HAL
open science

Robust Human Body Shape and Pose Tracking

Chun-Hao Huang, Edmond Boyer, Slobodan Ilic

To cite this version:

Chun-Hao Huang, Edmond Boyer, Slobodan Ilic. Robust Human Body Shape and Pose Tracking. 3DV - International Conference on 3D Vision - 2013, Jun 2013, Seattle, United States. pp.287-294, 10.1109/3DV.2013.45 . hal-00922934

HAL Id: hal-00922934

https://inria.hal.science/hal-00922934

Submitted on 31 Dec 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Robust Human Body Shape and Pose Tracking

Chun-Hao Huang
CAMP
TU München
huangc@in.tum.de

Edmond Boyer
INRIA Rhône-Alpes
Grenoble
edmond.boyer@inrialpes.fr

Slobodan Ilic
CAMP
TU München
slobodan.ilic@in.tum.de

***Abstract*—In this paper we address the problem of markerless human performance capture from multiple camera videos. We consider in particular the recovery of both shape and parametric motion information as often required in applications that produce and manipulate animated 3D contents using multiple videos. To this aim, we propose an approach that jointly estimates skeleton joint positions and surface deformations by fitting a reference surface model to 3D point reconstructions. The approach is based on a probabilistic deformable surface registration framework coupled with a bone binding energy. The former makes soft assignments between the model and the observations while the latter guides the skeleton fitting. The main benefit of this strategy lies in its ability to handle outliers and erroneous observations frequently present in multi-view data. For the same purpose, we also introduce a learning based method that partition the point cloud observations into different rigid body parts that further discriminate input data into classes in addition to reducing the complexity of the association between the model and the observations. We argue that such combination of a learning based matching and of a probabilistic fitting framework efficiently handle unreliable observations with fake geometries or missing data and hence, it reduces the need for tedious manual interventions. A thorough evaluation of the method is presented that includes comparisons with related works on most publicly available multi-view datasets.**

***Keywords*-human motion capture; non-rigid surface deformation; pose estimation**

I. Introduction

Marker-less human motion capture from multiple camera videos is a fundamental task in many applications including sport science, movie industry, and medical diagnostics. Since human motion is defined by both articulated motion and surface deformation they should ideally be estimated simultaneously. However, this requires sophisticated physics-based models that capture the real relationships between pose and shape. Since such models are hard to build and also involve complex parametrization, researchers often decouple them and treat each problem separately. One line of approaches considers only the estimation of surface deformations by fitting a reference model to the incoming image observations, e.g. [5], [6], [8], [9]. Another line of approaches parameterizes the model deformations with an articulated human skeleton represented as a kinematic chain [10], [13], [18], [21]. While the latter are less generic and strongly depend on the skeleton parametrization, the former are more generic and require less priors, hence allowing for larger classes of model deformations. Since the human anatomical structure can not be perceived by traditional visual sensors such as color cameras, approaches that model and track shapes instead of internal and unobserved skeletons tend to give more reliable results with visual data. Nevertheless, in many graphical applications that involve human body models, the pose is required as much or more than the shape surface. To this objective, we introduce a method that simultaneous recovers both the shape surface, in the form of a mesh, and its pose with articulated skeleton parameters. This method builds on two related works. First, the patch-based deformable surface registration framework proposed in [6] that relies on soft observation assignments and handles outliers. Second, the bone binding energy presented in [19] that forces the skeleton model to stay inside the deformed human body shape. The combination of these two strategies allows us to devise an approach that benefits from a robust surface registration when recovering human body pose and without the need for complex inverse kinematic parametrizations. Furthermore, in order to reduce complexity and to better handle erroneous observations, we investigate a learning based strategy that partition the input data into rigid body parts as learned from the reference model. This strategy limits the search when assigning observations to the model as well as enabling a better discrimination between inliers and outliers in the input data.

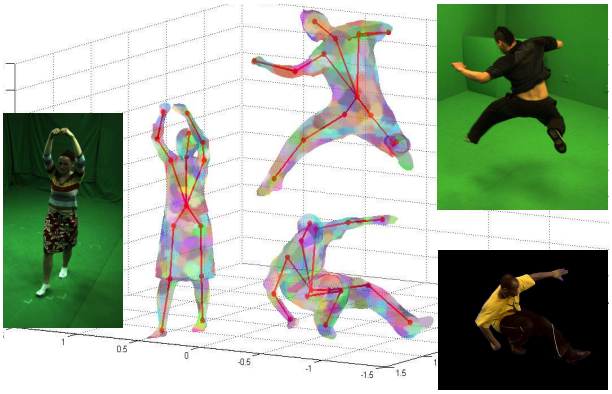


Figure 1. Our approach tracks both the shape and the pose of humans simultaneously. Results with three different standard datasets are shown above. Left: Skirt in [10]. Top right: Bouncing in [21]. Bottom right: Free in [17].

This paper has several contributions. Different from [6] or [10], pose and shape are recovered at the same time. Second, a robust framework is presented that improves over [19] by combining probabilistic assignments between the model and the observations with a Support Vector Machine (SVM) based classification that partitions observations so that they can be exploited more efficiently. Third, a thorough evaluation with various public dataset is presented [10], [15], [17], [21]. This evaluation validates the effectiveness of the proposed method.

The rest of this paper is organized as follows. In Section II we review the most relevant related work. Details of the proposed method are described in Section III. Validation experiments and results are provided in Section IV, and we conclude the paper in Section V.

