Multilinear Pose and Body Shape Estimation of Dressed Subjects from Image Sets

Nils Hasler[*][†][§], Hanno Ackermann[‡], Bodo Rosenhahn[‡], Thorsten Thormählen[*], Hans-Peter Seidel[*]
Max-Planck-Institut Informatik (MPII),[*] Saarland University,[†]
Leibniz University Hannover,[‡] Weta Digital[§]
hasler@mpi-inf.mpg.de

Abstract

In this paper we propose a multilinear model of human pose and body shape which is estimated from a database of registered 3D body scans in different poses. The model is generated by factorizing the measurements into pose and shape dependent components. By combining it with an ICP based registration method, we are able to estimate pose and body shape of dressed subjects from single images. If several images of the subject are available, shape and poses can be optimized simultaneously for all input images. Additionally, while estimating pose and shape, we use the model as a virtual calibration pattern and also recover the parameters of the perspective camera model the images were created with.

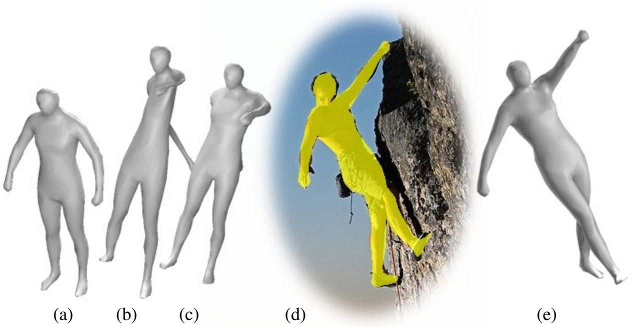

Figure 1. The estimation of a subject's pose and shape starts with the average person (a). Laplace deformation is applied to align the shape to image based constraints (b). This step can severely distort the shape. By fitting a shape/pose model to the deformed mesh a realistic human is created (c). Iterating these steps results in a shape as overlaid on the photograph (d) and shown on the right (e).

1. Introduction

This work focuses on the estimation of 3D shapes from silhouettes in single or multiple images. We derive a bilinear model which explains pose and shape variations of 3D scans, and learn the parameters of this model from a database of registered meshes of 114 subjects. Each person was scanned in several of 34 poses.

Starting with an initial 3D shape, we automatically compute correspondences between a silhouette and the shape. The parameters of pose and shape are subsequently determined to optimally explain the observed silhouette. Iterating both steps, we recover a 3D shape of the photographed subject. This pipeline is also visualized in Fig. 1. Additionally, the parameters of the perspective camera model the images were created with are estimated. We use this to further increase the accuracy of the estimated 3D shape.

1.1. Pose and Shape from 3D Scans

Given a database of 3D body scans, the problem is to estimate the parameters of a model which explains the variability among all scans. Once these parameters are known, new 3D shapes not in the database can be created. This is known as generating animatable 3D models from 3D scans.

To generate an animatable 3D model, Dekker *et al*. make heuristic assumptions about pose and shape of the scanned person to extract anthropometric landmarks and to segment the scan [8]. The resulting model can easily be animated but can only handle standing subjects.

An inverse approach is taken by Seo *et al*. who fit a skeleton driven parametric template model to an input scan [16]. Similarly, Allen *et al*. fit a rigged, skinned template model to input data [1]. The resulting meshes can be interpolated to create new shapes. Since the template model is articulated, and the embedded skeleton is scaled during shape conformation, it is possible to animate resulting models.

The popular SCAPE [2] model solves the problem in a similar way. Shape variation is encoded by analyzing principal components (PCA) of the 3D vertex positions of the training set. Pose variation is modeled by an embedded skeleton. Similar approaches for fast animation have recently been presented by Weber *et al*. [20] and Wang *et* *al*. [19].

In the approach by Hasler *et al*. [12], pose and shape variations are expressed by a differential encoding invariant to rotation and translation. The main drawback of this approach is that pose and shape cannot be analyzed independently.

In this work, we derive an analytical, bilinear model which explains poses and shapes of a database of registered meshes. The model consists of two independent sets of parameters which can be interpreted as pose and shape. Since the derived model is not based on skeletons, both sets can be estimated *linearly*, and skeleton-based constraints such as limb lengths may remain unspecified.

Our approach does not require every subject to be scanned in every pose. Subjects can be transformed to any pose by simple multiplication of shape and pose parameters to obtain 3D models not in the database. It is also possible to linearly interpolate between poses and shapes to create entirely new configurations. Furthermore, the linearity of the model permits simple integration of image-derived constraints into the estimation of 3D shapes.

1.2. Shape from Image Silhouettes

Estimating human pose and body shape from single images or image sets has not yet been solved satisfactorily. Yet, if parameters such as gender, or common body measures can be extracted reliably a number of tasks can be approached. For example a security system could improve the performance of its face recognition by considering body measures, or a medical application which detects posture problems could be built.

