging' around a 'support' leg whose foot, or 'pivot', stays in contact with the ground at any instant [22]. Therefore, the detection of this pivot point from a video sequence provides a very important biomechanics cue which is present in most motion processed by the tracker.

Knowledge of the precise position of the pivot foot also allows using different strategies for tracking either the 'support' or the 'swinging' leg, which enhances the power of our 2D model. Moreover, positions of consecutive striking feet provide some information about the subject's trajectory in the image plane which supplies clues regarding the relative camera-subject position. Consequently, detection of this permits a significant reduction of the complexity of the tracking task.

In addition to the 'pivot' foot constraint, the 'support' leg has another property: upper and lower legs are supposed to be aligned during the pivot motion around the static foot. Therefore, estimate of the locations of the associated knee and hip can be refined if they do not form a straight line with the pivot foot.

In our framework, the static foot is detected using the algorithm proposed in [4]. It is based on the biomechanics of gait motion. During the strike phase, the foot of the striking leg stays at the same position for half a gait cycle, whilst the rest of the human body moves. The pivot foot is detected using a low-level feature: corners produced by the Harris corner detector. Outliers due to cluttered backgrounds are filtered out by using a background subtraction algorithm. Corners associated to the pedestrian of interest are accumulated across several frames (i.e. 20 in our implementation). The region where the leg strikes the ground must have a high density of corners. Although this approach is usually efficient (when an individual motion is parallel to the camera plane, the static foot is detected easily), motions towards or away from the camera produce many points seen as static on the body due to the influence of the perspective. We deal with this by removing outliers and false positive by maintaining both temporal and spatial coherences of the 'pivot' point.

Corners, C, are accumulated across several frames using equation (2):

$$C = \sum_{t=1}^{N} (H(I_t) \wedge L_t) \tag{2}$$

where H is the output of the Harris corner detector, I_t is the original image at frame t, L_t is the pedestrian blob at frame t and $\wedge$ is the logical conjunction operator. Although we only consider one pedestrian, as commented in the introduction of this chapter, the pivot point detection algorithm could be extended to deal with multiple people by selecting an appropriate association algorithm.

Dense areas of corners are located using a measure for density of proximity, d_p. The value of proximity at point p depends on the number of corners within the region R_p and their corresponding distances from p. R_p is assumed to be a circular area centred in p, whose radius, r, is determined as the ratio of total image points to the total of corners in C. Corner proximity values, d_p, are computed for all regions R_p in C using equation (3).

$$\begin{cases} d_p^r = \frac{N_r}{r} \\ d_p^i = d_p^{i+1} \frac{N_i}{i} \end{cases} \tag{3}$$

where d_p^i is the proximity value for rings of radius i away from the centre p, and N_i is the number of corners at the distance i from the centre, rings are single pixel wide.

Starting from a radius r, the process then iterates to accumulate all the densities for the subregions R_p for all points p into a matrix to produce the corner proximity matrix of the frame. Highest values in the matrix generally correspond to the heel strike areas.

C. Position tracking based on Kalman filter

Using a Kalman filter we track the bounding box of the person under observation. The state vector is $x_t = [x^{ext}, y^{ext}, \dot{x}^{ext}, \dot{y}^{ext}, l^{ext}, \dot{l}^{ext}]$, where $[x^{ext}, y^{ext}]$ is the global location in the (x,y)-coordinates, l^{ext} is the blob height and $\dot{x}, \dot{y}, \dot{l}$ their derivatives. The likelihood function is based on a motion detector that extracts the blob corresponding to the subject.

D. Multiple particle filter tracking based on 2D articulated model

1) 2D asymmetrical articulated model informed by trajectory information: Our model aims to track simultaneously the relative positions of the different parts of the limbs. Thus, the tracker state vector is composed of the image coordinates of the hip points and the parameters which model the relative motions and positions, such as angles and lengths in the image plane. In order to introduce the biomechanics constraints, which rely on a relative independence between both legs, both hip points are employed as references and the angles of both legs with respect to the hips are included in the state vector. The state vector of each leg is described by the following equation:

$$X'_{leg} = [x_{hip}, y_{hip}, \dot{x}_{hip}, \dot{y}_{hip}, \theta_{hip-thigh}, \theta_{knee}, \dot{\theta}_{hip-thigh}, \dot{\theta}_{knee}, l_{femur}, l_{shin}, \dot{l}_{femur}, \dot{l}_{shin}] \quad (4)$$

where x and y are the coordinates in pixels, θ is the angle between a limb and the x axis and l is the length of the limb (see figure 2).

By including the global location of the person provided by the Kalman Filter, we obtain our articulated model state in absolute image coordinates substituting the Eq. 4 in Eq. 1:

$$X_{leg} = [x^{ext} + x_{hip}, y^{ext} + y_{hip}, \dot{x}_{hip}, \dot{y}_{hip}, \theta_{hip-thigh}, \theta_{knee}, \dot{\theta}_{hip-thigh}, \dot{\theta}_{knee}, l_{femur} \cdot l^{ext}, l_{shin} \cdot l^{ext}, \dot{l}_{femur}, \dot{l}_{shin}] \quad (5)$$

Using the pivot point as a constraint, the "support" leg is first estimated. Then, the 'swinging' leg is calculated. To perform a robust estimation, the hip point position of the "support" leg is used to constrain the other hip point. The distance between the two hips is set at a fixed anthropometric value D_0 during initialisation as a proportion of the width of the legs. Moreover, we assume the two hips points share the same y coordinate where the y-axis is defined as the axis which goes along the dorsal spine of the subject. This assumption is reasonable if the camera is sufficiently far away from the subject and does not provide a zenithal view, which is usually the case in visual surveillance. Although generally this y axis corresponds to the vertical axis of the image, its direction can be determined more precisely by calculating the momenta of the human figure where the y-axis is the larger axis of the ellipse that surrounds the subject.

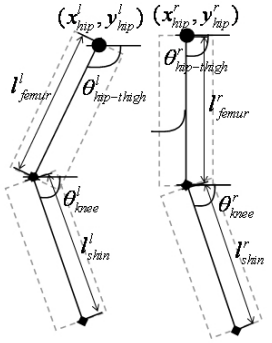


Fig. 2. 2D articulated model.