II. Related work

Human motion tracking/capturing has been long studied in both computer vision and graphic communities. Based on the way of parameterizing motion, existing works can be categorized into three classes:

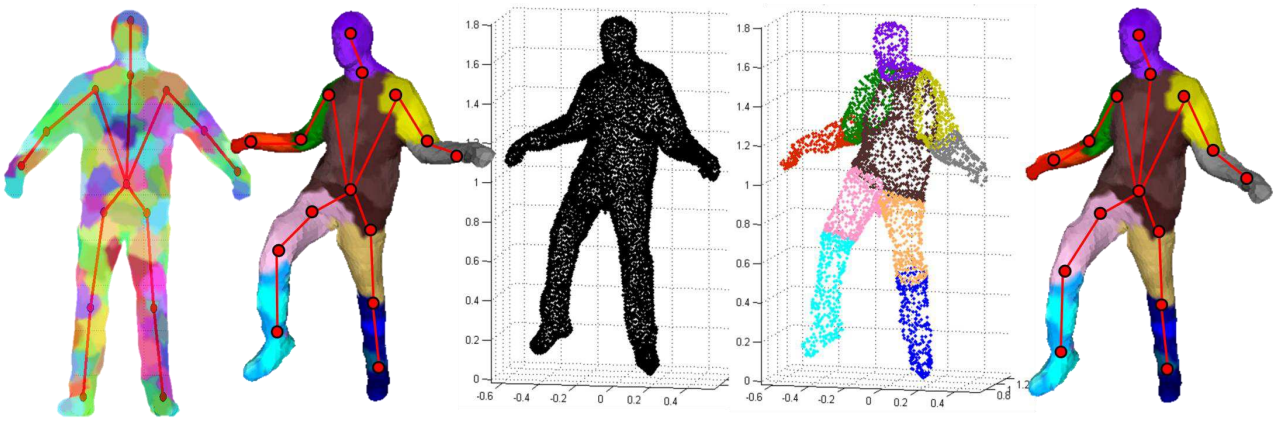
Mesh-based approaches: In this class of methods, motion is solely parameterized on the humanoid surface which evolves in time, without incorporation of a skeleton model. Authors usually introduce some constraints among vertices such that implausible deformations are avoided. Aguiar *et al.* [9] propose a scene-flow-based deformation scheme. To overcome the accumulated flow estimation error, they utilize Laplacian deformation framework [4] as a refinement step. In their follow-up work [8], they first deform a low-resolution tetrahedral mesh to roughly estimate the pose, and then transfer it to a high-resolution scanned model. Surface details are again preserved by Laplacian constraint. Cagniart *et al.* [5] advocate to divide the mesh into small cells called *patches*. A rigidity constraint is imposed among neighboring patches which smooths model deformation. In [6], they further improve the data term and the whole deformation framework acts like a probabilistic iterative closest point (ICP) approach. The advantage of these purely-mesh-based methods is that they can generalize to non-humanoid surface tracking, and they better handle non-rigid deformation such as loose apparel.

Skeleton-based approaches: Since human motion is highly articulated, many authors use skeleton-based models. Motion is then parameterized in a low-dimensional pose parameter space. However, in the observations, whether 3D point clouds or silhouettes, one does not observe the skeleton directly. A mesh surface is still needed for the fitting purpose but it is controlled by the underlying skeleton. As a result, skeleton plays the role of prior deformation model. From this point of view it is actually much more constrained than purely-mesh-based methods. There are mainly two concerns in this family of work: first, how to parameterize motions in terms of the skeleton, and second, how this skeleton should interact with the reference mesh. Vlasic *et al.* [21] parameterize motions as transformation of local coordinate of each joint. Vertex transformation is computed by the linear combination of different joint transformations, known as linear blend skinning [13]. With similar parameterizations, Gall *et al.* [10] adopt quaternion blend skinning [12] which produces less artifacts. In both methods, the skeleton acts as a kinematic chain where local transformations are transferred from the parents to the children. Energies between mesh and observations are defined in pose parameter space, based on the simple assumption that surface deformation is explained only by the skeleton. A second stage of surface refinement is usually required.

Hybrid approaches: The first category of approaches emphasizes more on the surface consistency, whereas the second category of approaches focuses on the pose. Straka *et al.* [19] advocate the integration of both categories into one energy function. They introduce differential bone coordinates as an implicit skinning approach, and therewith they formulate a skeleton-binding energy term defined on the parameters of both mesh surface and skeleton. This allows them to jointly estimate pose and shape, and they show that optimizing in this coupled space results in more robustness. Moreover, skeletons are parameterized only in terms of joint location. Although losing some rotational degree of freedom (DoF) for each joint, this makes the energy term quadratic in terms of both, body joint positions and mesh vertex positions. Therefore, the optimal solution can be obtain via standard optimization method. The difference of our approach compared to [19] is that we compute the bone energy per patch rather than per vertex. In addition, our observations are 3D visual hull reconstructions instead of 2D silhouettes. With 3D information, we are able to handle ambiguous situations that one cannot do with only image observations. Furthermore, we partition observations into body part regions according to learned partitioning in the previous frames. This allows us more efficient matching of the reference mesh and the input 3D observations, which combined with optimization with soft assignments from [6] makes it more robust to outliers and missing data.

