Multi-view 3D Human Pose Estimation combining Single-frame Recovery, Temporal Integration and Model Adaptation

Michael Hofmann
TNO Defence, Security and Safety
The Netherlands
K.M.Hofmann@uva.nl

Dariu M. Gavrila
Intelligent Systems Laboratory
Faculty of Science
University of Amsterdam (NL)
D.M.Gavrila@uva.nl

Abstract

We present a system for the estimation of unconstrained 3D human upper body movement from multiple cameras. Its main novelty lies in the integration of three components: single-frame pose recovery, temporal integration and model adaptation. Single-frame pose recovery consists of a hypothesis generation stage, where candidate 3D poses are generated based on hierarchical shape matching in the individual camera views. In the subsequent hypothesis verification stage, candidate 3D poses are re-projected to the other camera views and ranked according to a multi-view matching score.

Temporal integration consists of computing best trajectories combining a motion model and observations in a Viterbi-style maximum likelihood approach. Poses that lie on the best trajectories are used to generate and adapt a texture model, which in turn enriches the shape component used for pose recovery. We demonstrate that our approach outperforms the state-of-the-art in experiments with large and challenging real-world data from an outdoor setting. The new data set is made public to facilitate benchmarking.

1. Introduction

The recovery of 3D human pose is an important problem in computer vision with many potential applications in animation, interactive games, motion analysis (sports, medical) and surveillance. 3D pose also provides meaningful, view-invariant features for a subsequent activity recognition step. Despite the considerable advances that have been made over the past years (see next Section), the problem of 3D human pose recovery remains essentially unsolved. The challenges involve estimating articulated motion of bodies of which the exact proportions are not known in advance, dealing with the underconstrained nature of the problem due to loss of depth information and/or (self) occlusion, and performing foreground-background segmentation.

This paper presents a multi-camera system for the estimation of 3D human upper body movement which specifically addresses the combination of single-frame pose recovery, temporal integration and model adaptation. See Figure 1. Using input from three calibrated cameras, we are able to handle arbitrary movement (i.e. not limited to walking and running) in cluttered scenes with non-stationary backgrounds. We do not require particular initial poses to jumpstart the system. A further appealing aspect of the system is that, for single-frame pose recovery, the computational burden is shifted as much as possible to the off-line stage, so that on-line processing is optimized. Algorithmic complexity is sub-linear in the number of body poses considered as a result of a hierarchical representation and matching scheme. Moreover, by fusing information between cameras at the pose parameter level rather than at the feature level, inherent parallelism is increased.

The proposed system also has some limitations. Like previous 3D pose recovery systems, it currently cannot handle a sizable amount of external occlusion. It furthermore assumes the existence of a 3D human model that roughly fits the person in the scene (we are able to use the same generic model for different persons in the experiments).

2. Previous work

There is meanwhile a very extensive literature on 3D human pose estimation. Space limitations force us to make a selection which we consider is most relevant to this paper. For a more exhaustive listing, see recent surveys [9, 22].

One line of research has focused on 3D model-based tracking; i.e. given a reasonably accurate 3D human model and an initial 3D pose, predict the pose at the next time step using a particular dynamical and observation model [5, 7, 8, 11, 13, 25, 30, 35, 36, 37]. Multi-hypothesis approaches based on particle filtering [5, 7, 25, 37] or non-parametric belief propagation [33] are used for increased

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

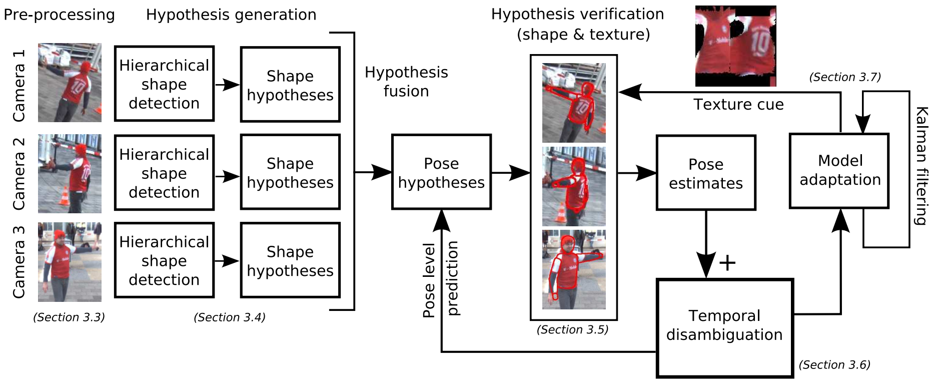


Figure 1. System overview. For details, please refer to the text, Section 3.1.

robustness. However, the high dimensionality of the pose parameter space necessitates researchers to employ strong motion priors (i.e. known action classes such as walking, running) and/or various sequential sampling techniques. In practice, tracking soon goes astray if no recovery mechanism is added.