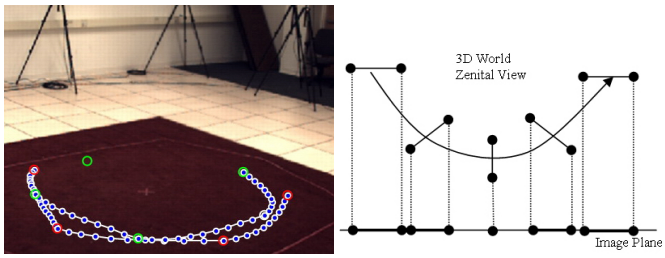
Due to its nature, 2D tracking allows a higher flexibility and simplicity of use and initialisation than 3D tracking. However, in 2D it is not possible to introduce traditional constraints, such as motion dynamic or kinematics. Instead, we transfer 3D properties to the 2D world. In the 3D world, the distance between the hips remains constant over time. However, when this fixed distance is projected in the camera plane, its value is changed by two different parameters: the location and the orientation. Whereas the location introduces a factor of scale which is estimated with the global size of the legs, the orientation distorts this distance in a non-linear way which depends on the view point.

Because of the stochastic nature of our tracking algorithm, the exact value of this distance is not required. Given the poses of the hips at the beginning and the end of a step, values of the hips between these two frames are estimated. In fact, the distance is correlated to the angle of the step trajectory in a non-linear manner as shown on Figure 3a. We approximate this correlation function using a function which models a S-curve.

$$D(\theta) = D_0 \cdot \frac{1 - e^{-\alpha\theta}}{1 + e^{-\alpha\theta}} \quad (6)$$

where D_0 is the maximum size of the hip distance with respect to the size of the leg (in our implementation, it is half the value of the sum of the thigh widths), θ is the angle between the trajectory and the x axis in the image plane and α is an empirical factor which controls the speed of the curve descent.

Therefore, hip distance is estimated at each frame based on the trajectory angle. This is performed by fitting cubic splines to all pivot points (see figure 3a,b).



(a) (b)

Fig. 3. (a) Interpolated trajectory (blue dots) of the pivot points (red and green dots). (b) Correspondences during a turn between the hip distances in a zenithal view and in the image plane.

At the end of the step, the new pose of the 'swinging' leg is known, i.e. positions of hip, knee and ankle. Therefore, sizes of both limbs at the beginning and end of the step are available. Their values are used to constrain limb size parameters during the whole step.

Consequently, our model allows introducing two gait constraints which help both loop and back tracking processes to improve the results of the 'swing leg': the size and the hip distance constraints. Since this information is known a posteriori, it can only be applied to the auxiliary tracking process.

2) Multiple particle filter tracking: One of the most challenging problems of 2D tracking is to deal with the perspective effect which amplifies changes in trajectories and, therefore, can create major variations in the target's size. Therefore, the usage of a simple first order model does not allow representing size dynamics adequately. Since our tracking framework is based on a full 'step' where heel strike positions are known, the final position of a step is partially reinitialised. Consequently, information is available to define the trajectory of the target during each step. Moreover, new tracking constraints are derived regarding maximum and minimum apparent limb sizes and distances between the hip points during the step.

This last constraint provides a reference point for the "swing" leg similarly as the pivot point restricts the location

of the 'support' leg. These new constraints, which were not initially available when the standard tracker operated, reduce significantly the complexity of the tracking problem. Furthermore, when using a particle filter based tracker, the probability of divergence increases after each prediction: the closer a frame is to the initialisation frame, the more accurate the estimation is likely to be.

In order to take advantage of these new constraints and tackle this inherent tracker weakness, we propose that once the standard tracker has processed a full "step", two new trackers are launched in parallel. These trackers have the same configuration and dynamical models enhanced by the constraints extracted from the output of the standard tracker. Whereas the 'forward' tracker starts from the first frame of the step, the 'backward' tracker begins at the last frame and tracks targets backwards.

Our algorithm is described in Figure 4. The observation process provides the information to the three concurrent particle filters. At time t, the standard particle filter (blue segment), makes a first estimation p_t^1 of the pose x_t during the current step. Simultaneously, the loop particle filter, or "forward" tracking (orange segment), refines the estimation of the previous step poses p_t^2 by means of the introduction of constraints, i.e. the size and the trajectory, that are only known once the step is completed. In parallel, a new back particle filter, or"backwards" tracking (green segment), also reestimates the same previous step p_t^3 but processing from the last pose to the first. When forward and backwards trackers meet, a decision should be done to decide which one is likely to refine the most the initial tracking. The winner tracker will continue its tracking until reaching a time step when its contribution is expected to be worse than the opposite tracker. At that point, the pose estimation p_t of the whole step is accepted as definitive.

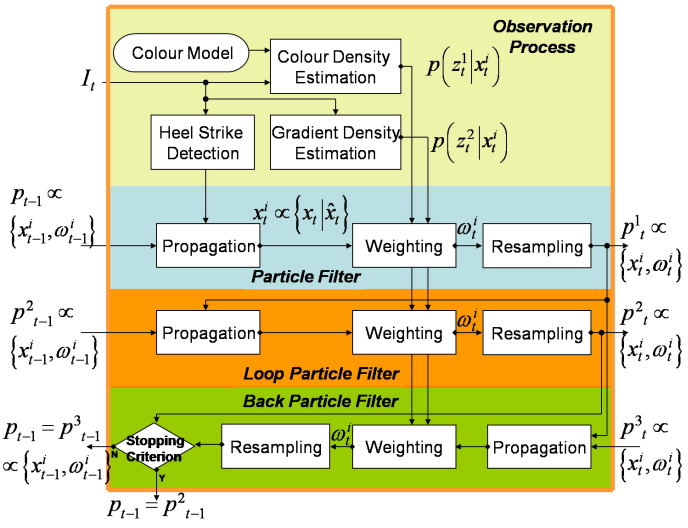


Fig. 4. Multiple particle filter framework.

A comparison criterion has been designed to decide at which frame the backward tracker is more likely to provide more accurate estimates than the forward tracker. First, although the particle filter does not provide an actual estimation for each frame, a weighted mean estimation is extracted combining all the hypotheses. By using this temporal estimation, the measurement of its likelihood function is obtained. Secondly, we express the intuitive idea that a tracker's reliability decreases after initialisation by introducing an exponential decreasing function that multiplies the estimation likelihood (see Figure 5). Therefore, estimation at a given time step about the most accurate tracker can be made according to the sign of the following quantity, q:

$$q = p(z_t|E[x_t^b]) \cdot f_{rel}(T-t) - p(z_t|E[x_t^f]) \cdot f_{rel}(t) \quad (7)$$

where $p(z_t|E[x_t^b])$ and $p(z_t|E[x_t^f])$ are respectively the measurement values (based on colour and edge likelihoods) of the backward and forward trackers, T is the length of the step in frames, x_t^b and x_t^f are the state vectors of the backward and forward tracking respectively, and f_{rel} is the decreasing function that takes into account the reduction of reliability over time

$$f_{rel}(t) = e^{-\beta(1-t)} \quad (8)$$

where β is an empirical factor which models accuracy degradation. It is set to 1 for walking sequences.