III. Method

To facilitate human motion tracking in synchronized and calibrated multi-view sequences, a 3D point cloud $\mathcal{T}^t$ is first reconstructed using silhouette observations. Our body model deforms according to these observations on the frame-to-frame basis. The model comprises a reference triangle mesh surface $\mathcal{M}$ and an intrinsic tree-structured skeleton. We adopt the patch-based mesh deformation model proposed in [5]. In this framework, vertices are grouped into N_P patches and deformation of $\mathcal{M}$ is parameterized in terms of $\mathbf{\Theta} = \{(\mathbf{R}_k, \mathbf{t}_k)\}_{k=1}^{N_P}$ where $\mathbf{R}_k$ and $\mathbf{t}_k$ are rotation and translation of patch P_k respectively. Our skeleton is a set of N_J 3D joint coordinate positions $\mathbf{J} = \{\mathbf{x}_j\}_{j=1}^{N_J}$ where N_J



(a) Patched mesh and skeleton (b) Model in frame t (c) Observed point cloud of frame t+1 (d) Partitioned point cloud (e) Estimated model

Figure 2. Illustration of our pipeline. In (b), patches attached to same joint are encoded in similar colors. Incoming observations (c) are partitioned and subsampled into (d) by the SVM classifier trained with (b). By minimizing Eq. (1) define between (b) and (d), the model deforms as in (e). In this example, we reduce the amount of target point from 7823 (c) to only 3682 (d).

is 15. The root of the tree is set at the pelvis, as in Figure 2 (a). Parameterizing directly on their position leads to a quadratic energy term that keeps optimization feasible [19]. The skeleton is manually rigged to the mesh. Each vertex v is associated with a non-leaf-node joint that has the largest skinning weight from [3]. By taking the majority vote, each patch is also associated to a joint as in Figure 2 (b). This association is fixed throughout the whole sequence and is used as the rigid body part label to classify the incoming observations.

Therefore, given a model which is properly registered in the first frame, the remaining task is to determine how the mesh and skeleton deform based on every $\mathcal{T}^t$. We approach this by minimizing a energy function defined as:

$$E(\mathbf{\Theta}, \mathbf{J}) = \lambda_r E_r(\mathbf{\Theta}) + \lambda_d E_{data}(\mathbf{\Theta}) + \lambda_b E_{bone}(\mathbf{\Theta}, \mathbf{J}). \quad (1)$$

E_r prevents neighboring patches from having different transformations; E_{data} serves as a data term measuring how well the configuration of patches explains the observations, and E_{bone} favors bones to follow the patches attached to them. λ_r, λ_d and λ_b are corresponding weights that adjusting the influence of each term.

In addition, we partition input visual hull observations $\mathcal{T}^t$ into different rigid body parts by a linear multi-class Support Vector Machine trained on the shape of the reference model fitted to the previous frame $\mathcal{M}^{t-1}$. This body part information allows us to exploit $\mathcal{T}^t$ more efficiently when defining E_{data}. The outline of our method is given in Figure 2. In the remainder of this section, we briefly review the framework of [5] for the sake of completeness. We explain each energy term and describe how we partition the target point cloud in detail.

A. *Rigidity term* E_r

Cagniart *et al.* [5] proposed to decompose the reference mesh into a number of patches. Without prior knowledge of the motion, patches are preferred to be distributed uniformly on the surface. A rigidity constraint is exerted among them. The idea is that neighboring patches should agree on their prediction of the future position of each other. Specifically, let us consider a patch P_k, a patch P_l in its neighborhood N_k, and let $\mathbf{x}_l(v)$ denote the predicted position of v from P_l. The rigidity energy enforces the predicted position $\mathbf{x}_k(v)$ and $\mathbf{x}_l(v)$ to be consistent:

$$E_r(\mathbf{\Theta}) = \sum_{k=1:N_P} \sum_{P_l \in N_k} \sum_{v \in P_k \cup P_l} w_{kl} \|\mathbf{x}_k(v) - \mathbf{x}_l(v)\|^2. \quad (2)$$

$\mathbf{\Theta}$ is implicitly encoded in $\mathbf{x}_k(v)$ and $\mathbf{x}_l(v)$. This energy is quadratic in terms of $\mathbf{\Theta}$ so its minimum can be found via standard Gauss-Newton method.

B. *Data term* E_{data} *and point cloud partitioning*

The role of the data term is to connect the observations and the model. Generally speaking, one first estimates which target point the vertex or patch belongs to and defines a distance between the correspondences. Next, by minimizing this distance, the model deforms closer and closer to observations. Alternating between these two phases is known as ICP approach. Correspondence estimation can be treated as a classification problem. The output of the classifier is either dense labels for each vertex [20] or sparse labels for some feature points [2]. Cagniart *et al.* [6] adopt probabilistic ICP in patch-based deformation framework. Instead of a deterministic correspondence, each target point has a soft assignment to every patch as in Figure 3 (a). The method can be viewed as Expectation-Maximization algorithm in Bayesian maximum likelihood estimation. In E-step, soft

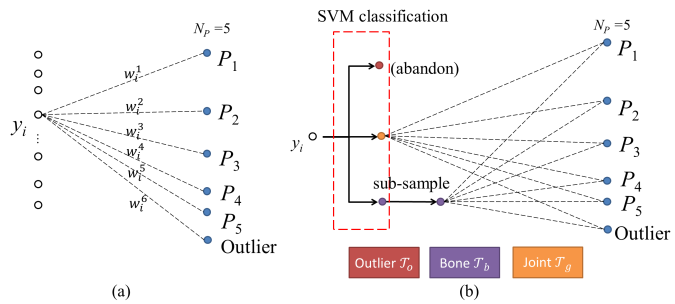


Figure 3. Illustrations of the classification scheme in (a) Cagniart *et al.* [6] and (b) our method. By performing SVM classification before the classification in [6], our method rules out fake geometries and provides a mechanism to work in the tradeoff between speed and accuracy.

