Computer Vision and Image Understanding 115 (2011) 1559–1570

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

3D Human model adaptation by frame selection and shape–texture optimization

Michael Hofmann [a], Dariu M. Gavrila [a,b,*]

[a] Informatics Institute, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands
[b] Environment Perception Department, Daimler Research and Development, Wilhelm Runge St. 11, 89081 Ulm, Germany

ARTICLE INFO

Article history:
Received 10 November 2010
Accepted 10 August 2011
Available online 18 August 2011

Keywords:
Human motion analysis
3D articulated pose estimation

ABSTRACT

We present a novel approach for 3D human body shape model adaptation to a sequence of multi-view images, given an initial shape model and initial pose sequence. In a first step, the most informative frames are determined by optimization of an objective function that maximizes a shape–texture likelihood function and a pose diversity criterion (i.e. the model surface area that lies close to the occluding contours), in the selected frames. Thereafter, a batch-mode optimization is performed of the underlying shape- and pose-parameters, by means of an objective function that includes both contour and texture cues over the selected multi-view frames.

Using above approach, we implement automatic pose and shape estimation using a three-step procedure: first, we recover initial poses over a sequence using an initial (generic) body model. Both model and poses then serve as input to the above mentioned adaptation process. Finally, a more accurate pose recovery is obtained by means of the adapted model.

We demonstrate the effectiveness of our frame selection, model adaptation and integrated pose and shape recovery procedure in experiments using both challenging outdoor data and the HumanEva data set.

1. Introduction

Markerless 3D human pose recovery has many potential applications in areas such as animation, interactive games, motion analysis and surveillance. Over the last decade, considerable advances have been made in the area of pose initialization and tracking, however, the problem of 3D human shape estimation has received less attention in comparison. Suitable models are often being assumed given or acquired in a separate, controlled model acquisition step. This decoupling is not surprising, given the high dimensionality of the joint pose and model space; typical 3D human pose recovery systems have 25–50 DOFs, and at least that many parameters for modeling shape.

The premise in this paper is that, due its high dimensionality, we cannot accurately estimate the human pose and model space jointly from scratch. We will show, however, that given an approximate initial human model (e.g. anthropometric mean) and a state-of-the-art multi-view pose recovery system [13,26], we can obtain pose estimates which are, at least for a subset of frames, accurate enough to start an optimization process to improve the human model used. This subsequently facilitates more accurate pose recovery.

Given that a person naturally takes different poses and positions over time, one can selectively choose the frames of a sequence at which to adapt a model. Not all frames of a sequence are equally suited; some frames involve poses with depth ambiguity, others involve cases where body parts are self-occluded. Overall, one would like to select a (preferably small) set of frames that contains a certain diversity in poses being observed. This paper describes how to select a suitable subset of frames and how to perform the optimization of the shape and pose parameters robustly, based on shape and texture cues. As we will see in the experiments, our frame selection not only improves system efficiency (by processing only a subset of a sequence), it also leads to a better behaved optimization.

2. Previous work and contributions

There is an extensive amount of literature on 3D human pose recovery, see recent surveys [13,26,35]. Due to space limitations, we focus on literature that recovers both 3D human pose and shape, i.e. the main paper scope.

Previous work on 3D human pose and shape recovery can be distinguished by the type of shape representation used and the way in which shape model adaptation takes place. 3D human shape models come in roughly three categories. The first category represents 3D human shape by a set of voxels [8,16,39]; eigenspace models can be used to introduce priors for the segmentation [16] and to perform body part labeling [39]. The second category represents the various body parts by volumetric primitives [21,25,37]. For example, Mikic et al. [25] uses ellipsoids, whereas Sminchisescu and Triggs [37] chooses for superquadrics with

* Corresponding author at: Environment Perception Department, Daimler Research and Development, Wilhelm Runge St. 11, 89081 Ulm, Germany.
E-mail addresses: mhofmann.uva@gmail.com (M. Hofmann), d.m.gavrila@uva.nl (D.M. Gavrila).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.08.002

global deformations. The third category uses surface models (e.g. meshes) to describe human shape [1–4,10,15,17,18,30,33,38,40]. The use of eigenspace surface models for shape estimation (SCAPE) was popularized by Anguelov et al. [1] and later used in Balan and Black [2], Balan et al. [3], Guan et al. [17] for joint pose and shape estimation under clothing.

Each above-mentioned category of human 3D shape representation comes with its own benefits and drawbacks. The choice for a particular representation should be guided by the application at hand. For example, an application requiring a detailed recovery of the 3D human shape will favor the use of finely tesselated mesh models, especially if loose (non-rigid) clothing is to be considered. The high expressivity of mesh models comes, however, at the price of having to estimate comparatively large number of parameters. In practice, this means that 3D measurement needs to be precise and at a high spatial resolution. This in turn leads to in-door settings with controlled backgrounds, and the availability of a large number of cameras (>5) relatively up-close, for an accurate volume carving [10,38,40]. In some cases, range data is measured by means of stereo vision [30,38]. Yet other work assumes that a detailed initial 3D shape model has been obtained off-line by laser scanning [10,18,40]. Human shape models based on volumetric primitives come typically with a lower number of parameters and are more suited for those applications where high realism is not feasible or not the primary scope, e.g. person tracking from a larger distance with a low number of cameras (1–3) in a surveillance context. Global, eigenspace-based surface models (e.g. SCAPE [1–3,17]) have the ability to reduce the number of parameters, possibly below the number of parameters needed for the local models using volumetric primitives; this as long as a particular test instance is well described by the modes of variation in the training set (a training set of body scans might however not compactly represent all clothed humans, e.g. those in wide jacket and tight pants).

Previous work that performs human 3D shape model adaptation, does this for an initial frame [18,25,34,38], after each frame incrementally [3,10,15,21,25,37,40], or based on several frames, in batch mode [2]. Model and pose initialization is obtained by manual annotation (e.g. registration between 3D model and 2D scene features) [3,4,15,18,30,37,38] or using automatic methods. The latter methods still have some limitations, such as the requirement of specific poses [10] or pre-defined motion scripts [21]. A number of methods are demonstrated on a wider range of initial (upright) poses [15,25,34,40], albeit under somewhat favorable environmental conditions (e.g. uniform background, larger number (>3) of overlapping cameras).

Our main contribution is an automatic frame selection method for 3D human shape and pose adaptation by means of discrete optimization over a pose error estimate (i.e. shape–texture likelihood) and a pose diversity criterion (i.e. the model surface area that lies close to the occluding contours). The idea of selective or "opportunistic" model acquisition was inspired by work by Ramanan et al. [32]. In that work, color information was added to 2D pictorial structure models in frames depicting "easy" poses (what easy poses were was defined heuristically, i.e. person standing with legs apart). Here, we offer a more principled and automatic frame selection approach for 3D shape model adaptation and show that it improves performance over simpler frame selection approaches. We integrate frame selection in a multi-step procedure for automatic shape and pose estimation through initial pose recovery, shape model adaptation and final pose recovery. Our approach provides a robust solution to the problem of estimating both human shape and pose.

A secondary, more limited contribution concerns the definition of a differentiable objective function for 3D human shape model and pose adaptation, based on both contour and texture cues. Previous methods using differentiable objective functions have considered 3D pose adaptation based on contour cues (e.g. [22,31,36]). We perform stochastic gradient-based optimization, similar to Kehl and Van Gool [22], and add vertex resampling to account for shape model changes. The objective functions used in Balan and Black [2], Balan et al. [3] involve contour features only and are non-differentiable. As will be seen in the experiments, the proposed optimization technique based on closed-form expressions for the Jacobian matrix compares favourably, in terms of convergence and processing cost.

Given that our primary application concerns 3D pose tracking in surveillance context (i.e. few overlapping cameras, outdoor scenario with dynamic and cluttered backgrounds, uncooperative subjects at some distance) we opted for a 3D human shape representation by means of volumetric primitives, see earlier discussion. In this set-up, we do not model variation of body shape with pose (e.g. Balan et al. [3]). Note that our contributions in this paper are to a large degree independent of the particular 3D human shape representation used; they could be applied to mesh models (e.g. SCAPE). For frame selection, we require that it is possible to determine vertices lying on the occluding contours, and that we can define a 2D distance measure between vertices lying on the surface. Regarding the objective function, we require the existence of a generation function for the model vertices that is differentiable with respect to pose and shape parameters.

3. 3D model shape and pose adaptation

3.1. Overview

Fig. 1a shows the proposed procedure for automatic human pose and shape estimation. In a first stage, an existing state-of-the-art pose recovery method that relies on a fixed ("generic") shape model is applied to estimate human pose on a "training" sequence (of interest are methods like [19], that are fully automatic and do not require manual intervention or particular intialization poses). Given that the shape model is imprecise, however, we expect pose recovery quality to be lacking. In a second stage, model adaptation is performed by the techniques introduced in this paper, utilizing this "training" sequence. In a final stage, the beforementioned state-of-the-art method is applied to estimate pose on a "test" sequence, now using the optimized model (the "test" sequence could well equal the "training" sequence).