Another line of research has dealt with 3D pose initialization. Work in this category can be distinguished by the number of cameras used. Multi-camera systems for 3D pose initialization were so far applied in controlled indoor environment. The near-perfect foreground segmentation resulting from the stationary background, together with the many cameras used (> 5), allows to recover pose by Shape-from-Silhouette techniques [6, 15, 21, 34].

Single camera systems for 3D pose initialization can be sub-divided whether they use generative or learning-based approaches. Learning-based approaches construct a mapping between 3D pose and 2D image observables using machine learning techniques [1, 3, 14, 32]. These approaches are conceptually appealing and fast, but questions still remain regarding their scalability to arbitrary poses. Certainly, a large number of examples would be needed in that case to allow for successful regression, given the ill conditioning and high dimensionality of the problem (most experimental results involve restricted movements, i.e. walking). Furthermore, learning-based approaches tend to rely on good foreground segmentation.

On the other hand, pose initialization using 3D generative models [17, 19] involves finding the best match between model projections and image, and retrieving the associated 3D pose. Pose initialization using 2D generative models [23, 28] involves a 2D pose recovery step followed by a 3D inference step with respect to the joint locations. In order to reduce the combinatorial complexity associated with pose recovery, previous generative approaches apply part-based decomposition techniques [33]. This typically involves searching first for the torso, then arms and legs [23, 24, 28]. This decomposition approach is error prone, in the sense that estimation mistakes made early on based on partial model knowledge cannot be corrected later on. In practice, this means that instances with an appreciable amount of torso movement and rotation are difficult to handle.

Methods for pose initialization can serve to initialize the beforementioned trackers. An increasingly popular alternative is their use in “Tracking-as-recognition” approaches, especially when no strong motion priors are available. Here, pose estimates obtained independently at each time instant are integrated to consistent trajectories, taking into account a more generic motion model. This is typically achieved by Markov chain optimization [10, 20, 24, 26].

The contributions of this paper are two-fold. The main contribution is the integration of three components into one system: single-frame pose recovery, temporal integration and texture-based model adaptation. The enforcement of temporal coherence results in a *set* of most likely pose trajectories, which are used to generate predictions. These are subsequently integrated into the pose estimation process. This is unlike [24], where the computation of the optimal pose trajectory is solely a post-processing step, decoupled from the estimation process. Model adaptation in our approach does not require a pre-defined key pose (i.e. feet apart) [10, 28] or a scripted initialization movement [13]. Instead, the likelihood of a shape model match is used as weighting factor for texture-based model adaptation, if the former is above a certain threshold. To further reduce the chance of a wrong model update, we update only for those poses which lie on the most likely pose trajectory (i.e. we perform batch-mode temporal integration before model adaptation, rather than model adaptation at each time instant independently [2]). We do not use strong motion priors [33].

The second contribution concerns the way multi-camera pose recovery is performed. The error-prone foreground segmentation resulting from operating in dynamic outdoor environments together with the lower number of cameras used prevents solving matters by Shape-from-Silhouette (SfS) techniques [6, 15, 21, 34]. Inverse kinematics techniques [13, 16], on the other hand, require close initial estimates. We propose to perform 3D pose recovery for each camera independently and fuse information at the pose parameter level (high-level). This improves the system scalability with respect to the number of cameras (e.g. allowing optimized per-camera matching, improved algorithm parallelism). We only apply SfS to determine a rough ROI (Section 3.3). Furthermore, on a practical note, we demonstrate that an exemplar-based approach for single-frame pose recovery can be used to describe the articulations of the upper-body as a whole. This ensures that upon matching, all available model knowledge is used at the same time, avoiding some of the drawbacks of the part-based decomposition approaches [23, 24, 28] discussed earlier.

3. 3D Pose Estimation

3.1. Overview

Figure 1 presents an overview of the proposed system. Image pre-processing determines a rough region of interest in each 2D image and in the 3D scene, based on background subtraction (Section 3.3). In the hypothesis generation stage, candidate 3D poses are generated based on hierarchical shape matching in the individual camera views (Section 3.4). In the subsequent hypothesis verification stage, the candidate 3D poses are augmented with those derived from a prediction step (except at the first frames). The resulting pose candidates are projected to all camera views and ranked according to a multi-view matching score based on shape and (possibly) texture information (Section 3.5). Temporal integration consists of computing best trajectories in batch mode using a Viterbi-style maximum likelihood approach (Section 3.6). Poses that lie on the best trajectories are used to generate and adapt a texture model (Section 3.7). This provides the beforementioned texture component in the multi-view matching score of hypothesis verification (while no texture model is available during the first frames, the multi-view matching score is based on the shape component only).