assignments w_i^k are computed and in M-step, the energy is minimized in terms of model parameters $\mathbf{\Theta}$. Given the observations $\mathcal{T} = \{\mathbf{y}_i\}_{i=1}^{N_T}$, they define a data term:

$$E_{data}(\mathbf{\Theta}) = \sum_{i=1}^{N_T} \sum_{k=1}^{N_P+1} w_i^k \|\mathbf{y}_i - \mathbf{x}(v_i^k)\|^2, \tag{3}$$

where $\sum_k w_i^k = 1$, v_i^k is the corresponding vertex in P_k for $\mathbf{y}_i$, which is chosen considering closeness of both normals and distances.

Compared to the deterministic approach, the probabilistic association offers more robustness. The drawback, however, is the computational overhead. The data term as defined in Eq.(3) requires traversing all target points, which is computationally expensive. Also it does not incorporate any body part information due to the purely-surface-based assumption. Therefore, we advocate a *hierarchical classification scheme* in which unreliable or redundant observations are culled out before entering the classification in [6], as shown in Figure 3 (b). Since the vertex-joint association in $\mathcal{M}$ is fixed throughout the whole sequence, one knows the distribution of each rigid body part in $\mathcal{M}^{t-1}$. Meanwhile, in the context of tracking, it is pratical to assume that $\mathcal{M}^{t-1}$ and $\mathcal{T}^t$ distribute similarly. One can thus predict the rigid body part for each instance in $\mathcal{T}^t$ based on $\mathcal{M}^{t-1}$.

Specifically, for each vertex $v \in \mathcal{M}^{t-1}$, we use its 3D position $\mathbf{x}(v)$ as feature, and the associated joint j_v as class label. Let N_V denote the total number of vertex in $\mathcal{M}^{t-1}$. With these N_V training pairs $(\mathbf{x}(v), j_v)$ we aim to train a classifier. Any multiclass classifier with probability output serves our purpose. We suggest linear SVM [1] as a preferable choice because it provides a good compromise between accuracy and training time. In the case of binary classification (i.e., $j_v \in \{-1, +1\}$), it aims at finding a hyperplane with coefficients $\mathbf{w}$ that satisfies the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N} \xi_i \tag{4}$$

$$\text{s.t.} \quad j_v(\mathbf{w}^\top \mathbf{x}(v) + b) + \xi_i \geq 1, \ \xi_i \geq 0, \ \forall v \in \mathcal{M}^{t-1}.$$

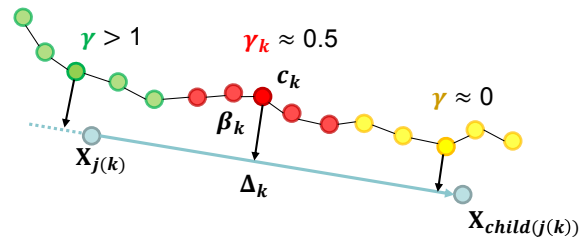


Figure 4. Patch-based β coordinate. γ contains information of the position of patches along the bone, which is used to determine the w_k in E_{bone}.

Parameters C and ξ_i are penalty weight and slack variables. The decision function for classifying target points in $\mathcal{T}^t$ is:

$$f(\mathbf{y}) = \text{sgn}\ (\mathbf{w}^\top \mathbf{y}_i + b), \ \forall \mathbf{y}_i \in \mathcal{T}^t. \tag{5}$$

For more details of multiclass SVM formulations we refer the interested readers to [11].

Each target point $\mathbf{y}_i$ has a rigid body label predicted from the SVM, and $\mathcal{T}^t$ is then partitioned into N_{RG} subsets ($N_{RG} = N_J -$ # non-leaf-node joints). By checking the probabilities from classifier, we further distinguish between target points that are near the joints, on the bones, and outliers. Let p_i denote a posteriori probability that $\mathbf{y}_i$ belongs to its predicted label. We apply the following criteria:

$$\mathbf{y}_i \in \begin{cases} \mathcal{T}_b & \text{if } 0.9 < p_i \\ \mathcal{T}_g & \text{if } 0.5 < p_i \leq 0.9 \\ \mathcal{T}_o & \text{if } p_i \leq 0.5, \end{cases} \tag{6}$$

where suffix b means right on bones, g means near around joints, and o means outliers. $\mathcal{T}$ is thus represented as $\{\mathcal{T}_b^j \cup \mathcal{T}_g^j \cup \mathcal{T}_o^j\}_{j=1}^{N_{RG}}$. We keep all $\mathbf{y}_i$ in $\mathcal{T}_g$, exclude all $\mathbf{y}_i$ in $\mathcal{T}_o$ and subsample $\mathbf{y}_i$ in $\mathcal{T}_b$. For instance, we keep all $\mathbf{y}_i$ near the knees and only a portion of $\mathbf{y}_i$ along the thighs and the calves, as in Figure 2 (d). This is because patches on a bone often move rigidly together. Only a small amount of target points are required to register some of them, and the rest of the patches can just follow. On the contrary, patches upon knees cannot be well predicted by those on the calves or thighs. They need more observations to be registered.