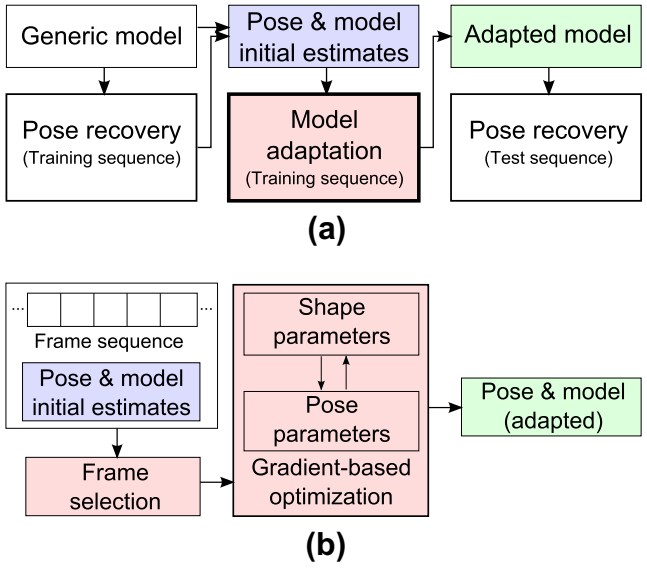

Fig. 1. (a) Overview of the combined pose and shape estimation process. (b) Overview of the model adaptation approach, see Section 3.1.

Fig. 1b shows the model adaptation process (the middle section in Fig. 1a) in more detail. The input consists of a sequence of frames, initial pose estimates for each frame and an initial shape model estimate. In a frame selection step, the sequence is analyzed and a subset of informative frames is selected based on maximizing a likelihood term based on silhouette fit, and based on a criterion that captures pose diversity (Section 3.3). Thereafter, a gradient-based model and shape optimization procedure is performed on the selected frames. The objective function incorporates the same silhouette fit term as used previously for frame selection (Section 3.4). We output the optimized shape model and the poses for the selected frames; the optimized model can be used for a subsequent pose recovery/refinement step.

3.2. Shape and pose representation

We use an articulated model with linearly tapered superquadrics [20] as geometric primitives for torso, neck, head, upper and lower arm, hand, upper and lower leg, and foot, assuming body symmetry. The parameter space $\vec{s}_i$ of each superquadric i comprises parameters for length (a_1,a_2,a_3), squareness (e_1,e_2) and tapering (t_x,t_y). Its generating function for a surface vertex is

$$\zeta(u,v)=\begin{pmatrix} a_1\cos^{e_1}u\cos^{e_2}v(t_x\sin^{e_1}u+1)\\ a_2\cos^{e_1}u\sin^{e_2}v(t_x\sin^{e_1}u+1)\\ a_3\sin^{e_1}u\end{pmatrix},\quad -\frac{\pi}{2}\leqslant u\leqslant\frac{\pi}{2},\quad -\pi\leqslant v\leqslant\pi \tag{1}$$

where u, v are angle parameters.

Joint articulation is described by homogeneous coordinate transformations $\mathbf{x}'=H\mathbf{x}$, where $H:=H(R(\phi,\theta,\psi),\vec{t})$ contains a 3×3 rotation matrix and a 3×1 translation vector. Limb transformations not at the model root (i.e. the torso center) are represented by a kinematic chain $\mathbf{H}=H_1H_2\cdots H_k$ along the respective joints. We make a few simplifying assumptions and represent articulation only at the neck, shoulder, elbow, pelvis and knee joints. A full body pose is then described by a 25-dimensional pose vector

$$\vec{p}=(t_x,t_y,t_z,(\phi,\theta,\psi)_{\text{torso}},(\phi,\theta,\psi)_{\text{head}},(\phi,\theta,\psi)_{\text{l.sh.}},\theta_{\text{l.elb.}},(\phi,\theta,\psi)_{\text{r.sh.}},\theta_{\text{r.elb}},(\phi,\theta,\psi)_{\text{l.pelv.}},\theta_{\text{l.knee}},(\phi,\theta,\psi)_{\text{r.pelv.}},\theta_{\text{r.knee}}) \tag{2}$$

where t_x, t_y, t_z describe global translation and the joint angles describe body part orientation. A sequence of poses is represented by a combined parameter vector $\vec{q}_{\text{pose}}$ (Eq. (3)). A complete shape model $\vec{q}_{\text{shape}}$ is represented by the parameters for each superquadric (Eq. (4)). Note that the resulting transformation matrices H_i (for a joint i) depend both on pose and shape parameters, i.e. $H_i\equiv H_i(\vec{q}_{\text{shape}},\vec{p})$.

$$\vec{q}_{\text{pose}}=(\vec{p}_1,\vec{p}_2,\ldots,\vec{p}_k) \tag{3}$$

$$\vec{q}_{\text{shape}}=(\vec{s}_{\text{torso}},\vec{s}_{\text{neck}},\vec{s}_{\text{head}},\vec{s}_{\text{u.arm}},\vec{s}_{\text{l.arm}},\vec{s}_{\text{hand}},\vec{s}_{\text{u.leg}},\vec{s}_{\text{l.leg}},\vec{s}_{\text{foot}}) \tag{4}$$

Our overall parameter vector $\vec{q}=(\vec{q}_{\text{shape}},\vec{q}_{\text{pose}})$ is composed of the global shape parameters and the pose parameters for each selected time step.

To evaluate our objective function in the context of gradient-based optimization of pose and model parameters, we need to sample visible surface points of our model by sampling points on each superquadric. Linear sampling of the two angle parameters in Eq. (1) results in very unevenly spaced vertices, due to the varying local curvature of the object. We therefore use a first-order differential model [29] to obtain vertices with (almost) constant distance D_{vertex} on the superquadric surface. However, the number of vertices and their relative position on the surface will vary with the state of the shape parameter vector $\vec{q}_{\text{shape}}$. Section 3.4 describes how to integrate this property into the optimization process.

3.3. Frame selection

We now describe a method to perform frame selection, given a sequence of (multi-view) frames and respective initial pose estimates. By executing such a step prior to shape and pose adaptation, we not only increase efficiency (by limiting the number of frames processed), but also leave out frames that might be detrimental to the adaption process, as we will see in the experiments.

One important criterion for frame suitability is the error of the initial pose estimate. Of course, we do not know this error a-priori; we estimate it by evaluating a likelihood measure E_s based on the distance between occluding contour vertices on a model instance and the closest image edge features (see later Eq. (9) in Section 3.4). A simple and intuitive frame selection method would then be to choose a set $\mathbf{S}_K$ of K frames, where K is a constant set by the user, maximizing the mean value of E_s over the selected frames ("likelihood-based frame selection"):

$$p(\overline{E}_s|\mathbf{S}_K)\rightarrow\max,\quad\text{where}\quad\overline{E}_s=\frac{1}{K}\sum_{i=1}^{K}E_s^i \tag{5}$$

This turns out to be not ideal in a number of cases, the chief reason being that the likelihood might be high even though the underlying pose estimate is unreliable. This typically occurs for ambiguous poses with self-occlusions and limbs close to the body; uncertainty in the positions of the self-occluded limbs tends to be higher than uncertainty in the clearly visible and outstretched limbs. Another drawback is that the frames with the highest likelihoods tend to lie close together, therefore not yielding sufficient pose diversity for model adaptation. When selecting K frames, we want to capture as much information about the body shape as possible. Therefore, another important suitability criterion for frame selection is the diversity of the poses selected.

To take both criteria into account, we integrate the above mentioned $\overline{E}_s$ (i.e. representing the shape–pose likelihood, benefiting from a low value of K) and the combined contour distance E_{ccd} discussed below (i.e. representing the pose diversity, benefiting from a high value of K) in our frame selection algorithm. We model these as two independent random variables such that the joint likelihood is described as

$$p(E_{ccd},\overline{E}_s|\mathbf{S}_K)=p(E_{ccd}|\mathbf{S}_K)\times p(\overline{E}_s|\mathbf{S}_K) \tag{6}$$

We describe each of the likelihoods by an exponential distribution $\lambda e^{-\lambda x}$, such that minimizing the negative log-likelihood $-\log(p(E_{ccd}|\mathbf{S}_K))-\log(p(\overline{E}_s|\mathbf{S}_K))$ is equivalent to minimizing a cost function

$$E_{sel}=E_{ccd}+\tau\overline{E}_s\rightarrow\min \tag{7}$$

where τ is a weighting factor. We will iterate over several values of K in the experiments.

The combined contour distance E_{ccd} measures the average surface distance of vertices on the superquadric to the closest vertex that lies on the occluding contour at any of the selected frames and camera viewpoints in $\mathbf{S}_K$. By minimizing this distance, the model surface area that lies close to the occluding contours, i.e. where direct measurements are available, is maximized, and as such, pose diversity is increased.

To compute the combined contour distance, the first step is to determine the subset of occluding contour vertices among all sampled vertices, i.e. we determine which vertices are part of the outer projection rim. Two neighboring vertices are defined as being part of the occluding contour, if there is a sign change in the dot product between surface normal and a ray through vertex and camera center. To obtain only the outer rim, we exclude vertices that are occluded by other body parts in our model by making use of the inside–outside function of a superquadric [20]. All sampled surface

vertices form a 2D discrete space with neighborhood relations based on the sampling distance D_{vertex} (see Section 3.2).

We map the surface manifold of the vertex space of each body part onto a binary "feature" image and enable those pixels which correspond to contour vertices; we can now evaluate the surface distance of a vertex to its closest contour vertex efficiently, by computing the distance transform image (chamfering) and accessing the appropriate pixel. See Fig. 2 for a schematic illustration, and see Fig. 3 for an example with actual image data, before and after frame selection. In this case, the unit size (1 pixel) in the vertex space is $\approx D_{\text{vertex}}$.