3.2. 3D Shape model

Our 3D upper body model uses tapered superquadrics as body part primitives, yielding a good trade-off between desired accuracy and model complexity [11]. Articulation at each joint is represented using transformations of homogeneous coordinates $\mathbf{x}' = H\mathbf{x}$, $H = H(R(\phi, \theta, \psi), T)$, where R is a 3×3 rotation matrix determined by the Euler angles ϕ, θ, ψ, and T a constant 3×1 translation vector. We represent a 3D upper body pose as an 13-dimensional vector of joint angles

$$\begin{aligned}\boldsymbol{\pi} \quad = \quad & \big(\pi_{\text{torso}}(\phi, \theta, \psi), \pi_{\text{l.shoulder}}(\phi, \theta, \psi), \pi_{\text{l.elbow}}(\theta), \\ & \pi_{\text{r.shoulder}}(\phi, \theta, \psi), \pi_{\text{r.elbow}}(\theta), \pi_{\text{head}}(\phi, \psi)\big) \qquad (1)\end{aligned}$$

augmented by a three dimensional vector $\boldsymbol{x}$ denoting the position of the root of the articulated structure (in our case, the torso center).

3.3. Image pre-processing

The aim of pre-processing is to obtain a rough region of interest, both in terms of individual 2D camera views and in terms of the 3D space. For this, we fuse the computed foreground masks of the individual camera views [38] by means of volume carving [18]. After the necessary morphological operations, connected voxel components of a minimum height and size give an estimate of the number of people and their rough 3D location in the scene. This in turn yields information about the image scales and regions of interest to be used in the forthcoming hypothesis generation step (Section 3.4). Projecting the reconstructed voxels onto the camera images produces an improved foreground mask. Note that in the considered outdoor environment with dynamic background and a limited number of cameras (3) we do *not* obtain well segmented human silhouettes in a quality suitable for solving pose recovery by SfS techniques [6, 15, 21, 34] outright.

3.4. Single-camera hypothesis generation using a hierarchical exemplar shape representation

We follow an exemplar-based approach to 3D pose recovery, matching a scene image with a pre-generated silhouette library with known 3D articulation. To obtain the silhouette library, we discretize the state space between lower and upper bounds for each joint angle while allowing a 360° torso rotation along the major body axis, and render the 3D shape model (see Section 3.2) by orthographic projection. The four angles of each arm are discretized into 6 states each, the torso rotation into 15 states, and the remaining angles into 3 states each. The average angle delta is approximately 22°. We reduce the number of allowable joint angle combinations by pruning anatomically impossible poses (using rule-based heuristics) and by collision detection on the shape model. About 15×10^6 poses remain. We further reduce the number of exemplars in our pose library to about 180,000 by clustering the silhouettes based on shape similarity and keeping only the cluster representatives. The remaining silhouette exemplars now contain links to the underlying poses, improving compactness of representation (e.g. ambiguous front/back poses represented only once during silhouette matching). The silhouette exemplars furthermore contain the 2D location of common reference point (in our case, the torso center).

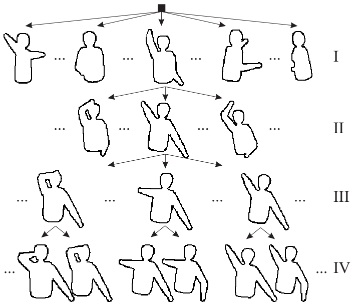


Figure 2. Schematized structure of the 4-level shape exemplar hierarchy. The exemplars at the leaf level represent a variable number of 3D articulations compactly grouped together (Section 3.4).

The exemplars in the library are organized in a (4-level) template tree following [12, 29, 35], see Figure 2. Online matching is efficiently implemented by a tree traversal; search is discontinued below those tree nodes where the match is below a certain (tree-level specific) threshold. Similar to [12], we use the chamfer distance $D_c(A, B)$ between two sets of binary edge features A and B as dissimilarity measure for clustering and matching. Instead of using silhouette exemplars of different scales, we rescale the scene image accordingly using information from the preprocessing step (see Section 3.3). After the matching step, we obtain a ranked lists of silhouette exemplars, one for each camera view. We select the N_i best matches for view i ($N_i = 300$ in the experiments) and extract the previously grouped original poses from each silhouette exemplar. Note that those original poses are in terms of 3D joint angles and do not encode depth due to the used orthographic projection. To obtain candidate 3D poses, we backproject the 2D location of the reference point (torso center) at various depths corresponding to the epipolar line in the other cameras in regions with foreground support. The possible 3D positions and poses are pruned by shape matching the respective projections with the silhouette exemplars which correspond to the found 3D joint angles in the other camera views, using $D_c(A, B)$). For efficiency, this mapping amongst the silhouette exemplars across the camera views is computed off-line by means of a look-up table. We now obtain a ranked list of candidate 3D poses that enter the hypothesis verification step.