Let N_T' denote the number of observations after subsampling and outlier removal, we then apply the "soft" classification scheme in [6], and Eq. (3) is revised as:

$$E_{data}(\mathbf{\Theta}) = \sum_{i=1}^{N_T'} \sum_{k=1}^{N_P+1} w_i^k \|\mathbf{y}_i - \mathbf{x}(v_i^k)\|^2. \tag{7}$$

C. Patch-based skeleton binding energy E_{bone}

In [19] Straka *et al.* introduce differential bone coordinates β for every vertex, defined as:

$$\beta_i = \mathbf{x}(v_i) - \sum_{j=1}^{N_J} \rho_{i,j}(\gamma_{i,j}\mathbf{x}_j + (1-\gamma_{i,j})\mathbf{x}_{child(j)}), \tag{8}$$

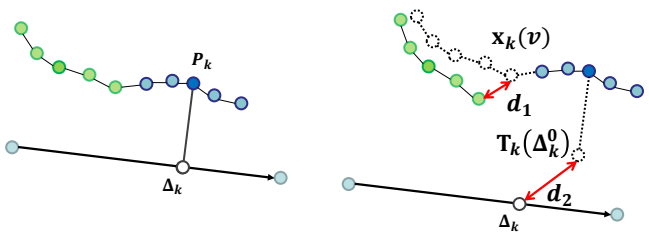


Figure 5. Illustration of the functions of E_{bone} and E_r. Left: initial configuration. Right: distances being minimized in the optimization. E_r minimize distance d_1 and E_{bone} minimize d_2. Both of them play the roles of regularization terms.

where $\mathbf{x}(v_i)$ represents 3D coordinates of the mesh vertex v_i. A bone is defined by joints $\mathbf{x}_j$ and $\mathbf{x}_{child(j)}$. $\gamma_{i,j}$ is chosen such that the vector between v_i and $\gamma_{i,j}\mathbf{x}_j + (1 - \gamma_{i,j})\mathbf{x}_{child(j)}$ is orthogonal to the bone. We follow this definition but compute it in a per-patch manner. Moreover, each patch considers only the attached joints $j(k)$ rather than all joints. We define our β coordinate as:

$$\beta_k = \Delta_k - c_k, \tag{9}$$

where $\Delta_k = \gamma_{k,j(k)}\mathbf{x}_{j(k)} + (1 - \gamma_{k,j(k)})\mathbf{x}_{child(j(k))}$ and c_k is the center location of patch, as shown in Figure 4. The patch-based skeleton binding energy keeps β from varying after transformation:

$$E_{bone}(\mathbf{\Theta}, \mathbf{J}) = \sum_{k=1}^{N_P} w_k \|\mathbf{T}_k(\beta_k^0) - \beta_k\|^2. \tag{10}$$

Here, we see that β_k is a function of $\mathbf{J}$ and transformation $\mathbf{T}_k$ includes rotation $\mathbf{R}_k$ and translation $\mathbf{t}_k$ of the patch, so E_{bone} is defined on the coupled space of mesh and skeleton. w_k are adjusted such that patches close to joints have less weight on where the associated joints should be, whereas patches in the middle of two joints contribute more. Such information is encoded in $\gamma_{k,j(k)}$.

We remark that Eq. (10) is better to be rewritten as:

$$E_{bone}(\mathbf{\Theta}, \mathbf{J}) = \sum_{k=1}^{N_P} w_k \|\mathbf{T}_k(\Delta_k^0) - \Delta_k\|^2, \tag{11}$$

because this way it can be interpreted easily together with E_r. As shown in Figure 5, when a patch moves to a new place, it predicts both the position of neighboring patches and Δ_k.

Combining Eq. (2), Eq. (7) and Eq. (11) into Eq. (1) we formulate our final energy function. We experimentally set $\lambda_d = 10$, $\lambda_r = 1$, and $\lambda_b = 1$ such that data term has higher importance than two smoothness terms that have equal influence. Our method is not overly sensitive to the exact values. Optimization of Eq. (1) is relatively standard since all the aforementioned energy terms are quadratic in terms of the model parameters. We therefore adopt Gauss-Newton algorithms to solve this unconstrained least-squares minimization. We implement SVM classification with the well-known library `libsvm` [1] which applies one-against-one multiclass strategy and provides the posterior class probabilities based on Platt scaling [1].

Sequence	Views	Frames	Patch #	Avg. spf.
Handstand1 [10]	8	401	144	6.07s
Wheel [10]	8	281	144	4.70s
Skirt [10]	8	721	219	3.57s
Dance [10]	8	574	223	3.68s
Crane [21]	8	175	125	3.55s
Handstand2 [21]	8	175	160	4.70s
Bouncing [21]	8	175	149	5.73s
Free [17]	8	500	172	5.46s
S4 walking [15]	4	349	186	4.76s

Table I
SEQUENCES USED FOR EVALUATION. WE FOLLOW [5] TO PATCH EACH REFERENCE MESH. TYPICALLY A NUMBER BETWEEN 150 AND 250 IS SUFFICIENT TO YIELDS DECENT RESULTS.