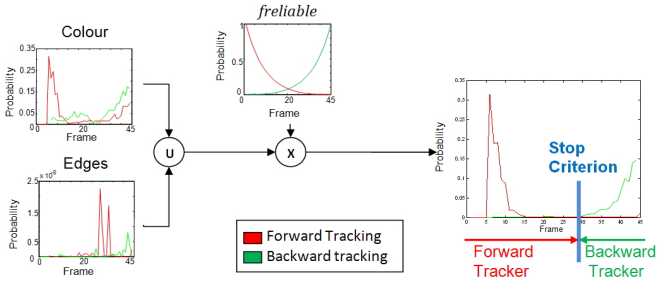


Fig. 5. Stopping criterion established to stop the auxiliary trackers by combining edge and colour likelihoods.

3) Predictive motion model and likelihood function of particle filters: We use simple first order dynamic models to track location, size and angular parameters since they are sufficiently accurate for modelling motion between successive frames during a single step and motion non-linearities is taken care of by the biomechanics constraints previously presented, i.e. the hip distance $D(\theta)$ and the size constraint in the auxiliary tracker.

We use a simple constant-acceleration dynamic model $X_{leg}^t = F \cdot X_{leg}^{t-1}$. We can express F as the dynamic matrix:

$$F = \begin{bmatrix} 1 & 0 & dt & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & dt & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & dt & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & dt & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & dt & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & dt \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (9)$$

where dt is the time lapse between two frames.

An adequate likelihood function must be applied to track the targets. In order to weigh each hypothesis, several visual features are combined, i.e. colour and edges (see Figure 6).

Colour is a discriminative feature which differentiates between object and background, but also between objects. Moreover, it is pose invariant. Edges also provide a good visual feature due to the continuity of the human limbs. Because of their invariance to colour, lighting and pose, they are especially useful to deal with self-occlusions between limbs [46].

Since we assume these features are independent from each other, we can combine them to obtain the observation probability:

$$p(z_t|x_t) = p(z_t^1|x_t) \cdot p(z_t^2|x_t) \tag{10}$$

where z_t^1 and z_t^2 are the colour and edge observations respectively and x is the state vector, $x_t = [X_{leg}^{left}, X_{leg}^{right}]$.

Colour features are obtained by sampling each region by a grid and expressing the colour information by RGB values subsampled to 4 bits per channel to filter out noise and small variations. The colour density is measured by comparing the colour feature of each region of the articulated model with its corresponding colour model. It is evaluated by estimating the Bhattacharyya coefficient between their histograms.

$$p(z_t^1|x_t) = \prod_{\forall r \in R(x_t)} (\sum_{h=1}^{H} \sqrt{s(h) \cdot q(h)})^{\alpha_c} \tag{11}$$

where r is each body part belonging to the set R of regions from the articulated model x_t, H are all the histogram bins, $s(h)$ is the current histogram, $q(h)$ is the reference model and $\alpha_c > 0$ is an empirical factor to strengthen the discriminative power of the feature.

A gradient detector is used to detect edges, and the result is thresholded to eliminate spurious edges. The Canny algorithm is applied for this purpose. The result is smoothed with a Gaussian filter and normalised between 0 and 1. The resulting density image P^e assigns a value to each pixel according to its proximity to an edge using the Euclidean distance transform.

$$p(z_t^2|x_t) = \prod_{\forall r \in R(x_t)} \frac{1}{N} \sum_{i=1}^{N} P^e(I_t, i)^{\alpha_e} \tag{12}$$

where I_t is the original image in RGB, r represents each of the regions which compose the articulated model, N are all the pixels which compose the region and $\alpha_e > 0$ is an empirical factor similar to α_c. By default, both factors are assigned the same value. However, their weight can be adjusted to bias the probability density function towards the feature which is believed to be the most informative in a given scene.

IV. Results

Our approach was evaluated over data sets which have been produced as benchmarks to the scientific community to evaluate and compare different tracking and pose recovery methodologies. First, we have used the HumanEva (HE) data sets I and II, where motion capture and video data were collected synchronously [5]. Since cameras are calibrated, motion capture data provides not only the groundtruth for 3D pose recovery, but also for 2D pose recovery by projection on the 2D sequences. A standard set of error metrics is also defined for evaluation of both pose estimations and tracking algorithms. Secondly, we have tested our solution with a well-known outdoor sequence [46], where groundtruth was obtained by annotating carefully the location of the limbs by hand.

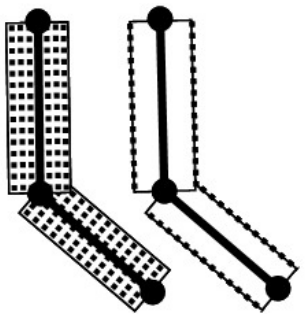


Fig. 6. Configurations of the pixel map sampling points for the colour and the edge measurements. The sampling points for colour measurements are defined by a grid sampling these regions, whereas the edge measurements are located along the contours of the regions which compose the articulated model.

A. Test sets and evaluation metrics

Our algorithms were tested with 3 indoor sequences from HumanEva data sets, i.e. *S2_Walking_1_C1*, *S2_Combo_2_C1* from HE I and *S2_Combo_1_C1* from HE II, and the outdoor Sidenbladh sequence. Since the *S2_Combo_1_C1* from HE II sequence is especially long, we divided it in two parts: part 1, which is at the beginning of the sequence and corresponds to walking, and part 2, which is at the end and shows some balancing, see Table I for details. These sequences were chosen to include a variety of movements (walking a complete circle and balancing) seen in indoor and outdoor environments from different points of view and happening mainly outside the camera plane (see Fig. 9, 10 and 13).