Algorithm 1 formalizes the evaluation of E_{ccd}, given a candidate frame subset $\mathbf{S}_K$. In the algorithm, $\vec{a}$ is a vector that stores the minimum distances for each vertex, while $\vec{d}_k$ is a distance map for frame k. For a set of several images, the distance maps are combined using a 'min' operator to determine the distance to the closest occluding contour. In lines 3–9, we first combine the contributions over each view, for each multi-view frame. Lines 10–12 describe the combination across the set of selected frames.

Algorithm 1. Evaluation of average distance to contour vertices

Input: Subset of K frames $\mathbf{S}_K$; model estimate $\vec{q}_{\text{shape}}$; pose estimates $\vec{p}_k$ (for each image)
Output: Average distance E_{ccd} of all model surface vertices to closest contour vertex over $\mathbf{S}_K$

```
1  Let ā and d⃗_k, k ∈ K be vectors of dimension |V|, where V is the
   set of all model surface vertices;
2  ā ← ∞;
3  foreach frame I_k ∈ S_K do
4  |  d⃗_{I_k} ← ∞;
5  |  Project model/pose estimate to each view of I_k;
6  |  foreach view c of I_k do
7  |  |  foreach vertex v ∈ V do
8  |  |  |  Compute distance d_{v,c} to closest contour vertex;
9  |  |  |  d_{I_k,v} ← min(d_{I_k,v}, d_{v,c});
10 foreach frame I_k ∈ S_K do
11 |  foreach vertex v ∈ V do
12 |  |  a_v ← min(a_v, d_{I_k,v});
13 return E_ccd ← mean(ā)
```

We still need to determine the image subset $\mathbf{S}_K$ that minimizes Eq. (7). Considering only the minimization of E_{ccd} (disregarding the term including E_{obj}^{mean}), a simplified version of this problem in which merely the number of combined contour vertices is to be maximized (i.e. not regarding distances) is a variant of the NP-hard maximum coverage problem. To find the global minimum or a close solution for the complete expression of Eq. (7), we use stochastic optimization in the form of Simulated Annealing [23]. Starting from a random frame selection $\mathbf{S}_K$, candidate moves (i.e. the exchange of one frame in the set with a randomly selected non-set frame) from selected frames $\mathbf{S}_K$ to $\widehat{\mathbf{S}}_K$ with cost difference ΔE_{sel} are accepted according to a probability $(1+e^{\Delta E_{sel}/T})^{-1}$ where T is a temperature parameter which is adjusted according to an exponential cooling schedule.

For our experiments (Section 4), the weighting parameter τ in Eq. (7) was determined experimentally on a validation data set and subsequently set to 0.6. One evaluation of Eq. (7) can be executed very efficiently, because both the distance maps $d_{I_k,v}$ in line 9 of Algorithm 1 (for evaluation of E_{ccd}) as well as the costs E_s for evaluation of $\overline{E}_s$ can be precomputed. In our experiments, we perform 25,000 iterations; this takes $\sim$2 s.

3.4. Gradient-based pose and model optimization

We now define an objective function that measures the quality of the model/poses fit using shape and texture cues, given the selected multi-camera input frames from Section 3.3 (in our case, three views per time step) and corresponding segmented edge features. Our objective function E_{obj} is a weighted linear combination of three terms, measuring shape fit, texture fit across images and taking into account anthropometric constraints.

$$E_{obj} = E_s + \lambda E_t + \mu E_p \rightarrow \min \tag{8}$$

The first term E_s measures the distance between occluding contour vertices on a model instance (see Section 3.3) and the closest image edge features. Let I be the set of input frames, C^i the set of camera views in frame i, and $V_{\text{rim}}^{i,c}$ the set of model vertices on the projected contour in view c of frame i. Furthermore, let $d_c(S^{i,c}, v)$ be a function that computes the distance between a vertex v and the closest edge in the feature image $S^{i,c}$, and P^c be the camera projection matrix of camera view c. The average distance over all camera views and frames is then computed as

$$E_s = \frac{1}{|I|}\sum_{i\in I}\frac{1}{|C|}\sum_{c\in C^i}\frac{1}{|V_{\text{rim}}^{i,c}|}\sum_{v\in V_{\text{rim}}^{i,c}} d_c(S^{i,c}, P^c \mathbf{H}^{v}\zeta^{v}) \tag{9}$$

$\mathbf{H}^{v}\zeta^{v}$ (see Section 3.2) denotes the 3D world coordinate position of vertex v after transformation along the articulated chain. We compute $d_c(S^{i,c}, v)$ by first pre-computing an oriented Chamfer distance transform for the feature image $S^{i,c}$ and then performing a look-up at the projected 2D position using bilinear interpolation. Gradients of Eq. (9) with respect to pose and shape parameters are as follows:

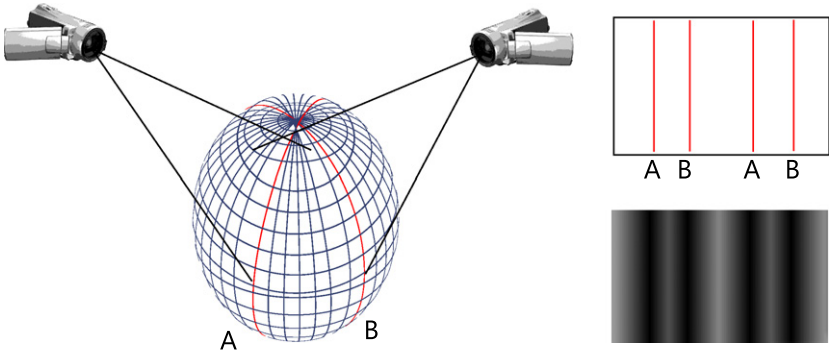

Fig. 2. Schematic illustration of the mapping of occluding contours of a quadric, as imaged by two cameras, onto a surface manifold. The top right image represents the binary feature image, where the occluding contours of the quadric correspond to pixels that are enabled. The bottom right image represents the corresponding distance image (larger distances are shown in a lighter gray shade).

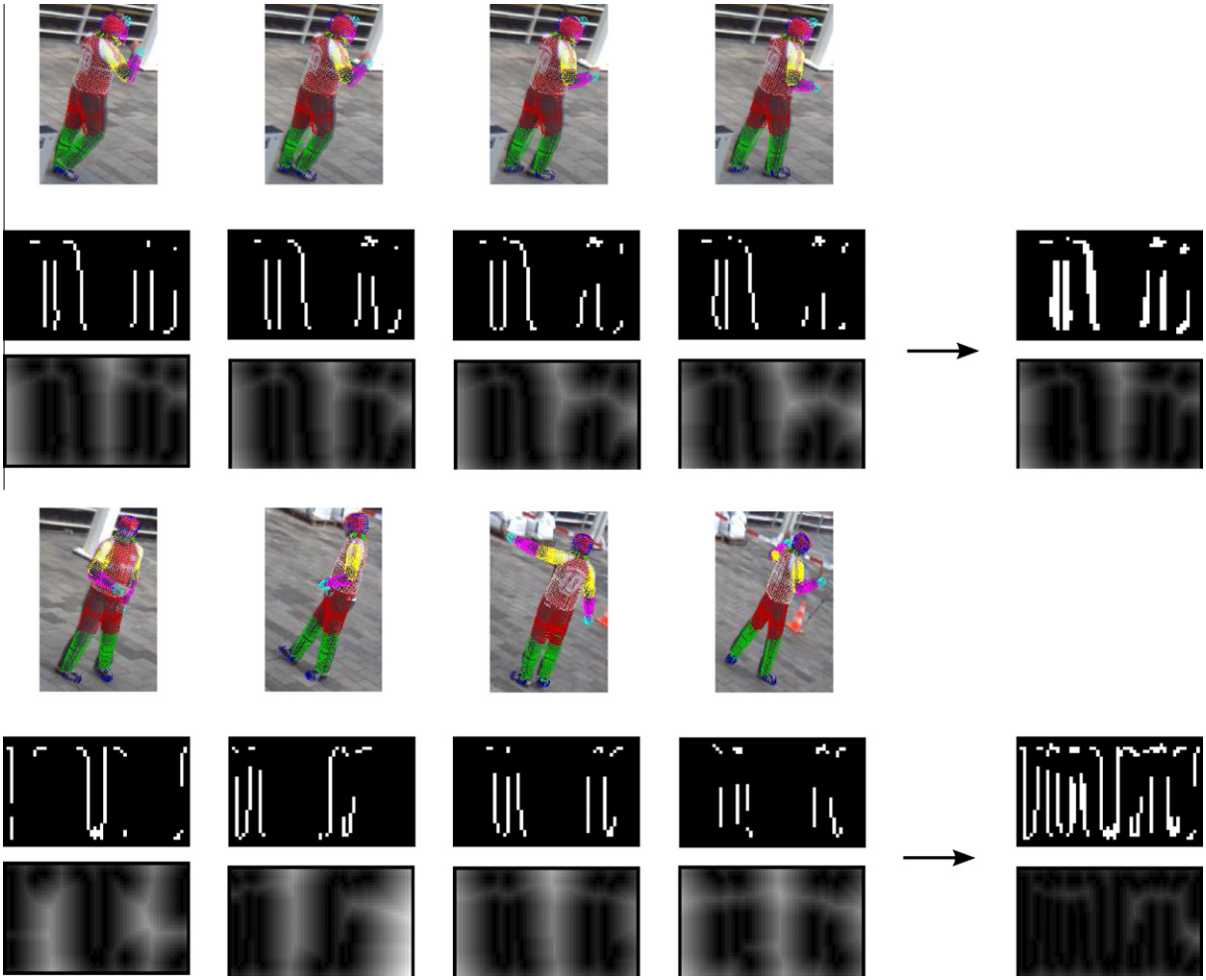

Fig. 3. Top: Initialization of frame selection algorithm with $K = 4$ consecutive frames. Bottom: Resulting selected frames after optimization. Below each image, we show an image of the torso surface space (occluding contours over all three cameras in white) and the respective distance image. See also Fig. 2. The resulting combined images are shown to the right. Note the changed distribution of the combined binary features after optimization. The mean grayscale value of the distance image after frame selection is significantly smaller (larger distances are shown in a lighter gray shade). It represents E_{ccd}; see also line 13 of Algorithm 1.