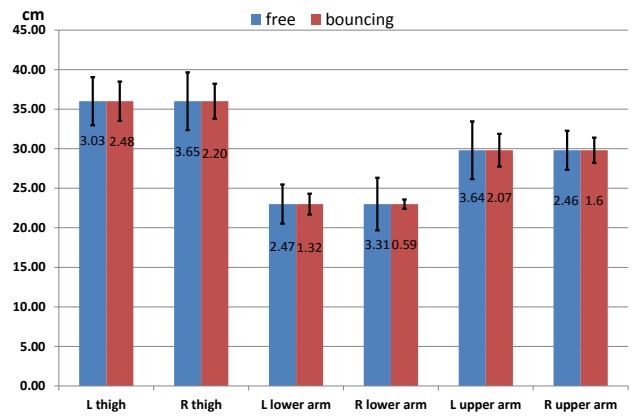


Figure 6. Averaged bone length of six body parts in free and bouncing sequences. Numbers in the bars are standard deviations.

IV. EXPERIMENTS

In this section we evaluate our method, both qualitatively and quantitatively. We test on 9 sequences from numerous public available datasets. These sequences range from those with rapid motions, e.g., Free [17] and Bouncing [21], to particularly articulated motions, e.g., Crane [21]. Table I lists the sequences and gives the average second per frame (spf) our method takes. We evaluate the pose and the shape separately, and demonstrate the effectiveness of the SVM-based matching scheme in terms of outlier rejection (Section IV-B) and target point subsampling (Section IV-C).

A. Evaluation on poses

With the S4 walking sequence from the `HumanEva-II` dataset [15], we compare the estimated joint locations with the ground truth obtained from markers. Skeletons in different datasets usually have different joint configurations in torso so we focus on limbs and head joints, for a total of 14 joints. Frames 298-335 are excluded due to the reported corruption of the ground truth in these frames. For the remaining frames, our approach presents an average total

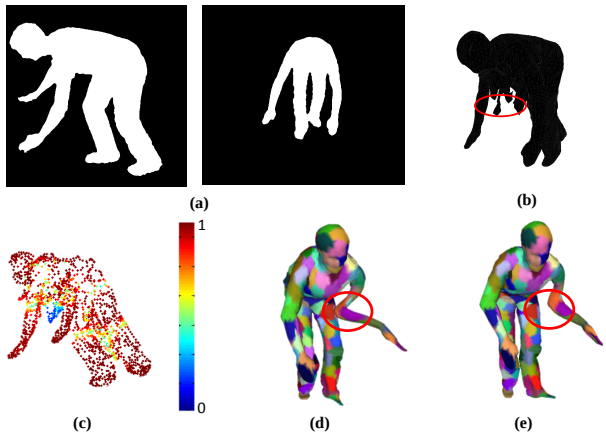


Figure 7. The effectiveness of SVM classification scheme. (a) Silhouette images in two views. (b) Reconstructed visual hull. Due to the ambiguity from 2D to 3D, there are some fake geometries in front of the chest. (c) Visual hull colored according to the probabilities from SVM classifier. (d) Estimated surface without filtering observations. (e) Estimated surface with outlier removal according to Eq. (6)

error of 70.86*mm*. According to [16], errors smaller than 80*mm* typically correspond to correct poses, which confirms the reliability of our method.

It should be noticed here that modeling a real human joint as a single 3D point is an over-simplified assumption. The numerical error in that case is only a coarse measure of how well the pose is estimated. Further optimizing on this error does not necessarily improve the estimation. On the other hand fixing the bone length increases robustness with respect to the noisy and occluded observations as claimed in [19]. In this work, in order to keep the bone length fixed, additional constraints are introduced and constrained least-square optimizations are performed. However, no reports on how much varying bone lengths influence the recovery of shape and pose is made. In our approach constant bone lengths are not explicitly constrained and we show variation of the averaged bone lengths of six body parts in the free and bouncing sequences in Figure 6. In order to make understandable and fair measurements, we map different scales of skeleton to centimeters using the anthropologic statistics from NASA [14]. Bar plots in Figure 6 show that our method exhibits significantly small bone length variations in spite of fast and large motions. This demonstrates that thanks to the stability of our method, bone lengths do not have to be additionally constrained, yielding a simpler and more efficient optimization.

B. Evaluation on shapes

1) Qualitative evolution: Our classification-based matching partitions incoming target observations into rigid body parts according to the reference shape model fitted to the previous frame. SVM classifier is trained using 3D coordinates of the reference mesh model in the previous frame with associated labels indicating the body part they belong. Here we demonstrate the effectiveness of this classification scheme. Artifacts appear in visual hulls due to silhouette ambiguities, as in Figure 7 (a) and (b). Using the visual hull vertices as target points, we obtain the results such as those in Figure 7 (d). However, the SVM classifier is able to identify artifacts and to give low probabilities to the corresponding points (Figure 7 (c)). Thus, outliers can successfully be removed based on Eq. (6). With the outlier rejection, our method estimates the surface as shown in Figure 7 (e), which demonstrates that the influence of fake geometries is alleviated.

2) Quantitative evolution: For quantitative evaluations, a commonly used metric is the *pixel overlap error* that measures the discrepancies between surface reprojections in the images and the corresponding input silhouettes. In Table II we show the ratio of erroneous pixels and the total number of pixels in the original silhouette. Since our approach builds on a patch-based deformation framework [6], we also compare to this method. As shown in Table II, our approach obtains better results than [6] in all sequences. This suggests that our SVM classification scheme helps in ruling out unreliable target points, a crucial feature when the input observations are noisy visual hulls. For further comparisons, we also implement a standard articulated ICP approach similar to [7]. Our method also shows better performances than this skeleton-based method as a result of a more flexible surface deformation model not constrained by pose space parametrization that is often insufficient in practice. These two comparisons show that our method outperforms both purely mesh-based approaches and simple skeleton-based methods.