An early method for estimating body shape from images has been presented by Hilton *et al*. [13]. They require three predefined orthogonal views to fit a deformable template model to silhouettes. Bălan *et al*. use the SCAPE model to estimate pose and shape from multiview video streams [6]. However, since SCAPE does not allow directly estimating pose or shape parameters an analysis-by-synthesis approach is employed. This method is limited in that multiple synchronized views of the subject are required.

Similarly, Sigal *et al*. [17] used the SCAPE model to estimate pose and body shape from still images by learning a direct mapping from silhouettes to parameters of the shape/pose model. Training data is generated by randomly creating models with the SCAPE model. Since the mapping can only create a rough approximation of the correct pose and shape, a stochastic optimization technique is consecutively used to fine-tune the result. This approach is limited in that they cannot take more than one image of the same subject into account at any one time.

Bălan *et al*. [5] approach a similar problem. They also perform pose and shape estimation from single images. However, their setup includes a light source that creates a hard shadow. During optimization they additionally estimate the position of the light source. This allows them to use the shadow as an additional projection of the subject which stabilizes monocular pose estimation significantly. A disadvantage of this approach is that its use is restricted to carefully controlled in-door environments.

More recently, Bălan and Black [7] proposed a solution robust to loose clothing. Each subject is photographed several times with a multi-camera setup, wearing different clothes and in different poses each time. By combining the gathered constraints they are able to generate a 3D shape of the subject. The optimization is improved by performing skin color detection in the images. A limitation is that the proposed technique requires multiple synchronized cameras to work.

In a recent work, Guan *et al*. [9] estimate pose and shape from single images given only a number of manual correspondences from the image and the height of the subject. After fitting the SCAPE model to the markers, the resulting mesh is used to initialize a graph cut based segmentation algorithm [14]. In addition to the silhouette they propose to use edges to improve fitting for overlapping body parts. Shading cues restrict them to naked subjects but improve the accuracy of the estimation. Like [17] this approach is unable to handle more than one image simultaneously.

In a similar vein, we use the learned model of poses and shapes to constrain the estimation of silhouettes in images. The two parameter sets of poses and shapes are determined such that the observed silhouette is explained best. The proposed method works for dressed and undressed subjects, requires fewer markers and requires neither edge constraints nor shading cues. Although we do not incorporate additional information such as height or gender as proposed by Blanz and Vetter [3], Scherbaum *et al*. [15], or Guan *et al*. [9] in various contexts, it is easy to enforce several such constraints for any PCA based approach [12]. If either pose or subject is known a-priori, the parameters for pose or shape, respectively, can be fixed. This enables us to simultaneously estimate a 3D model given different poses of the same subject.

From the automatically established correspondences between silhouette and shape we further estimate the parameters of the perspective camera model the image was created with.

We demonstrate our algorithm with challenging images chosen from the internet (*e.g*. Flickr). Some show uncontrolled outdoor environments, some are paintings, and all of them are of low quality.

Contributions

- A multilinear, analytical model of human pose and body shape is learned from a database of 3D models of many undressed subjects [12], each present in several poses. Our approach permits to independently specify pose and shape parameters. This allows optimizing several images of the same subject simultaneously.

- The model is applied to the challenging problem of estimating both pose and body shape of dressed subjects from multiple images, photographs, or paintings. No shape prior is used. The segmentation is started from an average shape.
- If several images of a subject are available, *e.g.* a video sequence, all images can be used concurrently to estimate the 3D model even if the subject is shown in different poses.
- The proposed model requires only simple input data: only silhouettes and a few initially placed markers are sufficient. No camera calibration or synchronization is necessary.
- Given only a human silhouette in an image, we are able to use the current estimate of the shape and its correspondences on the silhouette to improve the perspective camera model.

The paper is structured as follows: In Section 2 the bilinear model of human pose and body shape is derived. In Section 3 we introduce the approach for estimating pose and shape of a person from a single image or a set of images. Experiments are discussed in Section 4 and conclusions are drawn in Section 5.

2. A Bilinear Model for Human Body Pose and Shape

In this section, we derive a method for estimating pose and shape parameters from registered 3D meshes of many subjects in many poses. The main idea of the approach presented here is that both shape and pose variations can be represented as affine transformations, and the vertices of each triangle can be explained by multiplication of the two transformation matrices. This is a bilinear model whose parameters – pose and shape – can be estimated by a linear, non-iterative procedure. The introduced algorithm is robust to missing scans, *i.e.* not every subject has to be scanned in every pose. Furthermore, new meshes which are not in the database can be synthesized easily.

Compared to the well-known SCAPE model the proposed method requires significantly less manual input. Namely, it is not necessary to manually assign triangles to body parts nor is a skeleton a required input for the proposed method. The implicit assumption that body parts move approximately rigidly when introducing a skeleton still applies.

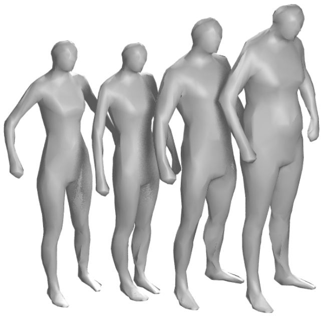

Figure 2. These subjects were asked to perform the same pose. Yet, variations in pose are significant.