$$\frac{\partial E_s}{\partial \vec{q}_{\text{pose}}} = \frac{1}{|I||C|} \sum_{\substack{i \in I, \\ c \in C^i}} \frac{1}{|V^{i,c}_{\text{rim}}|} \sum_{v \in V^{i,c}_{\text{rim}}} J_c J_P J^{\text{pose}}_{H^v} \zeta^v \tag{10}$$

$$\frac{\partial E_s}{\partial \vec{q}_{\text{shape}}} = \frac{1}{|I||C|} \sum_{\substack{i \in I, \\ c \in C^i}} \frac{1}{|V^{i,c}_{\text{rim}}|} \sum_{v \in V^{i,c}_{\text{rim}}} J_c J_P \left(J^{\text{shape}}_{H^v} \zeta^v + \mathbf{H}^v J_{\zeta^v} \right) \tag{11}$$

Here, J_c is the 1×2 Jacobian of the bilinear lookup function $d_c(S^{i,c}, v)$ while J_P is the 2×3 Jacobian of the projection operation. $J^{\text{pose}}_{H^v}$ is the Jacobian containing the partial derivatives of the entries of the cumulative transformation matrix $\mathbf{H}^v$ w.r.t. the pose parameters; analogous for $J^{\text{shape}}_{H^v}$.

The second term E_t of Eq. (8) measures the average variance of the texture that all visible surface vertices V_{all} project to, across several time steps.

$$E_t = \frac{1}{|V_{\text{all}}|} \sum_{v \in V_{\text{all}}} \sum_{\text{h}=1}^{3} \left[\frac{1}{|I_v|} \sum_{i \in I_v} \left(t^h(P^i \mathbf{H}^v \zeta^v) - \frac{1}{|I_v|} \sum_{j \in I_v} t^h(P^j \mathbf{H}^v \zeta^v) \right)^2 \right] \tag{12}$$

For each vertex v of the model, we compute the texture variance over the set of frames and views I_v in which the vertex is visible. The function $t^h(\cdot)$ computes a bilinear texture look-up using a smoothed version of the original image; the variable h denotes a summing over all three respective RGB color channels. Fig. 4 provides a visualization of the variance measurement.

With the term described in Eq. (12) added to our objective function (Eq. (8)), both shape and pose measurements are linked across different time steps. In contrast to this, using only Eq. (9) as objective function, the pose parameters could be optimized independently, as the inner sums of the gradient (Eq. (10)) are non-zero for non-overlapping parts of the pose state vector $\vec{q}_{\text{pose}}$.

Again, we compute derivatives of Eq. (12) with respect to pose and shape parameters. Eq. (13) gives the derivative w.r.t. pose parameters; the derivative w.r.t. shape parameters is similar.

$$\frac{\partial E_t}{\partial \vec{q}_{\text{pose}}} = \frac{1}{|V_{\text{all}}|} \sum_{v \in V_{\text{all}}} \sum_{\text{h}=1}^{3} \left[2 \frac{1}{|I_v|} \sum_{i \in I_v} \left(t^h(P^i \mathbf{H}^v \zeta^v) - \frac{1}{|I_v|} \sum_{j \in I_v} t^h(P^j \mathbf{H}^v \zeta^v) \right) \left(\frac{\partial t^h(P^i \mathbf{H}^v \zeta^v)}{\partial \vec{q}_{\text{pose}}} - \frac{1}{|I_v|} \sum_{j \in I_v} \frac{\partial t^h(P^j \mathbf{H}^v \zeta^v)}{\partial \vec{q}_{\text{pose}}} \right) \right] \tag{13}$$

The terms $\frac{\partial t^h(P^i \mathbf{H}^v \zeta^v)}{\partial \vec{q}_{\text{pose}}}$ and $\frac{\partial t^h(P^i \mathbf{H}^v \zeta^v)}{\partial \vec{q}_{\text{shape}}}$ can be computed analogously to the derivatives of the bilinear lookup function in Eqs. (10) and (11); i.e. through application of the chain rule by multiplying the respective Jacobians.

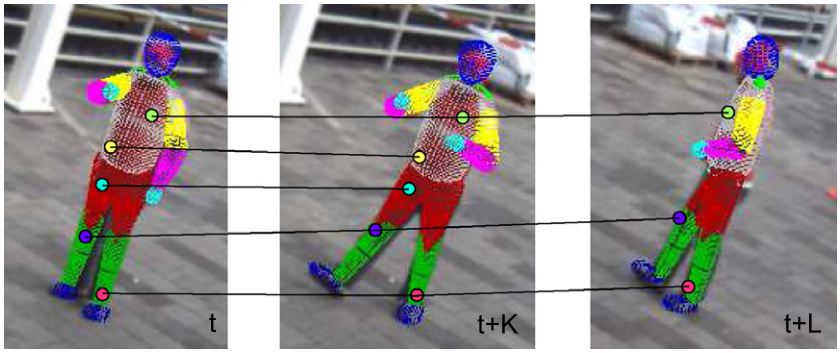

Fig. 4. Illustration of texture measurements at three time steps (e.g. selected frames). For each vertex, the texture variance is computed from measurements in frames where the respective vertex is visible.

The final term E_p of Eq. (8) adds penalty sub-terms, constraining the shape and pose parameters and preventing them from taking out-of-bound values.

$$E_p = \sum_{x \in \{\vec{q}_{\text{shape}}, \vec{q}_{\text{pose}}\}} \left(\frac{\max(x - u_x, l_x - x, 0)}{u_x - l_x} \right)^2 + \sum_{\vec{x} \subset \vec{q}_{\text{shape}}} \left(\max(0, d_M(\vec{x}, \mu_{\vec{x}}, \Sigma_{\vec{x}}) - 3) \right)^2 \quad (14)$$

where

$$d_M(\vec{x}, \mu, \Sigma) = ((\vec{x} - \mu)^T \Sigma^{-1} (\vec{x} - \mu))^{\frac{1}{2}}$$

The first part of Eq. (14) sums over each shape and pose parameter independently and adds a penalty if the respective parameter is outside of previously defined lower and upper bounds l_x and u_x. For example, the squareness parameters e_1 and e_2 of a superquadric are limited to the range $(0.0, \ldots, 1.0]$, etc. The second part of Eq. (14) sums over certain sets of shape parameter combinations and compares them to previously learned normal distributions (with mean $\mu_{\vec{x}}$ and covariance $\Sigma_{\vec{x}}$). Here, we only look at the lengths of certain limbs and penalize if the Mahalanobis distance d_M between current parameter vector and distribution is greater than 3. This allows us to enforce anthropometric constraints. We extracted measurements from a database of approx. 1500 people; in our case, we model the lengths of upper and lower arms and legs as a 2D normal distribution each, and the 3D extents of head and torso as a 3D normal distribution each. The derivative of Eq. (14) is

$$\frac{\partial E_p}{\partial \vec{q}} = \sum_{x \in \{\vec{q}_{\text{shape}}, \vec{q}_{\text{pose}}\}} \text{sign}(x - u_x) \frac{2}{u_x - l_x} \left(\frac{\max(x - u_x, l_x - x, 0)}{u_x - l_x} \right) + \sum_{\vec{x} \subset \vec{q}_{\text{shape}}} \left(\max(0, d_M(\vec{x}, \mu_{\vec{x}}, \Sigma_{\vec{x}}) - 3) \right) ((\vec{x} - \mu)^T \Sigma^{-1} (\vec{x} - \mu))^{-\frac{1}{2}} \times ((\vec{x} - \mu)^T \Sigma^{-T} + (\vec{x} - \mu)^T \Sigma^{-1})^T \quad (15)$$

We now address gradient-based local optimization strategies for our objective function (Eq. (8)). Quasi-Newton methods such as the L-BFGS algorithm [27] promise quadratic convergence under certain conditions. However, preliminary experiments using L-BFGS indicated frequent failure to converge to the global minimum, even with relatively close initializations. Because we sum over the number of visible vertices in Eqs. (9) and (12), both objective function and gradient are not smooth over the whole parameter space, despite being differentiable at each point (assuming vertex constancy). With every change in shape or pose parameters the set of selected vertices may change in a discrete manner, thus introducing discontinuities in the objective function. We are aware of approaches to explicitly model self-occlusions (e.g. [11]) in order to remove these discontinuities in pose tracking applications. However, integration with shape model adaptation under constantly changing model vertices is still unsolved. First-order optimization algorithms are better suited to cope with these noisy functions/gradients, and, while slower in comparison, we found that Gradient Descent (GD) successfully leads to convergence to the global minimum in a larger number of cases. Similar findings have been made in [5] on the problem of hand tracking from 3D measurements, and the authors propose a stochastic GD method with local step size adaptation that promises fast convergence and which we largely adopt. The GD update equation for a parameter vector $\vec{q}_i$ from iteration i to iteration $i+1$ is

$$\vec{q}_{i+1} = \vec{q}_i - \vec{a}_i \otimes \vec{g}_i \qquad \vec{g}_i = \frac{\partial E_{obj}(\vec{q}_i)}{\partial \vec{q}_i} \quad (16)$$

where $\otimes$ denotes a component-wise product and $\vec{a}_t$ is a step-size vector of local learning rates which is updated by a meta-level descent on the step-sizes.