Since the pose of a human body can be represented using M virtual markers, the state of the body can be written as $X = x_1, x_2, \dots x_M$, where $x_m \in \Re_2$ (2D body model is used) is the position of the marker m in the image. The error between the estimated pose $\hat{X}$ and the ground truth pose X can then be expressed as the average absolute distance between individual markers. To ensure fair comparison between algorithms using different numbers of parts, a binary selection variable per-marker $\hat{\Delta} = \hat{\delta}_1, \hat{\delta}_2, \dots \hat{\delta}_M$ was added [5]. Therefore, the final proposed error metric is:

$$D(X, \hat{X}, \hat{\Delta}) = \sum_{m=1}^{M} \frac{\hat{\delta_m} \|x_m - \hat{x}_m\|}{\sum_{i=1}^{M} \hat{\delta}_i} \tag{13}$$

where $\hat{\delta}_m = 1$ if the evaluated algorithm is able to recover marker m, and 0 otherwise.

For a sequence of T frames we can compute the average performance, ν_{seq}, and the standard deviation of the performance, σ_{seq}, using the following equations:

$$\nu_{seq} = \frac{1}{T} \sum_{t=1}^{T} D(X_t, \hat{X}_t, \hat{\Delta}_t) \tag{14}$$

$$\sigma_{seq} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \Big[D(X_t, \hat{X}_t, \hat{\Delta}_t) - \nu_{seq} \Big]^2} \tag{15}$$

B. Experimental results and discussion

We report experiments conducted first with the HE sequences and then with the outdoor data. Although experiments were performed with a number of particles in the Particle Filter ranging from 200 to 500, their number did not affect tracking accuracy. Since the pivot point detector can produce erroneous locations - an average error of 20 pixels was measured for the HE sequences -, this affects negatively the tracking module. To analyse independently the tracking algorithm, results are also provided where manual annotation was used to define pivot points (see Table I): the mean error increases from 13.5 pixels to 15.1 pixel when tracking is combined with automatic pivot point detection.

Figure 7 shows a frame by frame comparison of pose reconstruction errors between a single particle filter (without back-tracking or feedback) and two trackers built on our multiple particle filter framework with or without the addition of a Kalman filter, respectively called double or integrated trackers. Not only does our system perform significantly better than a single particle filter, but this chart also highlights one of the strength of our proposition: tracking is able to recover from serious divergence because of the partial reinitialisation provided by detection of pivot points and trajectory constraints. For example, although the integrated tracker starts diverging around frame 200, where limbs reach their apparent maximum size and are self-occluded, legs are accurately labelled on frames 219 and 242 (see Fig. 7 and 9). The figure also shows that the double tracker (T2) is more accurate than the integrated one (TI). However, since T2 relies on blob position, its incorrect estimation, e.g. around frame 250, may temporary cause poor pose reconstruction until the tracker's recovery. Analysis of the data of the column "Automatic pivot point detection" in Table I, which corresponds to the practical usage of our system, reveals that the double tracking strategy not only generally improves the mean accuracy of recovered poses, but also is much more stable than the integrated tracker (TI): T2 is in average 14% more accurate with a standard deviation 35% smaller in the case of automatic pose recovery.

Table II shows how our results compare with other techniques used to recover either 2D or 3D poses from the HumanEva data sets. When authors only provided mean errors for 3D poses, they were converted in pixels using approximate relationships between pixel and object lengths for each of the HumanEva data sets. Thus, for a subject height between 250 and 410 pixels and an assumed human height of 1.80 meters [11], a 1 pixel error is equivalent to an error of 4.4-7.2 mm, depending of the position of the person in the image and the perspective.

Most methods perform similarly to our on the HumanEva data sets, i.e. a pixel error in the 12-15 and 17-20 ranges for respectively HE I and HE II. Howe [9], [10], Poppe et al. [11] and Okada and Soatto [50] present example-based approaches to pose recovery, but use very different image descriptors, respectively silhouettes and histograms of oriented gradients (the last two papers). [49] has recently proposed a spatio-temporal 2D-model that allows a monocular pose recovery where the 2D limitations are tackled by the use of a probabilistic transition matrix. Finally, the hierarchical particle filter proposed by Husz et al. [13] relies on a motion model based on action primitives which predicts the next pose in a stochastic manner. Although their tracker performs similarly to the other methods when 2 or more camera sequences are available, its performances degrade significantly when processing a single sequence. The main drawback of all these methods is they are action specific and therefore they are not able to track individuals which display either unexpected motions or a combination of motions. The only approach which presents much more accurate results is proposed by Lee and Elgammal's [42]. Their work is based on a manifold whose topology is learned using a training set. Although they can claim a joint mean accuracy of 31 mm, i.e. 5 to 7 pixels, their approach relies on an even more constrained scenario: walking sequences or cyclic activities that have to be learnt explicitly. The outcome of this comparison is, first, that, since our framework is based on a generative approach, our approach is the only one which does not require any training phase, and therefore, is able to recover human poses of unusual movements as shown in Figure 10 and 12. Secondly, although our scheme does not rely on a constrained environment, it is

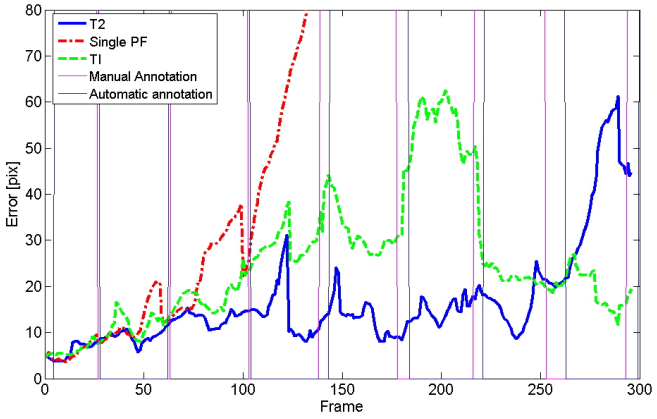


Fig. 7. Tracking error for each frame of the part 1 of S2_Combo_1_(C1) (HE II) sequence. Magenta and dark blue vertical lines are respectively the manual and automatic detection of the begining/end of a 'step'. Red dashed-dotted line is the error using a single particle filter, green dashed line shows the error using the multiple particle filter framework with the integrated tracking strategy (TI) and blue solid line shows the error using the multiple particle filter framework with the double tracking strategy (T2).