Assuming a bilinear model of pose and shape parameters implies that the vertices of each triangle can be factorized into

$$\mathbf{M}_{ijk} = \mathbf{P}_{ik}\mathbf{S}_{jk}\mathbf{T}. \tag{1}$$

Here, $\mathbf{M}_{ijk}$ is a 3×3 matrix consisting of the vertices of the kth triangle of subject j in pose i, and $\mathbf{T}$ is a canonical template triangle in the xy-plane. Matrices $\mathbf{P}$ and $\mathbf{S}$ are affine transformations applied to $\mathbf{T}$.

The problem defined by Eq. (1) is to decompose $\mathbf{M}_{ijk}$ into pose and shape components $\mathbf{P}_{ik}$ and $\mathbf{S}_{jk}$, respectively. The classical factorization algorithm [18] estimates them for all triangles of all scans simultaneously. Thereby, the constraint is imposed that pose i performed by one subject is identical to the same pose performed by a different subject. However, this implies that the differences between the two measurements are solely a result of body shape variations. Unfortunately, this prerequisite does not hold (*cf.* Fig. 2). Since individual poses vary, any algorithm must consider this during estimation. In the following, we will show that very few further assumptions are sufficient to obtain pose and shape parameters which satisfactorily explain the observed 3D meshes and which can be used to generate new 3D models not yet in the database.

The assumption that all poses of the individuals differ implies that each triangle $\mathbf{M}_{ijk}$ can be decomposed into a shape parameter $\mathbf{S}_{jk}$ and a shape-dependent pose parameter $\mathbf{P}_{ijk}$. The pose transformation $\mathbf{P}_{ijk}$ decomposes into a rotation matrix $\mathbf{R}_{ijk}$, and a deformation matrix $\mathbf{D}_{ijk}$ for shearing and scaling. Similarly, the shape transformation $\mathbf{S}_{jk}$ can written as the product of a rotation $\mathbf{R}_{jk}$, and a shearing-scaling deformation $\mathbf{D}_{jk}$

$$\mathbf{P}_{ijk} = \mathbf{R}_{ijk}\mathbf{D}_{ijk} \quad \text{and} \quad \mathbf{S}_{jk} = \mathbf{R}_{jk}\mathbf{D}_{jk}. \tag{2}$$

Hence, $\mathbf{M}_{ijk}$ can be written as

$$\mathbf{M}_{ijk} = \mathbf{R}_{ijk}\mathbf{D}_{ijk}\mathbf{R}_{jk}\mathbf{D}_{jk}\mathbf{T}. \tag{3}$$

For every triangle $\mathbf{M}_{ijk}$ of subject j, shape parameters $\mathbf{R}_{jk}$ and $\mathbf{D}_{jk}$ are computed as the mean rotation and mean deformation over all poses of each subject.

Having estimated $\mathbf{S}_{jk}$ it seems that we may compute $\mathbf{P}_{ijk}$ simply as $\mathbf{P}_{ijk} = \mathbf{M}_{ijk}\mathbf{T}^{+}\mathbf{S}_{jk}^{+}$ where $(\cdot)^{+}$ denotes the generalized inverse. Unfortunately, if $\mathbf{P}_{ijk}$ is computed as described above, it cannot be applied to another subject. Simply transferring a pose of one subject to any other person violates the implicit assumption that triangles depend on the shape of the subject.

To generalize $\mathbf{P}_{ijk}$ so that it may be applied to other subjects, we introduce the constraint that deformations $\mathbf{D}_{ijk}$ always act on triangles in the xy-plane. This idea is motivated by the fact that shearing and scaling are not rotation-invariant. Therefore we define

$$\mathbf{D}'_{ik} = \mathbf{R}_{jk}^{-1}\mathbf{D}_{ik}\mathbf{R}_{jk} \tag{4}$$

and insert it into Eq. (3). This reverses the order of shape rotation $\mathbf{R}_{jk}$ and pose scaling $\mathbf{D}_{ijk}$ so that deformations always act on triangles in the xy-plane.

To be able to solve for $\mathbf{R}_{ijk}$ we need to further define

$$\mathbf{R}'_{ik} = \mathbf{R}_{jk}^{-1}\mathbf{R}_{ik}\mathbf{R}_{jk}. \tag{5}$$

Finally, we obtain

$$\mathbf{M}_{ijk} = \mathbf{P}_{ik}\mathbf{S}_{jk} = \mathbf{R}_{jk}\mathbf{R}'_{ik}\mathbf{D}'_{ik}\mathbf{D}_{jk}\mathbf{T}. \tag{6}$$

The decomposition of Eqs. (2) is performed by polar decomposition. However, whenever it is used care must be taken that the obtained rotation has a positive determinant. If the determinant of any rotation matrix happens to be negative, the sign of the right singular vector corresponding to the smallest singular value can be safely reversed since it only affects the unused deformation component in z-direction.