$$\vec{a}_i = \vec{a}_{i-1} \otimes \exp(\mu_g \vec{g}_i \otimes \vec{v}_i) \qquad \vec{v}_{i+1} = \lambda_g \vec{v}_i + \vec{a}_i \otimes (\vec{g}_i - \lambda_g \vec{v}_i) \quad (17)$$

We choose suitable constant values for the meta-parameters ($\mu_g = 0.05$ and $\lambda_g = 0.6$). For efficiency reasons, we opt to update the gradient trace $\vec{v}_i$ as an exponential average of past gradients, as opposed to evaluating a multiplication of the Hessian with a vector [5] (the additional cost would be comparable to an additional gradient evaluation). We optimize iteratively over shape and pose parameters, spending a small but fixed amount of iterations for the optimization of each parameter set in a loop. The optimization process terminates either if $\|\vec{g}_i\| < \epsilon$, if $\max(\|\vec{q}_i - \vec{q}_{i-1}\|, \ldots, \|\vec{q}_i - \vec{q}_{i-K}\|) < \epsilon$, or after a maximum number of iterations has been reached.

The authors of [5] report that the stochasticity of their sampling approach helps avoid spurious minima, due to the local minima changing constantly. We randomly select $\frac{1}{3}$ of the previously sampled vertices for consideration in Eq. (8) at each iteration. Also, this optimization approach allows us to resample the model vertices using the constant-distance sampling described in Section 3.2 at each iteration, thus improving precision of the shape model representation under a changing parameter vector $\vec{q}_{\text{shape}}$. This makes the gradient function non-smooth, but the stochastic GD approach is able to tolerate this noise well. To further increase efficiency, we also modify the sampling step size D_{vertex} (see Section 3.2) during the optimization process in a coarse-to-fine approach, from $D_{\text{vertex}} = 6$ cm at the beginning to $D_{\text{vertex}} = 2$ cm near the maximum number of iterations.

Table 1
Quantitative evaluation of the pose input error of the frame selection process. Given is the average pose vertex error to ground truth over all poses in the selected frames after frame selection using the respective method.

	Mean avg. vertex error (cm) (std.dev. over trials)
Our frame selection	9.8 (1.1)
Likelihood-based selection	10.1 (1.4)
Random selection	13.7 (0.7)
Avg. vertex error (cm) of input poses (all frames) = 14.3 (std.dev. over *frames* 4.2)	

4. Experimental results

Our experimental data consists of recordings from three synchronized and calibrated color CCD cameras looking over a train station platform. In 13 sequences (S01–S13) of about 11 seconds on average (captured at 20 Hz), four actors (subjects P1–P4) perform unscripted movements such as waving, gesticulation and walking in front of a cluttered background.[1] Ground truth pose was manually labeled for all frames of the data set and for each of the persons; we estimate the ground truth accuracy to be within 4 cm. Furthermore, we used six sequences (S14–S19) from the HumanEva-I data set [35], for which good ground truth marker data was available (subjects P5, P6 == S1, S2 in HumanEva nomenclature; sequences 'Gestures', 'ThrowCatch', 'Walking' respectively; sequences resampled to 20 Hz).

We perform foreground segmentation for each frame using state-of-the-art background subtraction [41]. Because the segmented silhouettes are very noisy due to frequent lighting changes and people/trains moving in the background, we cannot rely only on silhouettes for edge features. Instead, we extract directed edge features (binary presence of edge, together with edge direction) from the input images using the Canny edge detector [7] and use the foreground silhouettes (after dilation with a 3×3 rectangular mask) as binary masks for the presence of relevant edge features.

4.1. Evaluation of frame selection and model shape and pose adaptation

To assess the properties of our proposed frame selection approach (Section 3.3) independently from the rest of the system, we performed a series of trials on a sequence of 200 frames and compare our approach to purely "likelihood-based" selection (as described in Section 3.3) and a random frame selection approach. In 50 separate trials, we added Gaussian noise to the ground truth pose and shape parameters and executed the different frame selection approaches for $K = 8$ frames. Table 1 summarizes the results by giving the average pose vertex error, i.e. the mean error between vertices of the model to the ground truth pose and model, averaged over all selected frames. We can see that random frame selection does not achieve a significant reduction of the error compared to the error over all frames of the sequence. In contrast, the likelihood-based and proposed frame selection approaches do result in selected frames with an approximately 4 cm lower input error. Fig. 5 illustrates the frame selection process for $K = 8$ on the above mentioned sequence, by means of the proposed approach (Section 3.3) and by means of the likelihood-based approach. As can be seen, the proposed approach (in green[2]), based on minimization of vertex distances to occluding contours, results in a more diverse set of poses selected, than the one based on the objective function (the latter would often pick several frames nearby the best matching one).

To assess the convergence properties of our model shape and pose adaptation approach (Section 3.4) in isolation, we also added different levels of Gaussian noise to our ground truth data of poses and associated shape models (in total we executed approx. 15,000 trials to account for varying K and varing input errors). The magnitude of perturbation was controlled independently for pose and shape parameters, i.e. creating initializations where either pose, shape or both parameter sets were perturbed. Evaluation consisted of plotting the average pose vertex error between the adapted parameter settings and the ground truth, for a set of selected frames and a given initial perturbation.

As a preliminary step, we experimentally determined optimized values for the weighting parameters λ and μ in Eq. (8) on a separate validation data set. Fig. 6 shows graphs for the best performing parameter combination ($\lambda_{opt} = 0.003$, $\mu_{opt} = 0.5$); one can see the beneficial influence of the texture term as well as the penalty term in Eq. (8) (i.e. comparing optimized value and value at zero).

Some example images depicting initial state and the result after running the model shape and pose adaptation are given in Fig. 7 (only a single view is shown). Fig. 8a shows a quantitative evaluation over all sequences, with a varying number K of frames used in the frame selection algorithm (Section 3.3). Standard deviations are given as bars for $K = 1$, 12. We can see that choice of K greatly influences optimization performance; using 4 and more multi-view frames improves results up to an average vertex input error of about 15 cm. Fig. 8b considers a subset for which only the shape parameters have been perturbed, thus removing the pose deviation influence. Convergence is quite stable, with the output error increasing only slowly with the input error. This is in part due to our penalty term in Eq. (8) which serves as a shape prior, but it is also apparent that using more input frames selected by our algorithm provides more information about the correct human shape.

Fig. 9a lists the convergence performance using the L-BFGS algorithm [27], using identical initial conditions and objective function as those used in Fig. 8a. As can be seen, performance of L-BFGS is generally worse, see also earlier discussion in Section 3.4. Fig. 9b shows the convergence performance using the likelihood described in [2] as objective function, whose contour term is similar to ours, but lacks a texture term and is non-differentiable. Because of the latter, it is optimized with a simplex method, as in [2]. Initial conditions are again identical to those used in Fig. 8a. As can be seen, performance of this approach is significantly worse; this is most likely due to the dimensionality of our state space which puts a challenge to the gradient-free optimization approach. Despite careful implementation, we were unable to obtain better results using the objective function and optimization approach of [2].

A batch optimization of eight frames (with three views each) requires about 20 s of processing time for the proposed approach (Fig. 8a), 15 s for the proposed objective function and L-BFGS (Fig. 9a) and 10 min for non-differentiable objective function and simplex method from [2], all in C++ code running on a 3 GHz Intel PC.

4.2. Evaluation of integrated approach

In order to realistically assess the benefits of frame selection and our overall approach, we now apply it in combination with the three-step approach described in Section 3.1 and Fig. 1a. For pose estimation given a model (generic or adapted), we are using a state-of-the-art (upper body) pose recovery system [19]. The

[1] Upon publication, the data set is made freely available for non-commercial research purposes to facilitate benchmarking.

[2] For interpretation of color in Figs. 1–10, the reader is referred to the web version of this article.

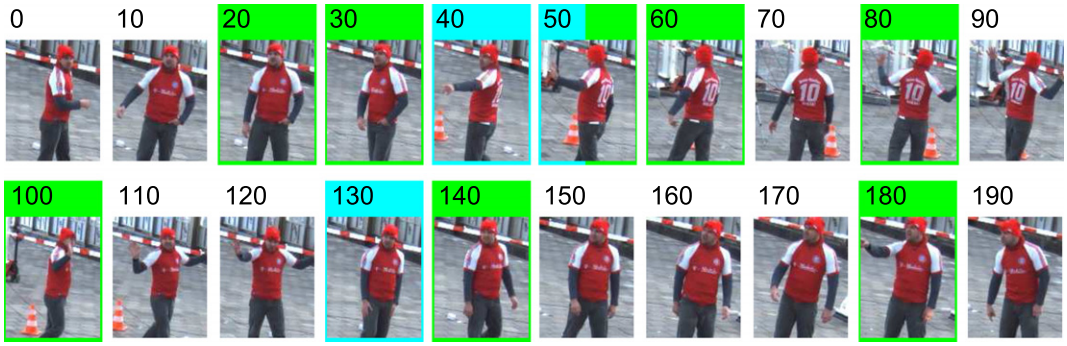

Fig. 5. Frame selection example for a sequence of 200 frames (every 10th frame, only one view shown). For $K = 8$, frames 22, 34, 48, 57, 80, 95, 141, 184 are selected by our frame selection algorithm (shown in green) and frames 40, 41, 43, 44, 45, 52, 54, 127 are selected using our likelihood criterion (shown in cyan). Only the closest frame multiple of 10 is colored in top rows. See Section 3.3.