Sequence	Frames	Absolute Mean Error [in pixels]	Standard Deviation [in pixels]
Sidenbladh	153	8.45	4.74

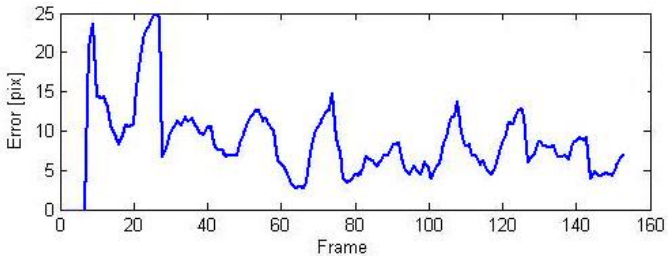


Fig. 8. Numerical results for H. Sidenbladh sequence (Two trackers strategy)

TABLE I
COMPARISON OF PERFORMANCES OF DOUBLE TRACKING (T2) AND INTEGRATED TRACKING (TI) STRATEGIES USING EITHER MANUAL OR AUTOMATIC PIVOT POINT DETECTION

Sequence (C1 camera)	Frames	Manual pivot point: Absolute Mean Error [in pixels]		Manual pivot point: Standard Deviation [in pixels]		Automatic pivot point: Absolute Mean Error [in pixels]		Automatic pivot point: Standard Deviation [in pixels]	
		T2	TI	T2	TI	T2	TI	T2	TI
S2_Walking1, HE I	[6, 418]	17.1	17.1	8.9	8.7	16.5	25.9	4.9	10.9
S2_Combo2, HE I	[1661, 2054]	11.6	9.3	7.4	5.8	12.0	9.1	7.7	6.2
S2_Combo1, HE II	[1, 307]	16.2	25.1	11.1	12.4	24.6	25.8	11.2	12.3
S2_Combo1, HE II	[747, 1202]	9.9	9.7	2.4	2.3	10.0	10.0	1.7	3.1
Total	1570	**13.5**	14.6	**9.0**	9.9	**15.1**	17.6	**9.5**	14.7

TABLE II
COMPARISON WITH STATE OF THE ART

Algorithm	Data set	Pix. error	Constraints	Training	Initialised
Manual pivot	HE I	13.2	Bipedal motion	No	Yes
	HE II	15.9	Bipedal motion	No	Yes
Automatic pivot	HE I	17.5	Bipedal motion	No	Yes
	HE II	17.7	Bipedal motion	No	Yes
Lee et al. [42]	HE I	5-7*	Activity specific & cyclic	Yes	No
Howe [9], [10]	HE I	12.5	Activity specific	Yes	No
	HE II	18.5	Activity specific	Yes	No
Poppe et al. [11]	HE I	10-14*	View and activity specific	Yes	No
	HE II	17-20*	View and activity specific	Yes	No
Husz et al. [13]	HE I	33	Single calibrated camera	Yes	Yes
	HE I	14.8	Multiple calibrated cameras	Yes	Yes
	HE II	19	Multiple calibrated cameras	Yes	Yes
Rogez et al. [49]	HE I	-	View and activity specific	Yes	Yes
	HE II	16.7	View and activity specific	Yes	Yes
Okada et al. [50]	HE I	6-9*	View and activity specific	Yes	Yes
	HE II	-	View and activity specific	Yes	Yes

* Pixel error estimated from 3D error

able to produce results whose accuracy is similar to most state of the art techniques.

Finally, our double tracking strategy was tested on outdoor data (Figure 13). Quantitative results for the Sidenbladh sequence confirm the accuracy and robustness of our method (Figure 8). Since the resolution of this data is about the half of the HumanEva data set's, pixel accuracy cannot be directly compared with those obtained with HumanEva. However, we could estimate that, at equal resolution, an accuracy of about 17 pixels would be achieved, which is in line with values shown in Table I. This experiment demonstrates the generality of our method to environment with different image resolutions, perspectives and illumination conditions, i.e. indoor and outdoor scenes.

V. CONCLUSION

This paper introduces a novel framework based on a set of Kalman and particle filters to track human body parts from a single camera. Its main contribution is the usage of a 2D articulated model constrained by human biomechanics. We have shown that such 2D model is as accurate at tracking 3D motions as 3D models. Not only does the use of a 2D model reduce the computation complexity of tracking human body parts, but also simplifies the tracker initialisation. Moreover, risks of divergence are reduced by our framework capacity of partial reinitialisation at each step.

As demonstrated in experiments with walking and balancing sequences, the main advantage of our system is that it is able to handle any bipedal motion and is not constrained to specific activities as most other methods. The only limitation of the system is that the pivot point should not be occluded for an extended period of time. To deal with this situation, more advanced reinitialisation methods should be integrated in the system [51].

ACKNOWLEDGMENT

This work was partially supported by the EPSRC sponsored MEDUSA and PRoCeSS projects (Grant No. EP/E001025/1 and EP/E033288 respectively), TIN2006-11044 and Feder from Spanish Ministry of Education, and FPI grant (BES-2004-3741) from MEyC.

REFERENCES

[1] Y. Wu, G. Hua, and T. Yu, "Tracking Articulated Body by Dynamic Markov Network", in *ICCV*, pp. 1094-1101 (2003).

[2] G. McAllister, S.J. McKenna, and I.W. Ricketts, "MLESAC-Based Tracking with 2D Revolute-Prismatic Articulated Models", in *ICPR*, vol. 2, pp. 725-728, Quebec (2002).

[3] P. Noriega, and O. Bernier, "Multicues 2D Articulated Pose Tracking Using Particle Filtering and Belief Propagation on Factor Graphs", in *ICIP*, vol. 5(2), pp. 725-728 (2007).

[4] I. Bouchrika, and M. S. Nixon, "People detection and recognition using gait for automated visual surveillance", in *IET Conf. on Crime and Security*, pp. 576-581 (2006).