3.5. Multi-camera hypothesis verification

For hypothesis verification at time step t, we rank all input pose hypotheses according to a descending probability $p(\{\boldsymbol{\pi}_t, \boldsymbol{x}_t\}|O_t)$ of a pose $\boldsymbol{\pi}_t$ with an associated 3D position $\boldsymbol{x}_t$, given an observation O_t. In the remainder, we simplify the notation $\{\boldsymbol{\pi}_t, \boldsymbol{x}_t\}$ to $\boldsymbol{\pi}_t$, making the 3D position implicit given the pose. We take $\boldsymbol{\pi}_t \in \Pi_t = \Pi_t^{det} \cup \Pi_t^{pred}$, where Π_t^{det} is the list of pose hypotheses from the detection step (see Section 3.4) and Π_t^{pred} the list of prediction hypotheses from the previous time step (see Section 3.6). We model the relation

$$p(\boldsymbol{\pi}_t|O_t) \quad \propto \quad p(O_t|\boldsymbol{\pi}_t)\, p(\boldsymbol{\pi}_t) \qquad (2)$$

according to Bayes' rule, where $p(\boldsymbol{\pi}_t)$ is a uniform prior over the space of anatomically possible poses.

The **observation likelihood** in Equation 2 is decomposed as

$$p(O_t|\boldsymbol{\pi}_t) \propto p(S_t|\boldsymbol{\pi}_t)\, p(T_t|\boldsymbol{\pi}_t) \qquad (3)$$

where S_t describes the shape similarity and T_t the texture similarity. We define

$$\begin{aligned} p(S_t|\boldsymbol{\pi}_t) &= p(D_s(S, E)) && (4)\\ p(T_t|\boldsymbol{\pi}_t) &= p(D_t(T, I)) && (5) \end{aligned}$$

$D_s(S, E)$ in Equation 4 is a shape-based multi-view similarity measure between the reprojected 3D model silhouette S in the pose $\boldsymbol{\pi}_t$ and detected image edges E based on the chamfer distance (see Section 3.4) over *K* cameras.

$$D_s(S, E) = \sum_{k \in K} D_c(S, E) \qquad (6)$$

$D_t(T_k, I)$ in Equation 5 is a measure for the similarity between the pixels u of the reprojected textured model T_k in camera k and the corresponding pixels v of the scene image I and is defined as

$$D_t(T, I) = \sum_{k \in K} \frac{1}{|T_k|} \sum_{u \in T_k} \left(\sum_{c \in \{r,g,b\}} (u_c - v_c)^2 \right)^{\frac{1}{2}} \qquad (7)$$

which is the sum of the average pixel-wise Euclidean distance in our color space (see Section 3.7), $|T_k|$ being the number of pixels on the reprojected model.

The individual probabilities $p(D_s(S, E))$ and $p(D_t(T, I))$ are each modeled using a gamma distribution whose parameters are considered independent from the pose $\boldsymbol{\pi}_t$ and estimated from the training data by maximum likelihood.

Hypothesis clustering. Evaluating the observation likelihood is an expensive operation since it involves rendering a 3D pose across multiple camera views. Assuming that $p(\boldsymbol{\pi}_t|O_t)$ in Equation 2 is a locally smooth function on a neighborhood of $\boldsymbol{\pi}_t$ in pose space, one can implement the following two-step approach to speed up the implementation. We cluster all pose hypotheses using a pose distance measure, evaluate all prototypes according to Equation 2 and only if the latter is above a threshold (obtained by cross-validation), do we evaluate the individual cluster elements.

The pose distance measure we use is

$$d_x(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \frac{1}{|B|} \sum_{i \in B} d_e(\vec{v_1^i}, \vec{v_2^i}) \qquad (8)$$

where B is a set of locations on the human upper body, $|B|$ the number of locations, $\vec{v^i}$ is the 3D position of the respective location in a fixed Euclidean coordinate system, and $d_e(.)$ is the Euclidean distance. For the set of locations, we choose torso and head center as well as shoulder, elbow and wrist joint location for each arm.

3.6. Temporal disambiguation and prediction

Temporal integration facilitates a further disambiguation among the candidate solutions obtained at a single time step, determining pose trajectories that match the observations well and exhibit coherent motion. This is formulated as the following optimization task: given a sequence of observations $O_0, \dots, O_T$ up to time step T, find the pose sequence $\boldsymbol{\pi}_0, \dots, \boldsymbol{\pi}_T \in \Pi_0, \dots, \Pi_T$ that maximizes

$$p(\boldsymbol{\pi}_{0:T}|O_{0:T}) \propto \prod_{t=1}^{T} p(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{t-1}) \times \prod_{t=0}^{T} p(O_t|\boldsymbol{\pi}_t) \qquad (9)$$

$p(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{t-1})$ denotes the pose transition probability as a first-order Markov chain, while $p(O_t|\boldsymbol{\pi}_t)$ is the observation likelihood of Equation 3. This type of problem is solved by application of the Viterbi algorithm [27] on the input data; in our case, this in done in a sliding window over the last 50 frames. We use a List Viterbi Algorithm (LVA) [31] implementation to compute not only the optimal, but the N best trajectories through the Viterbi trellis at each time step.