Pixel overlap errors are also shown for the methods [10], [19], [21]. Nevertheless we would like to point out that all these methods explicitly optimize silhouette reprojection errors in images, thus naturally yielding small pixel overlap errors. However, visual hulls are noisy observations and our contribution is clearly to identify and remove erroneous observations hence the pixel overlap error is not necessarily a relevant criterion in this context. We observed anyway that, on average, our approach provides results comparable with these methods with no more than 6% errors, which is within a reasonable margin of error for the silhouettes.

C. Benefits of subsampling target points

In Figure 8 we present another benefit where we subsample the observations on the bones. Blue and purple lines in Figure 8 correspond to the averaged second per frame (spf) and silhouette overlap error respectively when matching is done using standard closest compatible point search. Target observations are not partitioned and all of them are inspected for the closest compatible point. Red and green curves correspond to the averaged spf and silhouette overlap errors of our method when classification with SVM is used for improved matching between the reference model and the input target

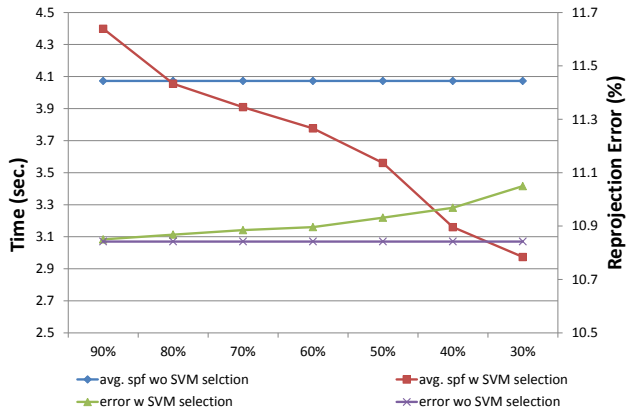


Figure 8. The tradeoff between time and overlap error on crane sequence. x axis is the subsample percentage in $\mathcal{T}_b$. Blue and red curves correspond to left y axis; green and purple curves correspond to right y axis.

observations. When most of the target observations $\mathbf{y}_i$ in $\mathcal{T}_g$ are kept (e.g., 90%), our approach requires more time due to the time-consuming SVM classification. In the case where the number of observations for training is reduced, the matching time decreases and the silhouette overlap error goes up. In the extreme case where 30% of $\mathbf{y}_i$ in $\mathcal{T}_g$ are used for classification, nearly 1.5 seconds per frame is gained while the increase in the error is only 0.2%. We remark here that the tradeoff between time and error is inevitable, but that our SVM-based approach provides a good compromise.

Lastly, some qualitative results are shown in Figure 9. Even the challenging Free sequence can be tracked properly. Our method is able to produce convincing results in terms of both shape and pose.

V. Conclusion and Future Work

We present an approach that captures marker-less human performances from multi-view sequences. Our method jointly estimates poses and shapes of the human body. To this end, we propose to use probabilistic deformable surface registration approach based on patched representation of the reference human body model [6] together with the bone binding energy [19]. In addition, we introduce a novel SVM-based classification scheme that partition target point clouds into rigid body parts and helps better correspondence search. We exploit posterior probabilities from classifiers to remove the redundant and unreliable observations and report speed up thanks to the use of the reduced set of observations for matching.

The reliability of the proposed method is verified by the experiments on sequences from various public datasets. Evaluations on `HumanEva-II` dataset show that our approach recovers the pose correctly. Our method does not rely on the bone-length constraint to obtain decent results. Evaluations on other sequences demonstrate that without explicitly optimizing on silhouettes, our approach still yield comparable results on shape estimation. Possible future directions include alleviating the requirement of background subtraction, and exploiting photometry information.

References

[1] Libsvm: a library for support vector machines. *ACM TIST*, 2(3):27, 2011. 4, 5

[2] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *ICCV*, pages 1092–1099. IEEE, 2011. 3

[3] I. Baran and J. Popović. Automatic rigging and animation of 3d characters. In *ACM Transactions on Graphics (TOG)*, volume 26, page 72. ACM, 2007. 3

[4] M. Botsch and O. Sorkine. On linear variational surface deformation methods. *Visualization and Computer Graphics, IEEE Transactions on*, 14(1):213–230, 2008. 2

[5] C. Cagniart, E. Boyer, and S. Ilic. Free-form mesh tracking: a patch-based approach. In *CVPR*, pages 1339–1346. IEEE, 2010. 1, 2, 3, 5

[6] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, pages 326–339. Springer, 2010. 1, 2, 3, 4, 6, 7, 8

[7] S. Corazza, L. Mündermann, A. Chaudhari, T. Demattio, C. Cobelli, and T. Andriacchi. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. *Annals of Biomedical Engineering*, 34(6):1019–1029, 2006. 6, 8

[8] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics (TOG)*, volume 27, page 98. ACM, 2008. 1, 2

[9] E. De Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *CVPR*, pages 1–8. IEEE, 2007. 1, 2

[10] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, pages 1746–1753. IEEE, 2009. 1, 2, 5, 6, 8