Summarizing, shape is computed as the average of all scans of a subject, and pose is considered to be a residual transformation. By immediately enforcing all constraints during the estimation procedure, a separate correction step after factorization is not necessary[1].

Principal component analysis is finally employed to learn a lower dimensional model of the parameters of pose and shape bases. Pose and shape bases are used to explain the observed 3D mesh (*cf.* Sec. 3). This requires that a linear combination of pose and shape rotations is defined. Since this cannot be done directly with rotation matrices, we represent rotations as log-quaternions, which can be interpolated safely. We do similarly for the parameters of deformation. This is also motivated by a further compression, namely that rotation and deformation are reduced to only 3 parameters each.

3. Estimating Pose and Shape from Silhouettes

In this section, a method is presented to estimate 3D meshes from image silhouettes using the pose and shape model (*cf.* Sec. 2) computed from our database of registered meshes. We also explain how the parameters of the perspective cameras are estimated.

We initialize the camera model by an orthographic camera, so all projection rays are orthogonal to the image plane. The average shape is used as a starting point for the optimization. This shape is optimized along with the pose to best explain the observed silhouette. Initially the average shape is rigidly rotated and translated to optimally fit to the orthographic projection rays. A few markers which correspond to certain points on the 3D mesh are selected in the image. We refine the initial 3D mesh by Laplace deformation [4, 21] so that the 3D correspondences of the markers are located on the projection rays of their 2D image positions. This constraint is satisfied if the cross product between the line 2D-3D point and the orthographic projection ray becomes zero

$$\lambda \times (\mathbf{m} - \mathbf{v})\,, \tag{7}$$

where λ denotes the orthographic projection ray of the 3D position of marker $\mathbf{m}$ in the image plane and the corresponding 3D mesh vertex $\mathbf{v}$. This is a linear constraint in the vertex position $\mathbf{v}$ which can be easily integrated into Laplace deformation methods.

An iterated closest point procedure is used to compute additional correspondences between silhouette and 3D mesh. However, instead of using constraints on the position of 3D vertices we use surface normals. They constrain the vertices of the 3D mesh to lie on the planar surface which locally approximates the measured silhouette. This improves convergence and is motivated by projective ambiguity: nothing is known about the position of any 3D point except that it is on its projection ray. Each known surface normal induces one equation

$$\nu \cdot (\mathbf{v} - \mathbf{o}) = 0, \tag{8}$$

where ν denotes the normal vector of the surface and $\mathbf{o}$ the closest point to $\mathbf{v}$ on the surface. The "$\cdot$" symbol denotes the inner vector product. Equation (8) is linear in the unknown vertex position $\mathbf{v}$ hence it can also be used to constrain a Poisson reconstruction.

This Poisson-reconstructed mesh has to be explained by pose and shape bases which implies that we impose priors on the mesh we expect to observe in the images. To explain the 3D model by pose and shape bases, it is necessary to determine the linear coefficients of all combinations of poses i and shapes j which optimally support the observed triangles. From these triangles, we extract their rotation and deformation parameters by polar decomposition of $\mathbf{M}_k\mathbf{T}^{+}$.

[1] In fact, the factorization algorithm as introduced in [18] estimates affinely distorted parameter sets. The original algorithm therefore requires a second stage called "metric upgrading" in which certain constraints are imposed on the model.

Here, $\mathbf{M}_k$ denotes the matrix consisting of the three vertices of the kth triangle.

To evoke linear interpolation between rotations we switch to log-quaternion algebra. Denote by $\mathbf{R}^m_k$ the rotation matrix induced by triangle $\mathbf{M}_k$, by ρ the function which maps any rotation matrix $\mathbf{R}$ to its log-quaternion representation, and by $\rho(\cdot)^{-1}$ the inverse mapping. Vectors $\mathbf{r}^p_{ik}$ and $\mathbf{r}^s_{jk}$ indicate the rotational parameters of pose i and shape j, respectively, for triangle k. Using log-quaternion algebra, we obtain from Eqs. (1) and (6)

$$\mathbf{R}^m_k = \rho\left(c^s_1 \mathbf{r}^s_{1k} + c^s_2 \mathbf{r}^s_{2k} + \ldots\right)^{-1} \rho\left(c^p_1 \mathbf{r}^p_{1k} + \ldots\right)^{-1}. \quad (9)$$

The coefficients of the linear combination are denoted by c^p_i for poses and c^s_j for shapes. We can rewrite Eq. (9) in the form of a bilinear equation system with parameter vectors $\mathbf{x} = [\, c^s_1 \;\ldots\; c^s_J \,]^\top$ and $\mathbf{y} = [\, c^p_1 \;\ldots\; c^p_I \,]^\top$

$$\mathbf{R}^m_k = \rho(\mathbf{x}^\top \begin{bmatrix} \mathbf{r}^s_{1k} \\ \vdots \\ \mathbf{r}^s_{Jk} \end{bmatrix})^{-1} \rho([\, \mathbf{r}^p_{1k} \;\ldots\; \mathbf{r}^p_{Ik} \,]\, \mathbf{y})^{-1}. \quad (10)$$

For the deformation parameters, we obtain a similar constraint on the coefficients of the linear combination.