With respect to the transition model, we make a number of simplifications to reduce the number of parameters involved. We assume the location of the root of the articulated structure to be independent of the joint angle configuration. We furthermore decouple the joint angles associated with the various body parts. Finally, we only consider parameter changes, i.e. we do not condition on specific previous values. We thus set

$$\begin{aligned} p(\vec{\pi}_t|\vec{\pi}_{t-1}) \propto \mathcal{N}(\Delta\vec{x}^{\text{root}}; \mu^{\vec{\text{root}}}, \Sigma^{\text{root}}) \times \qquad (10) \\ \mathcal{N}(\Delta\vec{\pi}_t^{\text{head}}; \vec{\mu}^{\text{head}}, \Sigma^{\text{head}}) \times \mathcal{N}(\Delta\vec{\pi}_t^{\text{torso}}; \vec{\mu}^{\text{torso}}, \Sigma^{\text{torso}}) \\ \times \mathcal{N}(\Delta\vec{\pi}_t^{\text{l.arm}}; \vec{\mu}^{\text{arm}}, \Sigma^{\text{arm}}) \times \mathcal{N}(\Delta\vec{\pi}_t^{\text{r.arm}}; \vec{\mu}^{\text{arm}}, \Sigma^{\text{arm}}) \end{aligned}$$

and estimated the parameters for the different normal distributions by maximum likelihood on the training data.

We generate pose predictions at every time step, which augment the detections of the next time step, as indicated in Figure 1, and are generated using whole trajectory information. To generate K pose predictions (in our system, $K = 200$) at time step t, K trajectories are sampled from the N best trajectories (we chose $N = 500$) with a probability proportional to the trajectory probability determined by the LVA algorithm. For each of these, both pose prediction $\tilde{\vec{\pi}}_{t+1}^k$ and position prediction $\tilde{\vec{x}}_{t+1}^k$ are determined using stochastic sampling. The pose prediction is generated as

$$\tilde{\vec{\pi}}_{t+1}^k := \vec{\pi}_t^k + \Delta\vec{\pi}_{t\rightarrow t+1} \qquad (11)$$

where the values of $\Delta\vec{\pi}_{t\rightarrow t+1}$ are drawn from the normal distributions described in Equation 10. The position prediction is drawn from $\mathcal{N}(\vec{\mu}_{\vec{x},t+1}^k, \Sigma_{\vec{x},t+1}^k)$, where $\vec{\mu}_{\vec{x},t+1}^k$ is the predicted state and $\Sigma_{\vec{x},t+1}^k$ the predicted covariance from Kalman filtering the available trajectory data.

We opted for the above sliding window batch-mode framework rather than a recursive framework because of increased estimation stability. Treating the tracking problem as a detection problem over a discrete pose space in every frame enables us to (re-)initialize the state of our system, while recursive filtering frameworks such as particle filtering might eventually fail and lose track. Furthermore, our prediction mechanism can increase tracking accuracy both by lifting the constraint of the discrete pose space and by "bridging gaps" where detections are poor.

3.7. Model adaptation using texture information

We now turn to augmenting our shape model with texture information in order to increase the discriminative power of hypothesis verification. Clearly, the outcome of texture mapping is very sensitive to the estimated pose of the shape model, and matching with a wrong texture model is truly damaging for pose estimation. In order to avoid incorrect texture model updates as much as possible, we decided not to perform these based on pose estimates at a single time instant, but rather based on the more reliable N trajectories computed in previous section (we currently maintain a single texture model associated to the optimal trajectory).

Given the optimal trajectory returned by the temporal disambiguation step, we evaluate the matching likelihoods $p(D_{s,torso}(S,E))$, $p(D_{s,l.arm}(S,E))$, $p(D_{s,r.arm}(S,E))$ for the chamfer match of each body part $\in$ {torso & head, left arm, right arm} for the last five poses of the trajectory. Similar to the probabilities in Equation 3, these are modeled using gamma distributions estimated from training data. For each body part, we then make a decision to adapt the texture model using the pose in the trajectory with the highest match likelihood, only if this likelihood is above a certain threshold.

In case of model adaptation, we acquire a texture map for the respective body part by sampling the visible area of the superquadrics for each camera view and storing the color values in a texture image. Collision detection on the ray from camera center to the points on the superquadric ensures that we do not sample in areas of self-occlusion through other body parts. The texture images are then combined by choosing for each pixel the sampled value for which the angle between superquadric normal vector and ray from camera center is smallest. Figure 3 shows an example of a reprojected texture map acquired from the depicted pose.

Because images from different cameras are effectively stitched together during the acquisition of a texture map,

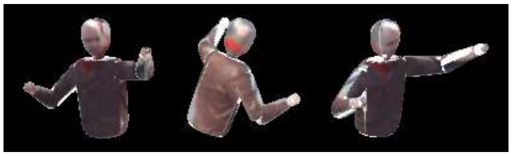


Figure 3. Example of shape model enriched with texture information, rendered from various viewpoints. Parts of the body that are occluded in all cameras stay untextured and are shown in white. Depicted color space is non-normalized RGB.