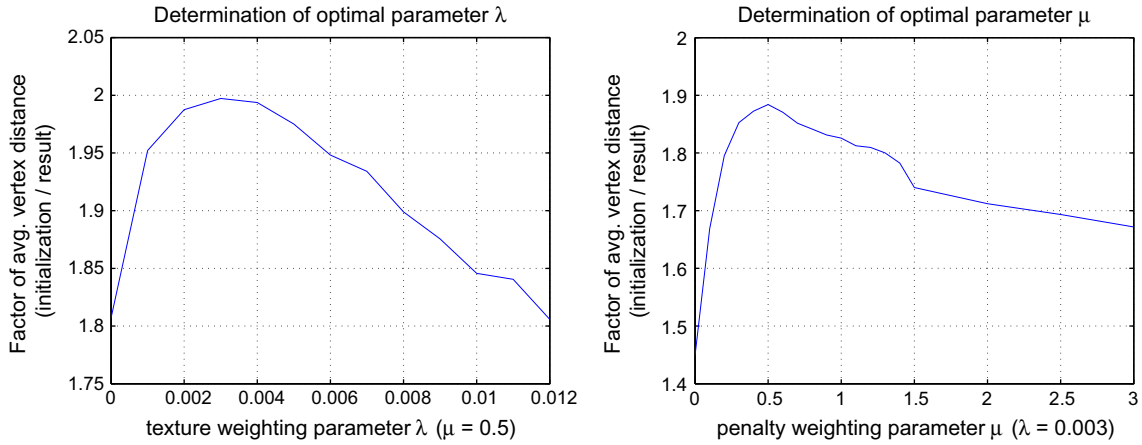

Fig. 6. Setting of optimized weighting parameters λ and μ (Eq. (8)). The y-axis shows the *factor* of the average pose vertex error between initialization and result (larger is better).

Fig. 7. Before and after model/pose adaptation. Reduction of avg. pose vertex error from 8.7 → 3.8, 16.1 → 5.0, 8.6 → 3.8, 8.8 → 3.8, 8.1 → 4.2, 10.1 → 4.9 (in cm, from left to right, 1st and 2nd row).

latter system uses a similar volumetric shape representation and utilizes a one-size-fits-all human shape model, i.e. the model is not adapted to the body shapes of the different persons. We opted for a pose recovery approach that follows a 'tracking as detection' paradigm and thus is able to automatically reinitialize upon temporary loss of track. This in turn increases the likelihood that

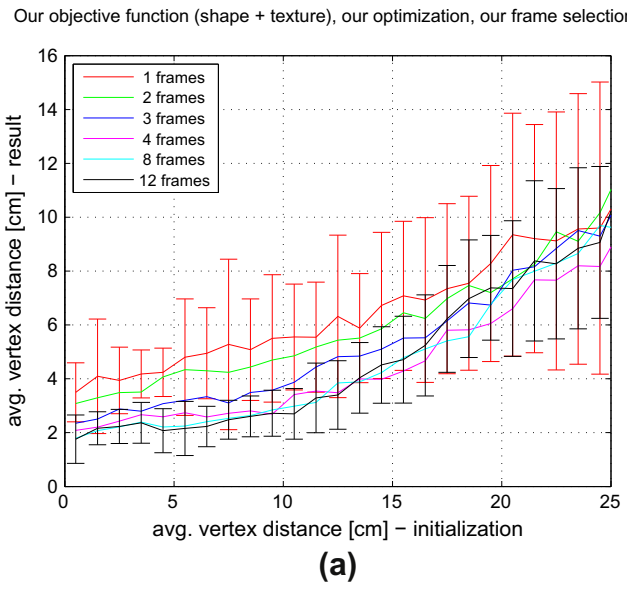

Fig. 8. Convergence performance of various configurations (see Section 4.1) (a) Our objective function and optimization. (b) Our objective function and optimization (shape perturbation only).

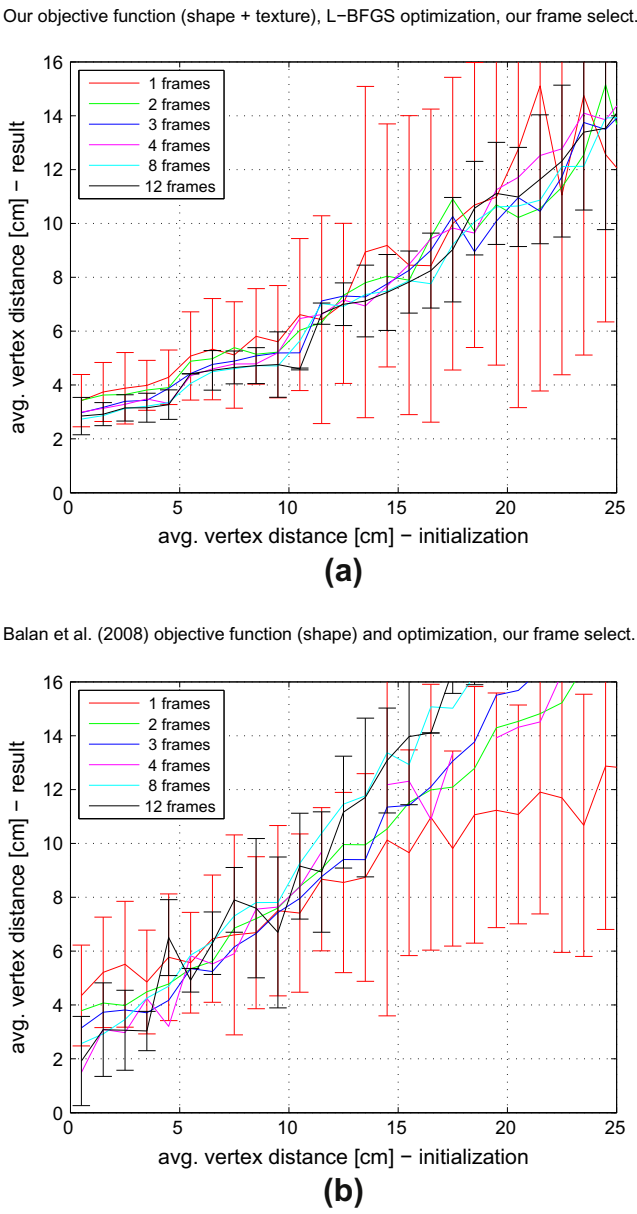

Fig. 9. Convergence performance of various configurations (see Section 4.1) (a) Our objective function and L-BFGS [27] optimization. (b) Objective function and optimization both from [2].

suitable poses for model adaptation are recovered within the error margin our adaptation algorithm can cope with.

Our three-step approach to automatic pose estimation works as follows: First, we recover pose over each sequence using a single "generic" (i.e. unadapted) model using [19]. Then, we use the recovered poses and the generic model as initializations for our shape and pose adaptation step (i.e. frame selection followed by optimization). Finally, we recover pose again with [19], this time using the adapted shape model.

We define the model vertex error as the mean error between vertices of a model to the ground truth model, both in a canonical pose. The measure captures the "distance" of a shape model to the ground truth shape model, irrespective of pose. Table 2 lists the model vertex error for the initial, generic model used for pose recovery (column "Generic"), and for the adapted models resulting from shape optimization using various frame selection variants (the proposed frame selection method "Our", likelihood-based frame selection "Likelih.", frame selection based on random choice "Random" and no frame selection "None", i.e. selection of all frames). Experiments were repeated six times per sequence; we list mean and standard deviation of the error. We observe that on average our proposed frame selection approach accounts for the largest reduction in model vertex error among all frame selection strategies. This is due to the selection of lower-error frames (cf. Table 1) in combination with a set of diverse poses that "explain" more of the shape model than randomly selected poses or poses selected solely on the basis of cost function ("likelihood"). We note that the average standard deviation using random frame selection is larger compared to the other approaches; this is due to the randomness of the selection process while the remaining deviations are mainly explained by the stochasticity of our optimization step.

Table 2
Quantitative results of shape model adaptation on 19 sequences (S14–S19 from HumanEva-I [35]). Given is the avg. model vertex error to ground truth (in cm) of the used generic model and the models obtained after executing the adaptation process using different frame selection approaches (six trials, standard deviation in brackets). See Section 4.2 for further details.