Given $\mathbf{y}$, we obtain the total pose rotation $\mathbf{R}^p_k$ by the inverse of ρ. The inverse of $\mathbf{R}^p_k$ is then right-multiplied to $\mathbf{R}^m_k$. Shape coefficients $\mathbf{x}$ can be solved for after mapping $\mathbf{R}^m_k \left(\mathbf{R}^p_k\right)^{-1}$ by function ρ to log-quaternion representation. The derivation for deformation constraints is similar.

Solutions to the vectors $\mathbf{x}$ and $\mathbf{y}$ can be estimated by iteratively solving for one given the other [10]. Vector $\mathbf{y}$ is initialized either to the average shape, or to the shape which explains the previous 3D mesh. According to our experience, this scheme converges reliably in few iterations.

Finally, with the coefficients $\mathbf{x}$ and $\mathbf{y}$, a new 3D model is synthesized. We use it as a "virtual calibration object" to estimate a perspective camera model (*cf.* Sec. 3.1). We iterate the steps between silhouette-mesh correspondence computation and camera model refinement until convergence.

3.1. Virtual Calibration Object

Having established 2D-3D correspondences between the silhouette and the estimated 3D shape, we may use the 3D shape as "virtual calibration object". Generally, the calibration of a perspective camera, *i.e.* determining its projection matrix and decomposing it into intrinsic and extrinsic parameters, may be performed by means of direct linear transformation once some 3D points of an observed object are known. This implies that we can use the estimated 3D shape as virtual calibration object since its 2D-3D correspondences are all known.

Since the projection ray of any 3D point $\mathbf{X}$ need be orthogonal to its 2D correspondence $\mathbf{x} = [x\ y\ 1]^T$ [11], we obtain

$$x \times \mathbf{P}\mathbf{X} = 0 \quad (11)$$

where $\mathbf{P}$ is the projection matrix of the perspective camera, and $\times$ denotes the cross product. This induces a linear equation system we solve for the unknown camera matrix $\mathbf{P}$.

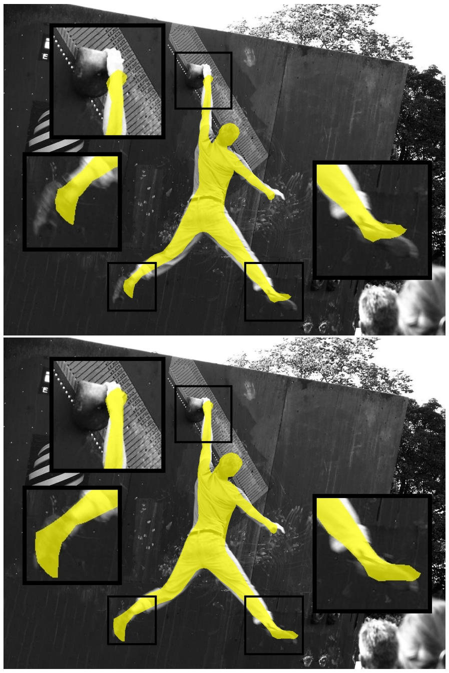

Figure 3. Estimation computed with an orthographic (**upper image**) and a projective camera (**lower image**).

Let $\mathbf{K}$ be the 3×3 upper-diagonal matrix of the intrinsic parameters, $\mathbf{R}$ the 3×3 matrix indicating the orientation of the camera, and $\mathbf{t}$ a 3-vector of the translation of the camera from the origin. Since

$$\mathbf{P} = \mathbf{K}\,[\mathbf{R}|\mathbf{t}]\,, \quad (12)$$

we can determine $\mathbf{K}$ and $\mathbf{R}$ from the first three columns of $\mathbf{P}$ using the fact that $\mathbf{R}$ is a rotation matrix and $\mathbf{K}$ symmetric by using singular value decomposition.

4. Experimental Results

In this section experimental evaluation of the proposed model is performed. We use the approaches described in the previous sections to estimate 3D shapes of several subjects from single images or image sets. The *Flickr*-images were taken under diverse, uncontrolled lighting conditions and the subjects wear unchecked clothes. Some of the images are of low quality.