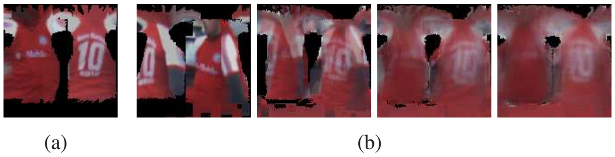


Figure 4. Model adaptation of torso (a) ground truth texture map (b) temporal progression of the texture map from an initial incorrect estimate to a correct (but somewhat blurred) estimate by Kalman filtering

there will be differences in luminance due to camera properties and scene illumination. We reduce the variation induced by global indirect illumination by working in a normalized RGB color space $r = R/L$, $g = G/L$, $b = B/L$, where $L := \frac{1}{|K|}\sum_{k\in K}(R_k + G_k + B_k)$ is the average luminance over the scene pixels K.

The texture model is implemented using Kalman filtering on each pixel of the texture map in order to become more robust to potential input from incorrect estimates. See Figure 4 for an illustration. At each new time step, the state of each filter is evaluated to generate a texture map for use in hypothesis verification (see Section 3.5).

4. Experiments

Our experimental data consists of recordings from three synchronized color CCD cameras looking over a train station platform. In 12 sequences (about 10s on average, captured at 20Hz), various actors perform unscripted movements, such as walking, gesticulation and waving. The setting is challenging; the movements performed contain a sizable amount of torso turning, the background is cluttered and non-stationary (people are walking in the background, trains are passing by), furthermore, there are appreciable lighting changes. The realism of the dataset in the context of surveillance was the key motivation for preferring it over the popular HumanEva dataset [33]. We make this novel data set public to facilitate benchmarking[1].

Cameras were calibrated using Bouguet's method [4]; this enabled the recovery of the ground plane. Ground truth pose was manually labeled for all frames of the data set (considering the quality of calibration and labeling, we estimate the ground truth accuracy to be within 3cm). The general motion model (Section 3.6) was derived from the aggregated CMU *MoCap* data[2]; after some conversions the latter yielded 756,844 frames for training.

Figure 6 shows examples of recovered poses, taken from the best trajectory using shape and texture cues, with the proposed approach; 3D pose is estimated quite well. The main failure mode concerns those ”ambiguous” poses with the hands close to the torso; the silhouette-based approach stands little chance in recovering exact hand position, furthermore, most clothing does not contain appreciable texture differences between torso and arms. Table 1 quantifies the results in terms of the deviation between estimated and ground truth 3D pose over the entire dataset. It shows the successive benefit of adding predictions (Section 3.6) and texture-based model adaptation (Section 3.7) to the single-frame pose recovery, resulting in a reduction of pose error from 12.7cm to 10.9cm.

We furthermore compared the various instantiations of our system with the hierarchical Partitioned Annealed Particle Filter (PAPF) [7]. This is a state-of-the-art technique for tracking high-DOF (unconstrained) articulated movement, which, unlike SfS approaches, does not require perfect silhouette segmentation. To focus on the essential differences, we implemented the PAPF using the same foreground segmentation (Section 3.3) and shape likelihood computation (Eq. 4). We also incorporated the CMU *MoCap* data in the PAPF when initializing the diffusion covariance. After some tuning, we selected a parameterization with 4 layers for our 13 DOF model (cf. 10 layers for a 30 DOF model in [7]). The number of particles per layer was set to 200, as in [7]. The PAPF was initialized with the ground truth in the first frame of each sequence, while our system self-initializes.

In experiments, we observed a suprisingly good performance of the PAPF on many sequences. For more difficult sequences (appreciable background clutter, ”ambiguous” poses as discussed above, fast torso turning), however, we found that the PAPF particles dissipated away from the correct solution after a while, with little chance for recovery. On average, we obtained a considerable outperformance of the proposed approach vs. PAPF (avg. pose error of 10.9cm vs. 14.6cm), even though the latter had a clear headstart starting from the ground truth pose.

Figure 5 provides a closer look at a challenging tracking sequence with two 360° torso turns in short succession (frames 70-150 and 160-420). The plot shows the pose error (Equation 8) for the Viterbi-best trajectory using various system configurations, as well as a comparison with the trajectory obtained by PAPF. The greyish backdrop encodes the distribution of the pose error over the single-frame de-

[1] The data set is made freely available for non-commercial research purposes. See http://www.science.uva.nl/research/isla/downloads/3d-pose-estimation/index.html or contact the second author

[2] http://mocap.cs.cmu.edu/

	our system (S & T, det. + pred.)	our system (S, det. + pred.)	our system (S, det. only)	PAPF [7]
avg	**10.9 (5.1)**	**11.6 (5.7)**	**12.7 (6.6)**	**14.6 (6.6)**