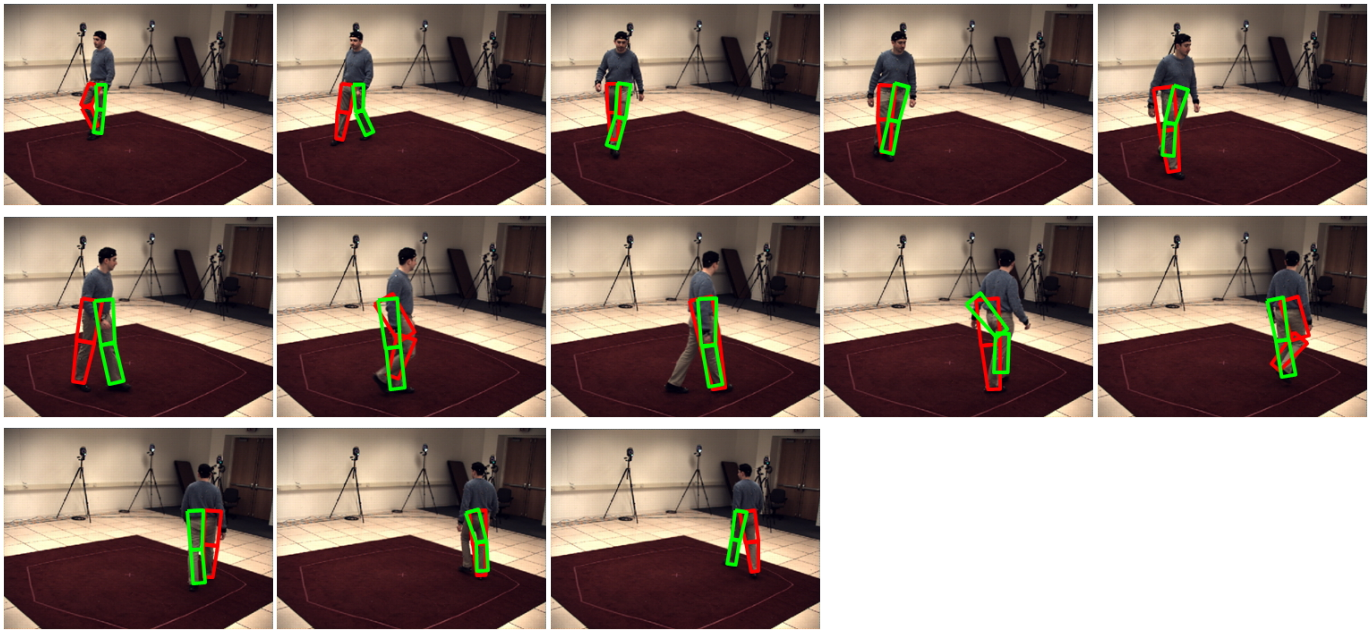


Fig. 9. Results for *S2_Combo_1_(C1)* (HumanEva II) sequence. Frames: 1, 26, 51, 76, 101, 126, 151, 176, 201, 226, 251, 276, 291.

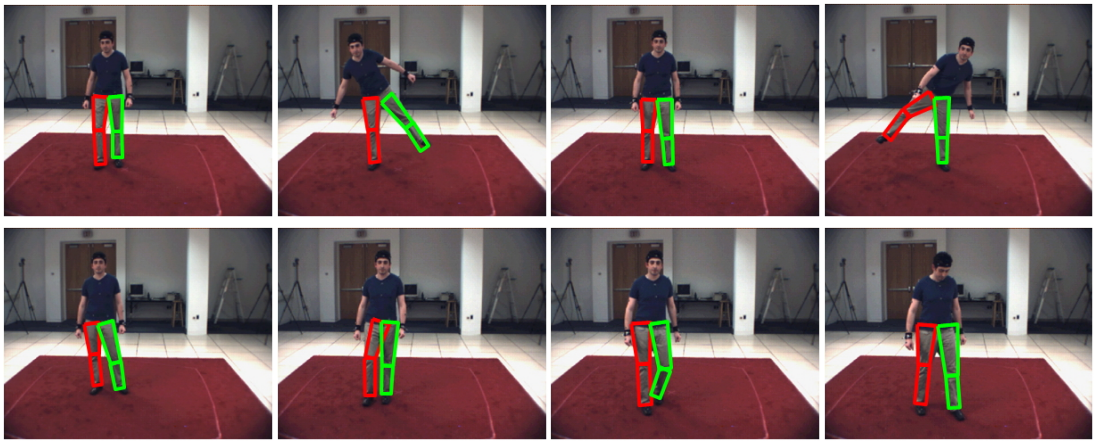


Fig. 10. Results for *S2_Combo_2_(C1)* (HumanEva I) sequence. Frames: 1661, 1731, 1801, 1871, 1941, 2011, 2041, 2071

[5] L. Sigal and M. J. Black, "HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion", in *Technical Report CS-06-08, Brown Univ.* (2006).

[6] J. Xue, N. Zheng, J. Geng, and X. Zhong, "Tracking Multiple Visual Targets via Particle-Based Belief Propagation", in *IEEE Transactions on Systems, Man and Cybernetics - Part B* , Vol. 38 (1), pp. 196-209 (2008).

[7] L. Li, W. Huang, I. Yu-Hua Gu, R. Luo, and Q. Tian, "An Efficient Sequential Approach to Tracking Multiple Objects Through Crowds for Real-Time Intelligent CCTV Systems", in *IEEE Transactions on Systems, Man and Cybernetics - Part B* , Vol. 38 (5), pp. 1254-1269 (2008).

[8] L. Sigal and M. J. Black, "Predicting 3D People from 2D Pictures", in *AMDO* (2006).

[9] N. R. Howe, "Evaluating Lookup-Based Monocular Human Pose Tracking on the HumanEva Test Data", in *Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2)*, (2007).

[10] N. R. Howe, "Recognition-Based Motion Capture and the HumanEva II Test Data", in *Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2)*, (2007).

[11] R. Poppe, "Evaluating Example-based Pose Estimation: Experiments on the HumanEva Sets", in *Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2)*, (2007).

[12] C.S. Lee, and A. Elgammal, "Body pose tracking from uncalibrated camera using supervised manifold learning", in *Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM)*, Whistler, Canada, (2006).

[13] Z. L. Husz, A. M. Wallace, and P. R. Green, "Evaluation of a Hierarchical Partitioned Particle Filter with Action Primitives", in *Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2)*, (2007).

[14] J. Deutscher, and I. D. Reid, "Articulated body motion capture by stochastic search", in *Int. Jour. of Computer Vision*, vol. 61(2), pp. 185-205 (2005).

[15] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition", in *Int. Jour. of Computer Vision*, vol. 61(2), pp. 55-79 (2005).

[16] M. W. Lee and I. Cohen, "Human body tracking with auxiliary measurements", in *IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, pp. 112-119 (2003).

[17] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking", in *ECCV*, vol. 2, pp. 3-19 (2000).

[18] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking", in CVPR, vol. 1, pp. 447-454 (2001).

[19] M. Isard and A. Blake, "Condensation: conditional density propagation for visual tracking", in *Int. Jour. of computer vision*, 29 (1) pp. 5-28 (1998).

[20] J. Deutscher, A. Blake and Ian Reid, "Articulated Body Motion Capture by Annealed Particle Filtering", in *CVPR* vol. 2, pp. 126-133 (2000).

[21] K. Choo and D. Fleet, "People Tracking Using Hybrid Monte Carlo Filtering", in *ICCV*, pp. 321-328 (2001).