4.1. Camera Estimation

In Fig. 3 a subject is shown jumping to a climbing hold. The two images correspond to two states of the optimiza-

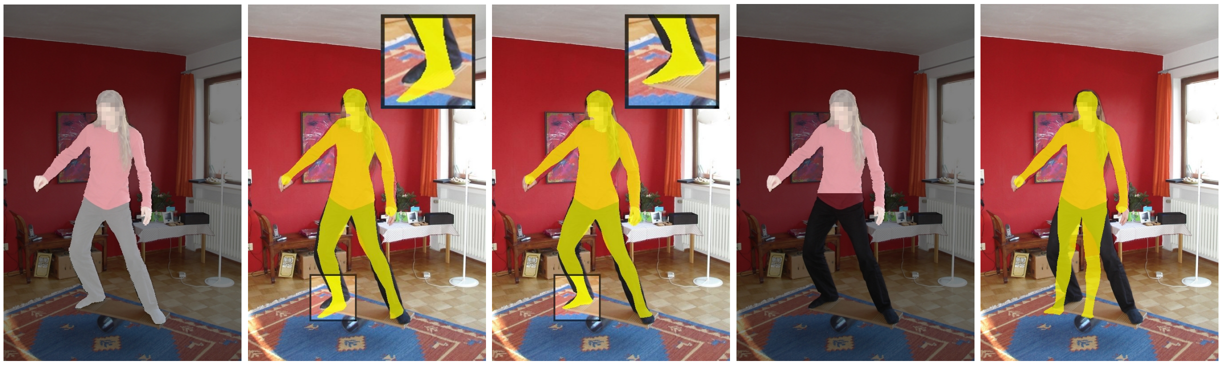

Figure 4. **Leftmost three images**: input silhouette, result with an orthographic (middle) and a projective camera (right). **Rightmost two images**: silhouette with simulated occlusion (from hips downwards) and estimation result.

tion procedure. For the upper image, shape is estimated using an initial orthographic camera. For the lower image, a projective camera was estimated which optimally fits the shape to the silhouette as explained in Sec. 3.1. Especially the parts of the subject that are most affected by perspective distortion, namely the feet, fit much better to the silhouette.

A similar result is shown in Fig. 4. The left three images show the same experiment as above. The model is fitted to the silhouette and results are shown before and after camera estimation. Again, the perspective correction improves the result significantly. The right two images correspond to an experiment where a partially visible silhouette was simulated by occluding the silhouette from the hips downwards. The estimated 3D shape fits well to the visible silhouette, and the legs are estimated to have a matching size. From hereon, only the results after camera estimation are shown.

4.2. Automatic Segmentation

In many of our examples, unchecked lighting and low image quality make manual segmentation of the input essential. In some cases, however, if the background is sufficiently different from the subject, we are able to perform automatic segmentation. The example in Fig. 5 shows such a case. The model is first fitted to four markers on the hands and feet of the subject. This first initialization does not fit the pose very well but it is sufficient to initialize the GrabCut segmentation algorithm [14]. The resulting silhouette is accurate in most parts. Only between the legs the trees are wrongly classified as foreground. This is the reason that the left foot of the snowboarder is estimated to be too high. Repeating the procedure with a better starting point resolves the segmentation issue, and the computed 3D shape finally fits the silhouette well.

4.3. Paintings

One difficulty with paintings is that the artist may have chosen to deform the subjects. For instance, in the painting in Fig. 6 (*Hay Harvest at Éragny* by Camille Pissarro) the women have unrealistically long arms. The effect on pose estimation is such that the fit seems to be of low quality at first glance but since the learned shape model forbids unrealistic human shapes, the computed mesh adheres to regular human proportions. For the image shown in Fig. 7 (Rembrandt's *Night Watch*), the main difficulty during estimation of the subjects lies in the irregular silhouettes. The militiamen wear hats, carry guns and swords. Still, a good approximation of the characters can be computed.

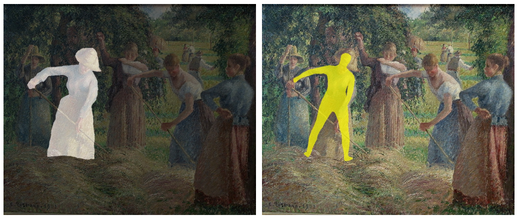

Figure 6. The women in the painting *Hay Harvest at Éragny* by Camille Pissarro have exceptionally long arms. This cannot be explained by the model. The learned shape model enforces a fit that strictly adheres to common human proportions.

4.4. Multiple Images

The previous examples showed pose and shape estimates computed from a single image. If, however, several images of one subject are available an improved shape estimate can be computed by considering all available information. This is demonstrated in Fig. 8. It shows the input silhouettes (left image of each block), estimation results by independently fitting a 3D shape to each image (middle image), and the results after enforcing that shape coefficients are equal for all input images (right). The figure also shows side-by-side comparisons of the estimated 3D shapes in the observed poses (lower right image block).