Table 1. Avg. pose error in cm (Eq. 8) over 12 test sequences; standard deviation in brackets (S:shape, T:texture)

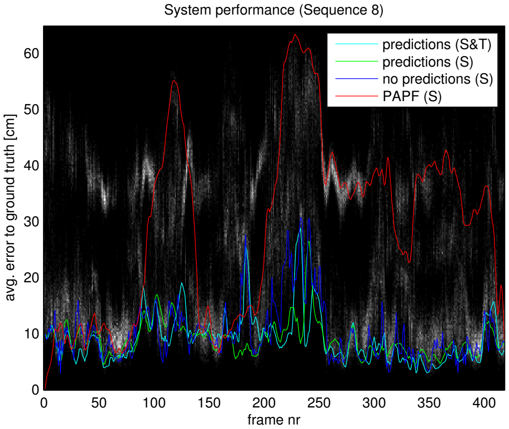


Figure 5. Pose error (in cm) for best trajectory for three system configurations (with and without prediction generation; S:shape, T:texture) and for the PAPF. The background shows histograms of the pose distance of the single-frame detections per time step (lighter shades indicate higher densities).

tections (lighter shades indicate higher densities). One can differentiate two bands of detections; the one with higher pose error (around 40cm) corresponds to poses which are similar in appearance except for a 180° turn of the torso. In this sequence, the Viterbi-based approaches correctly track the two 360° torso turns, whereas the PAPF estimates the torso orientation unchanged. We take this as an example of the increased robustness of the proposed Viterbi trajectory-based estimation (which combines detection and prediction) to momentarily incorrect pose estimates.

The current system requires 15-20s per image triplet to recover 3D pose, running with unoptimized C++ code on a 2.6 GHz Intel PC. Although not fast in absolute terms, it seems to compare favorably to the processing speeds previously reported in the literature concerning 3D pose recovery with generative models against non-stationary background, e.g. [2, 17, 19, 23]; yet direct comparisons are difficult (unconstrained upper body movement vs. whole-body walking). Our performance bottleneck is currently multi-camera hypothesis verification (Section 3.5) and, to a lesser degree, single-camera hypothesis generation (Section 3.4). These components are easily parallelizable, allowing a near-linear reduction of processing speed with available CPU/GPU cores.

5. Conclusion and Further Work

We proposed an integrated system for estimating 3D human upper body pose from multiple cameras. The system combines a hierarchical, exemplar-based single-frame pose recovery, Viterbi-style best trajectory estimation, and a filtering approach to 3D model texturing. We demonstrated an improvement versus the state-of-the-art in a dozen of challenging real-world sequences depicting different actors performing unscripted movements.

Future work involves the recovery of whole-body pose and that of multiple people. A direct extension of the chosen exemplar-based approach to whole-body recovery is possible but rather memory intensive. A more suitable solution, better able to deal with partial occlusion, is to recover upper and lower body pose separately and integrate results.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monoc. images. *TPAMI*, 28(1):44–58, 2006.

[2] A. Balan and M. Black. An adaptive appearance model approach for model-based articulated object tracking. In *CVPR*, 2006.

[3] A. Bissacco *et al.* Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *CVPR*, 2007.

[4] J.-Y. Bouguet. Camera calib. toolbox for matlab, 2003.

[5] M. Brubaker *et al.* Physics-based person tracking using simplified lower-body dynamics. In *CVPR*, 2007.

[6] K. M. G. Cheung et al. Shape-from-silhouette across time - parts I and II. *IJCV*, 62 and 63(3):221–247 and 225–245, 2005.

[7] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005.

[8] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *ICCV*, pages 315–320, 2001.

[9] D. Forsyth et al. Computational studies of human motion. *Found. Trends. Comput. Graph. Vis.*, 1(2-3):77–254, 2005.

[10] A. Fossati *et al.* Bridging the gap between detection and tracking for 3d monocular video-based mo tion capture. In *CVPR*, 2007.