[11] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002. 4

[12] L. Kavan, S. Collins, J. Žára, and C. O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics (TOG)*, 27(4):105, 2008. 2

[13] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172. ACM Press/Addison-Wesley Publishing Co., 2000. 1, 2

Sequence	Our approach	Related approach			
Handstand1	15.53%	20.13% [6]	23.04%	7.66% [10]	8.07% [19]
Wheel	10.28%	10.30% [6]	14.35%	6.42% [10]	7.30% [19]
Skirt	11.94%	12.55% [6]	21.43%	9.56% [10]	8.04% [19]
Dance	9.95%	9.90% [6]	15.01%	12.08% [10]	6.11% [19]
Crane	10.79%	11.20% [6]	16.33%	5.29% [21]	
Handstand2	12.84%	13.97% [6]	15.16%	5.73% [21]	
Bouncing	9.87%	9.95% [6]	14.64%	5.34% [21]	
Free	14.12%	14.69% [6]			

Table II

COMPARISON OF OUR APPROACH WITH OTHERS. REPORTED VALUES ARE MEAN SILHOUETTE OVERLAP ERROR. THE SECOND COLUMN OF RELATED APPROACHES IS STANDARD ARTICULATED ICP WHICH IS SIMILAR TO [7]. WITHOUT EXPLICITLY OPTIMIZING ON SILHOUETTE IMAGES, OUR METHOD YIELDS ERROR IN GENERAL NO MORE THAN 6% COMPARED WITH THOSE FROM [10], [19], [21].

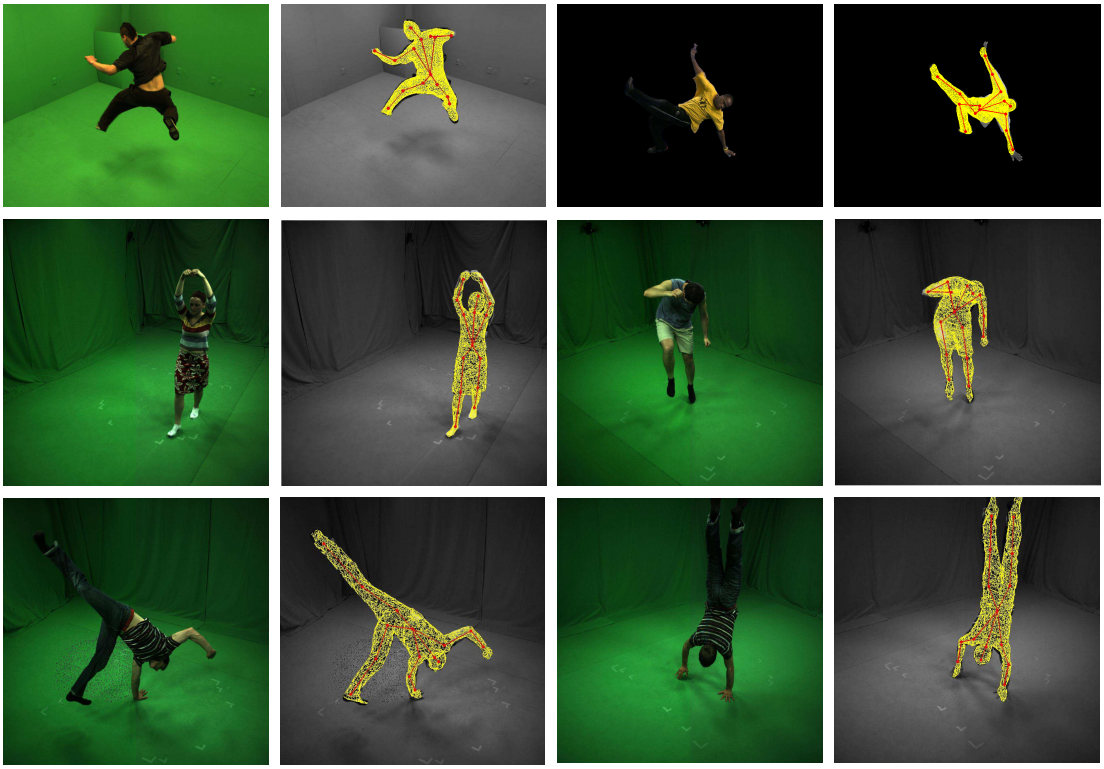


Figure 9. Example frames of input videos and overlaid estimated models

[14] NASA. Anthropometry and biomechanics. *Man-Systems Integration Standards*, 1. URL http://msis.jsc.nasa.gov/sections/section03.htm. 6

[15] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1):4–27, 2010. 2, 5

[16] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15–48, 2012. 6

[17] J. Starck and A. Hilton. Surface capture for performance-based animation. *Computer Graphics and Applications, IEEE*, 27(3):21–31, 2007. 1, 2, 5

[18] C. Stoll, J. Gall, E. De Aguiar, S. Thrun, and C. Theobalt. Video-based reconstruction of animatable human characters. In *ACM Transactions on Graphics (TOG)*, volume 29, page 139. ACM, 2010. 1

[19] M. Straka, S. Hauswiesner, M. Rüther, and H. Bischof. Simultaneous shape and pose adaption of articulated models using linear optimization. In *ECCV*, pages 724–737. Springer, 2012. 1, 2, 3, 4, 6, 7, 8

[20] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *CVPR*, pages 103–110. IEEE, 2012. 3

[21] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, volume 27, page 97. ACM, 2008. 1, 2, 5, 6, 8