Seq.	Subj.	Generic	Our	Likelih.	Random	None
S01	P1	7.4	3.5 (0.2)	4.6 (0.3)	4.0 (0.5)	4.0 (0.2)
S02	P1	7.4	3.1 (0.3)	3.2 (0.3)	4.1 (0.6)	4.0 (0.3)
S03	P1	7.4	4.0 (0.2)	4.0 (0.1)	3.9 (0.6)	3.6 (0.1)
S04	P2	5.3	3.8 (0.1)	4.1 (0.3)	4.0 (0.4)	4.0 (0.1)
S05	P2	5.3	3.0 (0.2)	3.9 (0.3)	3.8 (0.5)	4.3 (0.4)
S06	P2	5.3	3.1 (0.4)	3.3 (0.2)	3.9 (0.7)	4.5 (0.1)
S07	P2	5.3	3.7 (0.3)	4.4 (0.3)	4.5 (0.4)	4.0 (0.1)
S08	P3	4.4	2.9 (0.1)	3.2 (0.2)	3.0 (0.4)	2.8 (0.1)
S09	P3	4.4	2.5 (0.1)	2.5 (0.3)	2.7 (0.3)	2.5 (0.1)
S10	P3	4.4	2.6 (0.1)	2.9 (0.3)	3.1 (0.3)	3.2 (0.1)
S11	P3	4.4	2.7 (0.2)	2.8 (0.3)	2.6 (0.3)	2.6 (0.1)
S12	P4	5.3	5.0 (0.3)	5.4 (0.3)	4.9 (0.5)	5.6 (0.3)
S13	P4	5.3	4.8 (0.4)	4.9 (0.2)	4.4 (0.3)	4.2 (0.1)
S14	P5	12.6	3.3 (0.2)	3.5 (0.2)	3.8 (0.3)	3.6 (0.1)
S15	P5	12.6	3.6 (0.2)	4.5 (0.2)	3.7 (0.3)	3.8 (0.3)
S16	P5	12.6	4.2 (0.3)	5.0 (0.2)	5.4 (0.8)	5.2 (0.4)
S17	P6	7.8	4.1 (0.1)	3.7 (0.2)	4.7 (0.5)	4.8 (0.5)
S18	P6	7.8	4.3 (0.3)	4.7 (0.1)	4.5 (0.7)	4.2 (0.1)
S19	P6	7.8	4.1 (0.2)	4.2 (0.1)	5.5 (0.6)	5.5 (0.1)
Mean		7.8	3.6 (0.2)	3.9 (0.2)	4.0 (0.5)	4.0 (0.2)

Table 3
Quantitative results of pose recovery on 19 sequences (S14–S19 from HumanEva-I [35]). Given is the pose error to ground truth (in cm), as specified in [19]. See Section 4.2 for further details.

Seq.	Subj.	Tr.Seq.	Labeled	Generic	Our	Likelih.	Random	None
S05	P1	S10	7.7	16.9	8.9	15.1	13.7	18.5
S10	P1	S05	7.7	15.5	9.2	9.7	11.4	12.4
S11	P1	S10	7.6	11.9	9.1	13.8	11.7	11.9
S14	P2	S12	7.2	11.8	7.9	9.3	10.5	17.6
S08	P2	S12	7.9	14.9	11.6	16.1	13.7	12.2
S12	P2	S08	7.6	16.7	8.1	16.0	11.2	15.2
S15	P2	S12	5.0	7.5	5.8	7.8	11.4	8.2
S01	P3	S09	9.7	12.9	13.9	13.9	13.2	12.5
S06	P3	S01	11.4	13.8	13.8	13.9	13.9	13.8
S09	P3	S06	9.2	10.0	9.8	10.5	9.9	9.8
S13	P3	S06	10.2	12.8	12.2	12.9	12.2	12.7
S02	P4	S16	9.1	11.9	11.4	11.9	13.4	12.6
S16	P4	S02	7.5	13.8	9.5	13.2	13.2	9.6
HE01	P5	HE02	6.3	16.8	6.4	6.5	6.9	6.7
HE02	P5	HE03	5.8	11.5	10.1	9.5	9.5	9.1
HE03	P5	HE01	4.7	15.6	8.9	9.1	10.6	10.6
HE04	P6	HE05	5.3	8.2	6.9	6.3	10.3	6.4
HE05	P6	HE06	5.0	8.6	6.0	5.8	5.4	5.8
HE06	P6	HE04	4.6	8.3	4.6	5.4	5.5	5.4
Mean			7.3	12.6	9.1	10.8	11.0	11.1

Table 4
Quantitative results of pose recovery, per body part, averaged over all sequences. Given is the pose error to ground truth (in cm) per body part, as specified in [19]. See Section 4.2 for further details.

	Torso	Head	L.Sh.	L.Elb.	L.Wr.	R.Sh.	R.Elb.	R.Wr.
Labeled	3.8	4.6	5.2	9.0	11.8	5.3	8.6	10.5
Generic	5.0	6.2	9.2	16.0	23.2	8.7	14.2	18.2
Our	4.3	5.1	6.9	11.5	16.6	6.7	10.0	11.8
Likelih.	4.7	5.4	8.2	13.8	19.5	7.9	12.3	15.1
Random	4.7	5.6	8.3	13.8	20.3	8.0	12.2	14.5
None	4.7	5.6	9.1	13.8	18.6	8.7	13.8	15.1

Tables 3 and 4 now quantify the results of the subsequent pose recovery step using the different adapted models. Here we give the pose error in cm as specified in [19]; i.e. the average deviation between recovered poses and ground truth at a set of joint locations (torso center, head center, left/right shoulders, elbows, wrists). Column "Tr.Seq." lists the respective training sequence used for initial pose recovery and model adaptation. Each sequence was chosen to be different from the test sequence while featuring the same subject. Column "Labeled" shows the pose error using the labeled (ground truth) model, while column "Generic" contains the pose error using the initial generic model. The remaining columns list the pose error after model adaptation using the different approaches to frame selection. Table 4 additionally lists the error per measured body part, averaged over all sequences of Table 3, for each model type. Errors are generally lowest close to the root of the articulated model and higher at the limb ends. Regarding the four choices of frame selection for model adaptation, the error increases in the order listed in the table.

As can be seen, pose recovery using the model obtained by the proposed frame selection method outperforms the simpler variants overall (compare column "Our" with "Likelih.", "Random" and "None"). Pose recovery using this model does not quite reach the error that is achievable with the labeled 'ground truth' model (column "Labeled"); we believe that the latter error is close to the achievable optimum using mentioned pose recovery system. We clearly outperform pose recovery using the generic model, and on average we can also improve on pose recovery using the model generated by likelihood-based selection, random frame selection as well as no frame selection.

Results are comparable, though much closer, on the HumanEva-I sequences compared to our data set. Here, our proposed frame selection is on one level with likelihood-based selection but we can still clearly outperform pose recovery using the generic model and pose recovery using the model generated by random frame selection. We attribute this to the different properties of the data sets; in contrast to our own data set, each HumanEva-I sequence features not more than one actor in a controlled indoor scenario (which allows for easier foreground segmentation). The camera is significantly closer to the subjects, as the area covered by all cameras is approx. 2×2 m, vs. 4×5 m for our data set. Consequently, the average errors achieved are lower over all model types. Tracking errors of 8.13 cm and 11.2 cm were reported in [28] for the respective walking sequences S16 and S19 using Annealed Particle Filtering [12]; our error on these sequences is somewhat lower (8.9 cm and 4.6 cm). Further literature listing errors on either HumanEva-I and II data sets include [6,9], where errors of approx. 8 cm are reported. [14,24] report significantly lower errors around 3–4 cm; in the latter paper, a strong motion model for walking motions is enforced. Comparison remains difficult due to the comparison between upper body vs. full body pose recovery, frame-sequential tracking systems vs. systems that include initialization, systems using weak vs. strong motion priors, differing frame rates, and pose recovery systems optimized for varying levels of detail.

Overall, the results presented in Table 3 can be seen as a direct result of the improved model vertex errors observed in Table 2. Fig. 10 shows some example frames of recovered poses in different sequences using the "generic" model and the model adapted using the proposed frame selection.

Our current approach still has some limitations, which we plan to address in future work. While our system is designed to adapt the shape of one person, extension to multiple people in one frame can be achieved by executing the adaptation multiple times. However, we currently do not model occlusion of the subject by objects or other persons. Experiments indicate that the system can deal with a minor amount of occlusion, as long most selected frames are not affected. However, there is a need to incorporate prior