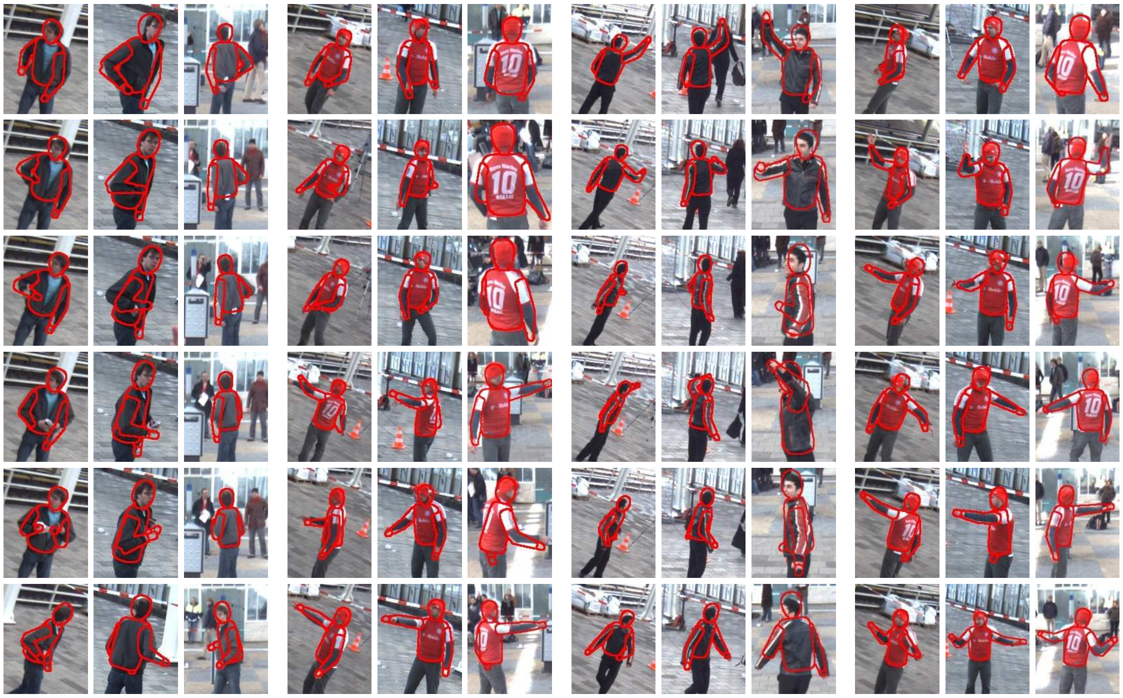
[11] D. M. Gavrila and L. Davis. 3-D model-based tracking of humans in action: a multi-view approach. In *CVPR*, 1996.

[12] D. M. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *ICCV*, pages 87–93, 1999.

[13] I. Kakadiaris and D. Metaxas. Model-based estimation of 3-D human motion. *PAMI*, 22(12):1453–1459, 2000.

[14] A. Kanaujia *et al.* Semi-supervised hierarchical models for 3d human pose reconstruction. In *CVPR*, 2007.

[15] R. Kehl and L. V. Gool. Markerless tracking of complex human motions from multiple views. *CVIU*, 103(2-3):190–209, 2006.



(a) Sequence 2 (b) Sequence 8 (c) Sequence 11 (d) Sequence 12

Figure 6. Examples of recovered poses in the three camera views for multiple persons (best trajectory, shape and texture cues). Shown are image cut-outs.

[16] D. Knossow et al. Human motion tracking with a kinematic parametrization of extremal contours. *IJCV*, 79:247–269, 2008.

[17] P. Kohli et al. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *IJCV*, 79:285–298, 2008.

[18] A. Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 16(2):150–162, 1994.

[19] M. W. Lee and I. Cohen. A model-based approach for estimating human 3D poses in static images. *TPAMI*, 28(6):905–916, 2006.

[20] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and Viterbi path searching. In *CVPR*, 2007.

[21] I. Mikic et al. Human body model acquisition and tracking using voxel data. *IJCV*, 53(3):199–223, 2003.

[22] T. B. Moeslund et al. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 103(2-3):90–126, 2006.

[23] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *TPAMI*, 28(7):1052–1062, 2006.

[24] R. Navaratnam et al. Hierarchical part-based human body pose estimation. In *BMVC*, 2005.

[25] E.-J. Ong *et al.* Viewpoint invariant exemplar-based 3d human tracking. *CVIU*, 104:178–189, 2006.

[26] P. Peursum *et al.* Tracking-as-recognition for articulated full-body human motion analysis. In *CVPR*, 2007.

[27] L. Rabiner. A tutorial on HMMs and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989.

[28] D. Ramanan et al. Tracking people by learning their appearance. *TPAMI*, 29(1):65–81, 2007.

[29] G. Rogez et al. Randomized trees for human pose detection. In *CVPR*, 2008.

[30] B. Rosenhahn et al. Scaled motion dynamics for markerless motion capture. In *CVPR*, 2007.

[31] N. Seshadri and C. Sundberg. List Viterbi decoding algorithms with applications. *Trans. on Communications*, 42:313–323, 1994.

[32] G. Shakhnarovich et al. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, pages 750–757, 2003.

[33] L. Sigal et al. Tracking loose-limbed people. In *CVPR*, 2004.

[34] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *ICCV*, pages 915–922, 2003.

[35] B. Stenger et al. Model-based hand tracking using a hierarchical Bayesian filter. *TPAMI*, 28(9):1372–1384, 2006.

[36] M. Vondrak et al. Physical simulation for probabilistic motion tracking. In *CVPR*, 2008.

[37] X. Xu and B. Li. Learning motion correlation for tracking articulated human body with a Rao-Blackwellised particle filter. In *ICCV*, 2007.

[38] Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In *ICPR (2)*, pages 28–31, 2004.