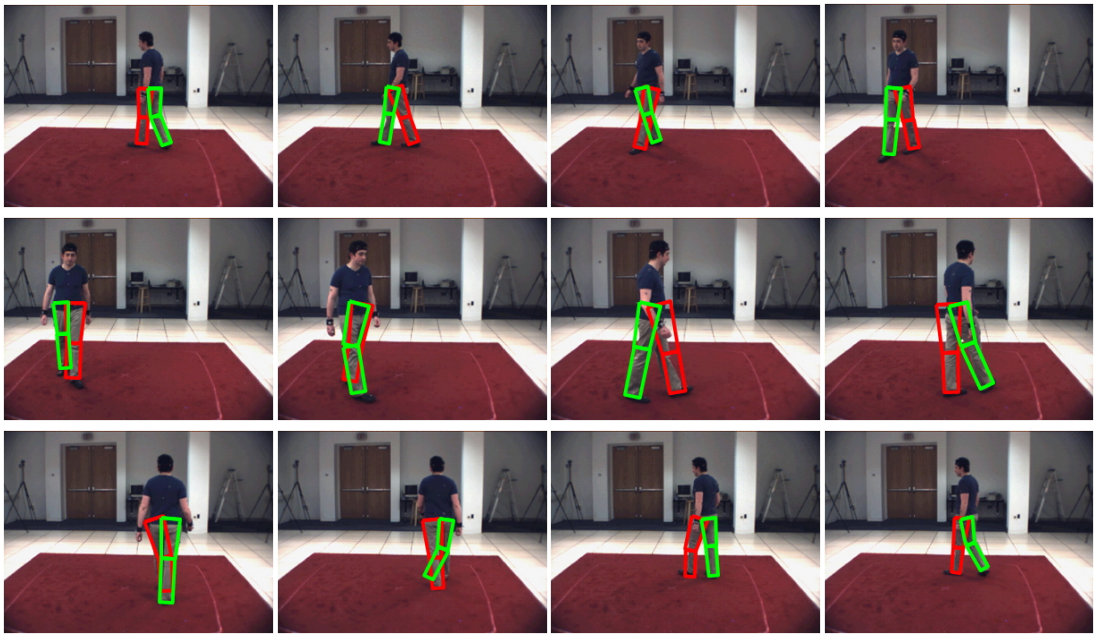


Fig. 11. Results for *S2_Walking_1_(C1)* (HumanEva I) sequence. Frames: 6, 46, 86, 126, 166, 206, 246, 286, 326, 366, 406, 423.

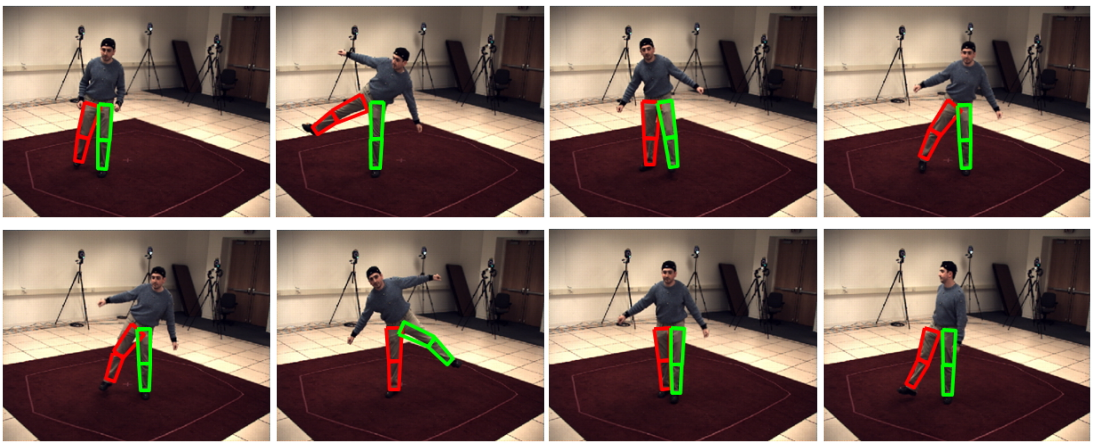


Fig. 12. Results for *S2_Combo_1_(C1)* (HumanEva II) sequence. Frames: 748, 818, 888, 958, 1028, 1098, 1168, 1224.

[22] C. M. Fryer, "Biomechanics of the lower extremity", in *Instruct Course Lect*, vol. 20, pp. 124-130 (1971).

[23] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated Image Motion", in *Int. Conf. on Automatic Face and Gesture Recognition*, pp. 38-44 (1996).

[24] J. O'Rourke and N.I. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation", in *IEEE PAMI*, Vol. 2(6), pp. 522-536 (1980).

[25] D. Hogg, "Model-based vision: a program to see a walking person", in *Image and Vision Computing*, Vol. 1(1), pp. 5-20 (1982).

[26] Z. Chen , H. Lee, "Knowledge-Guided Visual Perception of 3-D Human Gait from a Single Image Sequence", in *IEEE Transactions on Systems, Man and Cybernetics* , Vol. 22 (2), pp. 336-342 (1992).

[27] K. Rohr, "Towards model-based recognition of human movements in image sequences", in *CVGIP: Image Underst.*, Vol. 59(1), pp. 94-115 (1994).

[28] J. M. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects", in *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, pp. 612-617 (1995).

[29] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of humans in action: a multi-view approach", in *CVPR '96: Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 73 (1996).

[30] I. A. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion with occlusion based on active multiviewpoint selection", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 81-87 (1996).

[31] A. Elgammal and C. Lee, "Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 681-688 (2004).

[32] B.Li, Q. Meng, H. Holstein, "Articulated Pose Identification With Sparse Point Features", in *IEEE Transactions on Systems, Man and Cybernetics - Part B*, Vol. 34 (3), pp. 1412-1422 (2004).

[33] R. Li, M.H. Yang, S. Sclaroff and T.P. Tian, "Monocular Tracking of 3D Human Motion with a Coordinated Mixture of Factor Analyzers" in *ECCV06*, Vol. 2, pp. 137-150 (2006).

[34] Q. Wang, G. Xu, and H. Ai, "Learning object intrinsic structure for robust visual tracking", in *Proc. CVPR*, Vol. 2, pp. 227 (2003).

[35] A. Safonova, J. K. Hodgins and N. S. Pollard, "Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces", in *ACM Trans. Graph.*, Vol. 23(3), pp. 514-521 (2004).