By training a simple linear regressor on the database the method also allows us to estimate biometric parameters. The subject in Figure 4 was estimated to be 178 cm tall and to weigh 72 kg while he truly stands 180 cm tall and weighs

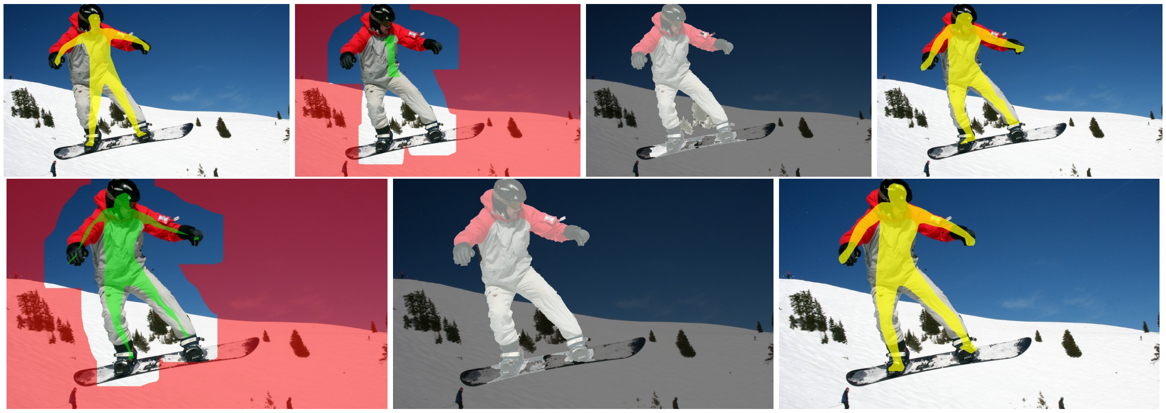

Figure 5. Automatic foreground segmentation. **Top row** (from left to right): Initial estimate using four markers, input to GrabCut, output of GrabCut, and estimation result. **Bottom row**: Second iteration of the approach. GrabCut input, GrabCut output, and result are shown.

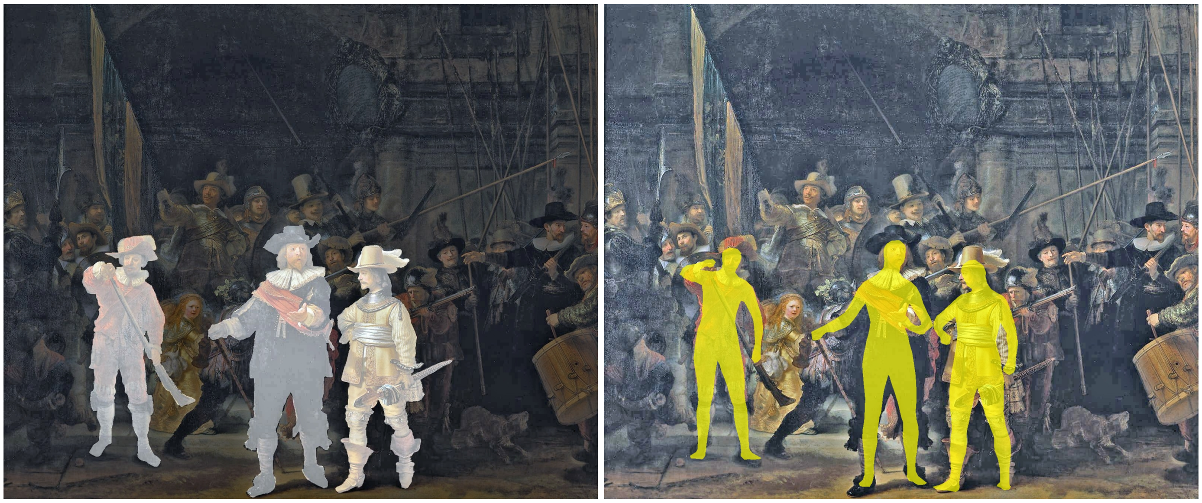

Figure 7. The poses of three subjects from Rembrandt's *Night Watch* are estimated. The difficulty here is that the silhouettes are highly irregular.

67 kg. While true height and weight of the subject in Fig. 8 are not known, we can ascertain that the estimates for the four images agree well. Standard deviation of height and weight are 1.2 cm and 2.2 kg respectively.

Acknowlegdements

We would like to thank the photographers Darren Copley (Fig. 3), Shay Haas (Fig. 5), and foto.pamp.es (Fig. 8).

5. Conclusions

In this paper a bilinear model of human pose and body shape is proposed. Its parameters are learned from a database of undressed subjects. The model is applied to the challenging problem of estimating 3D meshes of *dressed* subjects from single or multiple images or paintings. Since pose and shape parameters are separated in the proposed model, it is possible to simultaneously optimize a model from several images showing the same person in different poses. Given only silhouettes in the images we estimate parameters of full perspective cameras. We presented experiments with highly challenging and low quality image material.