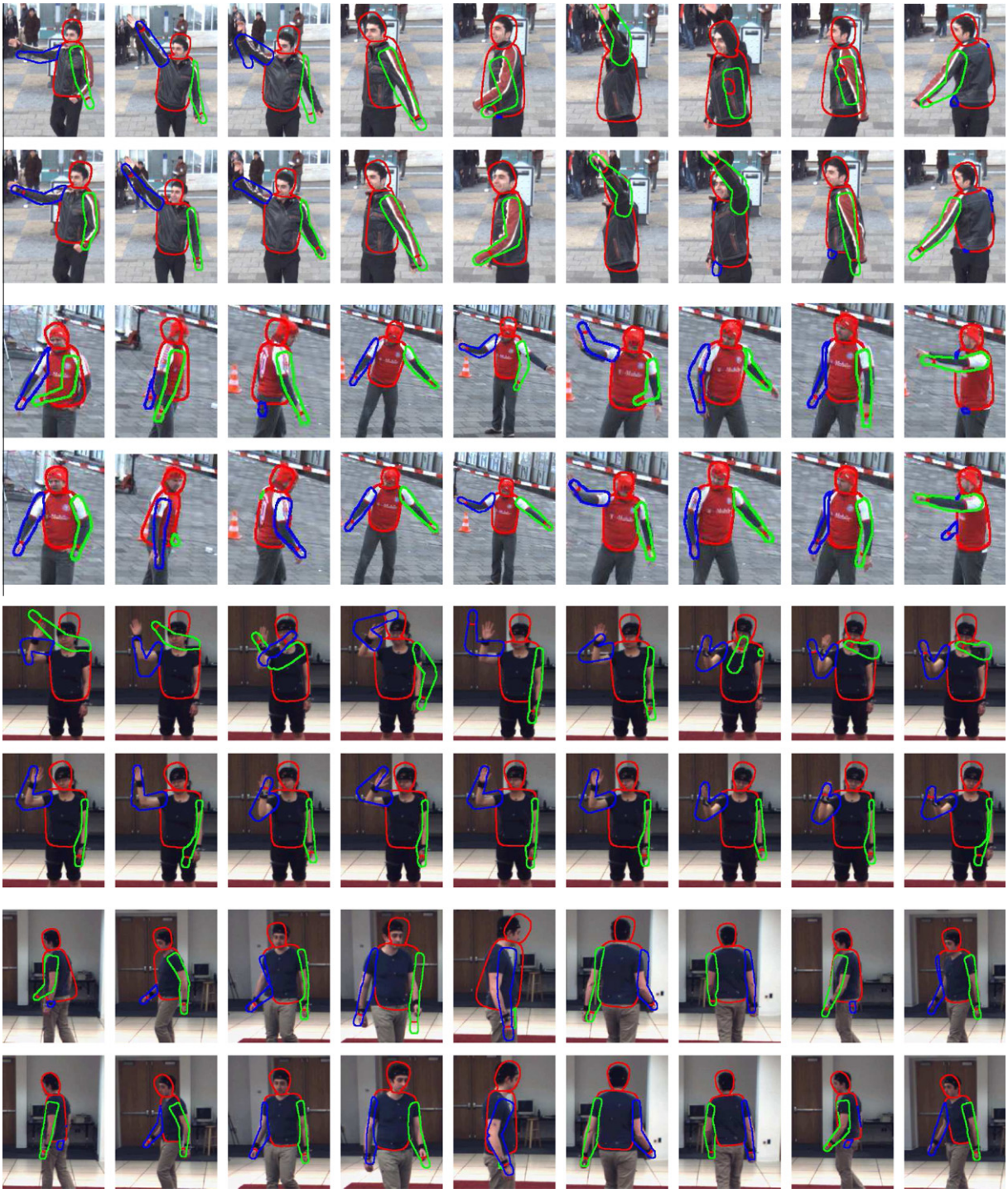

Fig. 10. Example frames from the pose recovery evaluation; shown are sequences S02, S06, S14, S19. Top: Generic model. Bottom: Optimized model (Our algorithm).

knowledge of occlusions directly into the optimization process when intending to handle sequences with groups of people. Also, our adaptation approach currently is not integrated into the pose recovery process "on-line", i.e. a complete pose recovery with an adapted model is split into three stages as outlined by Fig. 1a. A simple way to integrate shape adaptation into pose recovery could be to perform frame selection after each newly recovered pose and to then adapt the model if the minimum frame selection cost (see Eq. (7)) changes. We intend to investigate in future work how this additional feedback loop affects performance of both pose recovery and shape adaptation.

5. Conclusion

We presented an efficient method for 3D human shape and pose adaptation that addresses both the selection of a suitable set of input frames and an adaptation step. The latter is carried out as stochastic gradient-based optimization using an objective function based on shape and texture cues; we showed that the proposed optimization approach can handle input pose errors up to 15 cm well. In a series of experiments, we demonstrated that the main components of our approach outperform the state-of-the-art. In particular, regarding the proposed frame selection step, we made

the point that frame selection not only improves system efficiency (by processing a subset of frames), it can also lead to a better behaved optimization; in other words: it is quality, not size, that matters.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cviu.2011.08.002.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, J. Davis, SCAPE: shape completion and animation of people, ACM Transactions on Graphics 24 (2005) 408–416.

[2] A. Balan, M. Black, The naked truth: Estimating body shape under clothing, in: Proceedings of the European Conference on Computer Vision, 2008, pp. II: 15–29.

[3] A. Balan, L. Sigal, M. Black, J. Davis, H. Haussecker, Detailed human shape and pose from images, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[4] L. Ballan, G.M. Cortelazzo, Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes, in: Fourth International Symposium on 3D Data Processing, Visualization and Transmission, 2008.

[5] M. Bray, E. Meier, N. Schraudolph, L. Van Gool, Fast stochastic optimization for articulated structure tracking, Image and Vision Computing 25 (2007) 352–364.

[6] M.A. Brubaker, D.J. Fleet, A. Hertzmann, Physics-based person tracking using the anthropomorphic walker, International Journal of Computer Vision 87 (2010) 140–155.

[7] J. Canny, A computational approach to edge detection, in: RCV87, 1987, pp. 184–203.

[8] G. Cheung, S. Baker, T. Kanade, Shape-from-silhouette across time – Parts I and II, International Journal of Computer Vision 62 and 63 (2005) 221–247. 225–245.

[9] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, T.P. Andriacchi, Markerless motion capture through visual hull, articulated ICP and subject specific model generation, International Journal of Computer Vision 87 (2010) 156–169.

[10] E. de Aguiar, C. Theobalt, C. Stoll, H.P. Seidel, Marker-less deformable mesh tracking for human shape and motion capture, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[11] M. de la Gorce, N. Paragios, D.J. Fleet, Model-based hand tracking with texture, shading and self-occlusions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[12] J. Deutscher, I. Reid, Articulated body motion capture by stochastic search, International Journal of Computer Vision 61 (2005) 185–205.

[13] D.A. Forsyth, O. Arikan, L. Ikemoto, J. OBrien, D. Ramanan, Computational studies of human motion, Foundations and Trends in Computer Graphics and Vision 1 (2005) 77–254.

[14] J. Gall, B. Rosenhahn, T. Brox, H.-P. Seidel, Optimization and filtering for human motion capture, International Journal of Computer Vision 87 (2010) 75–92.

[15] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, H.-P. Seidel, Motion capture using joint skeleton tracking and surface estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1746–1753.

[16] K. Grauman, G. Shakhnarovich, T.J. Darrell, A Bayesian approach to image-based visual hull reconstruction, in: Proceedigns of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR, 2003), pp. I: 187–194.

[17] P. Guan, A. Weiss, A.O. Balan, M.J. Black, Estimating human shape and pose from a single image, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2009.

[18] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormaehlen, H.-P. Seidel, Estimating body shape of dressed humans, in: Shape Modeling International, 2009.

[19] M. Hofmann, D.M. Gavrila, Multi-view 3d human pose estimation combining single-frame recovery, temporal integration and model adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2214–2221, 2009.

[20] A. Jaklic, A. Leonardis, F. Solina, Segmentation and Recovery of Superquadrics, Kluwer, 2000.

[21] I. Kakadiaris, D. Metaxas, Three-dimensional human body model acquisition from multiple views, International Journal of Computer Vision 30 (1998) 191–218.

[22] R. Kehl, L. Van Gool, Markerless tracking of complex human motions from multiple views, Computer Vision and Image Understanding 103 (2006) 190–209.

[23] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.

[24] C.-S. Lee, A.M. Elgammal, Coupled visual and kinematic manifold models for tracking, International Journal of Computer Vision 87 (2010) 118–139.

[25] I. Mikic, M.M. Trivedi, E. Hunter, P.C. Cosman, Human body model acquisition and tracking using voxel data, International Journal of Computer Vision 53 (2003) 199–223.

[26] T. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 103 (2006) 90–126.

[27] J. Nocedal, Updating quasi-Newton matrices with limited storage, Mathematics of Computation 35 (1980) 773–782.

[28] P. Peursum, S. Venkatesh, G. West, A study on smoothing for particle-filtered 3d human body tracking, International Journal of Computer Vision 87 (2010) 53–74.

[29] M. Pilu, R.B. Fisher, Equal-distance sampling of superellipse models, in: Proceedings of the British Machine Vision Conference (BMVC), 1995.

[30] R. Plaenkers, P. Fua, Articulated soft objects for multiview shape and motion capture, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 1182–1187.

[31] E. Poon, D.J. Fleet, Hybrid monte carlo filtering: Edge-based people tracking, in: Motion, 2002, pp. 151–158.

[32] D. Ramanan, D.A. Forsyth, A. Zisserman, Tracking people by learning their appearance, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 65–81.

[33] B. Rosenhahn, U.G. Kersting, K. Powell, R. Klette, G. Klette, H.P. Seidel, A system for articulated tracking incorporating a clothing model, Machine Vision and Applications 18 (2007) 25–40.

[34] L. Sigal, A.O. Balan, M.J. Black, Combined discriminative and generative articulated pose and non-rigid shape estimation, in: J.C. Platt, D. Koller, Y. Singer, S.T. Roweis (Eds.), NIPS, MIT Press, 2007, p. 2007.

[35] L. Sigal, A. Balan, M. Black, HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, International Journal of Computer Vision 87 (2010) 4–27.

[36] C. Sminchisescu, Consistency and coupling in human model likelihoods, in: FGR, IEEE Computer Society, 2002, pp. 27–32.

[37] C. Sminchisescu, B. Triggs, Estimating articulated human motion with covariance scaled sampling, The International Journal of Robotics Research 22 (2003) 371–392.

[38] J. Starck, A. Hilton, Model-based multiple view reconstruction of people, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2003, pp. 915–922.

[39] N. Ukita, R. Tsuji, M. Kidode, Real-time shape analysis of a human body in clothing using time-series part-labeled volumes, in: Proceedings of the European Conference on Computer Vision, 2008, pp. III: 681–695.

[40] D. Vlasic, I. Baran, W. Matusik, J. Popovic, Articulated mesh animation from multi-view silhouettes, ACM Transactions on Graphics 27 (2008).

[41] Z. Zivkovic, Improved adaptive Gaussian mixture model for background subtraction, in: International Conference on Pattern Recognition, 2004, pp. 28–31.