[36] A. Rahimi, B. Recht and T. Darrell, "Learning Appearance Manifolds from Video", in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 868-875 (2005).

[37] R. Urtasun, D. J. Fleet and P. Fua, "Monocular 3-D Tracking of the Golf Swing", in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 932-938 (2005).

[38] R. Urtasun, D. J. Fleet and P. Fua, "3D People Tracking with Gaussian Process Dynamical Models", in *CVPR '06: Proceedings of the 2006*

Fig. 13. Results for H. Sidenbladh sequence. Frames: 15, 30, 45, 60, 75, 90, 105, 120, 135, 150.

IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 238-245 (2006).

[39] R. Urtasun, D. J. Fleet, A. Hertzmann and P. Fua, "Priors for People Tracking from Small Training Sets", in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, Vol. 1, pp. 403-410 (2005).

[40] C. Sminchisescu and A. Jepson, "Generative modeling for continuous non-linearly embedded visual inference", in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, pp. 96 (2004).

[41] S.B. Hou, A. and Galata, F. Caillette, N. Thacker and P. Bromiley, "Real-time Body Tracking Using a Gaussian Process Latent Variable Model", in *ICCV07*, pp. 1-8 (2007).

[42] C. Lee and A. Elgammal, "Body Pose Tracking From Uncalibrated Camera Using Supervised Manifold Learning", in *NIPS- Workshop on Evaluation of Articulated Human Motion and Pose Estimation. EHuM06* (2006).

[43] S. X. Ju, M. J. Black and Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated Image Motion", in *FG '96: Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 38 (1996).

[44] J.M. Rehg, D.D. Morris and T. Kanade, "Ambiguities in Visual Tracking of Articulated Objects Using Two- and Three-Dimensional Models", in *The International Journal of Robotics Research*, Vol. 22(6), pp. 393-418 (2003).

[45] E. A. Hunter, P. H. Kelly and R. C. Jain, " Estimation of Articulated Motion Using Kinematically Constrained Mixture Densities", in *NAM '97: Proceedings of the 1997 IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 10 (1997).

[46] H. Sidenbladh, M. J. Black and D. J. Fleet, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion", in *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pp. 702-718 (2000).

[47] T. Tian, R. Li and S. Sclaroff, "Articulated Pose Estimation in a Learned Smooth Space of Feasible Solutions", in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 50 (2005).

[48] J. Martínez, J.C Nebel, D. Makris, C. Orrite, "Tracking Human Body Parts Using Particle Filters Constrained by Human Biomechanics", in *BMVC'08*, Leeds, UK (2008).

[49] G. Rogez, C. Orrite and J. Martínez-del-Rincon, "A spatio-temporal 2D-models framework for human pose recovery in monocular sequences", in *Pattern Recognition*, Vol. 41, Issue 9, pp. 2926-2944 (2008).

[50] R. Okad and S. Soatto, "Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images", in *European Conference on Computer Vision (ECCV)* (2008).

[51] P. Kuo, A. Thibault, M Lewandowski, D. Makris, J.-C. Nebel, "Exploiting Human Bipedal Motion Constraints for 3D Pose Recovery from a Single Uncalibrated Camera", in *International conference on Computer Vision theory and Applications (VISAPP2009)*, (2009).

[52] A. Utsumi, H. Mori, J.Ohya and M. Yachida, "Multiple-View-Based Tracking of Multiple Humans", in *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition*, Vol. 1, pp. 597 (1998).

[53] L. Raskin, E. Rivlin, and M. Rudzsky, "Using Gaussian process annealing particle filter for 3D human tracking", in *EURASIP J. on Adv. in Sig. Proc.*, (2008).

[54] Z. Lu, M. C. Perpinan, and C. Sminchisescu, "People tracking with the Laplacian eigenmaps latent variable model", in *NIPS*, (2007).

[55] X. Li, S. J. Maybank, S. Yan, D. Tao and D. Xu,, "Gait Components and Their Application to Gender Recognition", in *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, Vol. 38(2), pp. 145-155 (2008).

[56] J. Darby, B. Li, N. Costen, D. Fleet and N. Lawrence, "Backing Off: Hierarchical Decomposition of Activity for 3D Novel Pose Recovery", in *BMVC'09*, London, UK (2009).

[57] J. Wang, and Y. Yagi, "Adaptive Mean-Shift Tracking With Auxiliary Particles", in *IEEE Transactions on Systems, Man and Cybernetics - Part B*, Vol. 39 (6), pp. 1578-1589 (2009).

Jesus Martínez del Rincón received the PhD degree from the University of Zaragoza specialising in Biomedical Engineering in 2008. He previously graduated from the University of Zaragoza in Telecommunication in 2003. He is currently a research fellow in the Faculty of Computing, Information Systems and Mathematics, Kingston University, London.His current research interests include aspects of computer vision such as human motion analysis, activity recognition and multi-target tracking in real time.

Dimitrios Makris received the diploma in Electrical and Computer Engineering from Aristotle University of Thessaloniki, Greece in 1999 and the PhD in Computer Vision from City University, London in 2004. He is currently a senior lecturer in the Faculty of Computing, Information Systems and Mathematics, Kingston University, London. His research interests are in the area of image processing, computer vision and machine learning, especially in motion analysis and pose recovery. He is a member of the IEEE and an elected member of the Executive Committee of the British Machine Vision Association (BMVA).

Carlos Orrite Uruñuela received the Master's degree in Industrial Engineering at the Zaragoza University in 1989. In 1994, he completed the Master's degree in biomedical engineering working in the field of medical instrumentation, and in 1997 he did his PhD on computer vision. He is currently an associate professor at the Department of Electronics and Communications Engineering, at the University of Zaragoza and carries out his research activities in the Aragon Institute of Engineering Research (I3A). His research interests are in the area of computer vision and human-machine interface.

Jean-Christophe Nebel received the MSc(Eng) degree in Electronics and Signal Processing in 1992 from the Institute of Chemistry and Industrial Physics, Lyon, France. He completed in 1997 the PhD degree in Parallel Programming from the University of St Etienne, France. He is currently an associate professor in the Faculty of Computing, Information Systems and Mathematics at Kingston University, London. His research interests include computer vision and bioinformatics. He was awarded in 2004 with co-authors the A. H. Reeve Premium by the Council of the Institute of Electrical and Electronics Engineers for a journal paper describing his pioneer work in developing a 3D Dynamic Whole Body Measurement System. He is a senior member of the IEEE.