References

[1] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM ToG*, 22(3):587–594, 2003. 1

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM ToG*, 24(3):408–416, 2005. 1

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH Papers*, pages 187–194, New York, NY, USA, 1999. ACM Press. 2

[4] M. Botsch and O. Sorkine. On linear variational surface deformation methods. *IEEE Trans. on Visualization and Computer Graphics (TVCG)*, 2007. 4

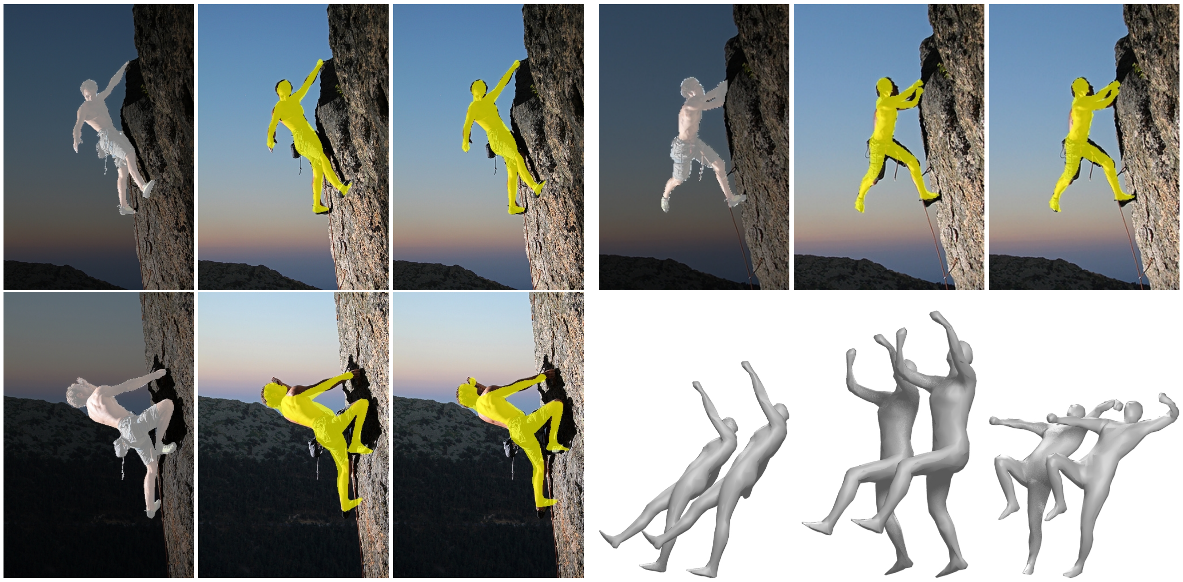

Figure 8. If multiple views of one subject (here Javier Clos) are available conjoint optimization improves the stability of the estimation. Three blocks (**left to right**) input silhouette, estimate with variable shape, result after jointly optimizing shape. The fourth block shows the estimated shapes before and after optimization while enforcing the same shape parameters for the four images.

[5] A. Bălan, M. Black, H. Haussecker, and L. Sigal. Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In *Proc. ICCV*, pages 1–8, Oct. 2007. 2

[6] A. Bălan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *Proc. CVPR*, pages 1–8, June 2007. 2

[7] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *Proc. ECCV*, volume 5303, pages 15–29, Oct. 2008. 2

[8] L. Dekker, I. Douros, B. Buston, and P. Treleaven. Building symbolic information for 3d human body modeling from range data. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, pages 388–397, 1999. 1

[9] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *Proc. ICCV*, Kyoto, Japan, Sept. 2009. 2

[10] R. Hartley and F. Schaffalizky. PowerFactorization: 3D Reconstruction with Missing or Uncertain Data. In *Australia-Japan Advanced Workshop on Computer Vision*, June 2002. 5

[11] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 5

[12] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Eurographics*, number 28, Mar. 2009. 2

[13] A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun. Virtual people: Capturing human models to populate virtual worlds. In *CA '99: Proceedings of the Computer Animation*, page 174, Washington, DC, USA, 1999. 2

[14] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH Papers*, pages 309–314, New York, NY, USA, 2004. ACM Press. 2, 6

[15] K. Scherbaum, M. Sunkel, H.-P. Seidel, and V. Blanz. Prediction of individual non-linear aging trajectories of faces. In *Eurographics*, volume 26 of *Computer Graphics Forum*, pages 285–294, Prague, Czech Republic, 2007. Blackwell. 2

[16] H. Seo and N. Magnenat-Thalmann. An automatic modeling of human bodies from sizing parameters. In *SI3D '03: Proceedings of the 2003 symposium on Interactive 3D graphics*, pages 19–26, New York, NY, USA, 2003. ACM Press. 1

[17] L. Sigal, A. Bălan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007. 2

[18] C. Tomasi and T. Kanade. Shape and Motion from Image Streams under Orthography: A Factorization Method. *International Journal of Computer Vision*, 9(2):137–154, November 1992. 3, 4

[19] R. Y. Wang, K. Pulli, and J. Popović. Real-time enveloping with rotational regression. In *ACM SIGGRAPH Papers*, page 73, New York, NY, USA, 2007. ACM Press. 2

[20] O. Weber, O. Sorkine, Y. Lipman, and C. Gotsman. Context-aware skeletal shape deformation. *Computer Graphics Forum*, 26(3):265–274, Sept. 2007. 1

[21] Y. Yu, K. Zhou, D. Xu, X. Shi, H. Bao, B. Guo, and H.-Y. Shum. Mesh editing with poisson-based gradient field manipulation. In *ACM SIGGRAPH Papers*, pages 644–651, New York, NY, USA, 2004. ACM Press. 4