Model-Based Generation of Realistic 3D Full Body Avatars from Uncalibrated Multi-view Photographs

Nicholas Michael, Andreas Lanitis

Model-Based Generation of Realistic 3D Full Body Avatars from Uncalibrated Multi-View Photographs

Nicholas Michael[a] and Andreas Lanitis[b]

Visual Media Computing Lab, Dept. of Multimedia and Graphic Arts,
Cyprus University of Technology, 3036 Lemesos, Cyprus
[a]nicholas.michael@cut.ac.cy, [b]andreas.lanitis@cut.ac.cy

Abstract. In today's world of rapid technological advancement, we find an increasing demand for low-cost systems that are capable of fast and easy generation of realistic avatars for use in Virtual Reality (VR) applications. For example, avatars can enhance the immersion experience of users in video games and facilitate education in virtual classrooms. Therefore, we present here a novel model-based technique that is capable of real-time generation of personalized full-body 3D avatars from orthogonal photographs. The proposed method utilizes a statistical model of human 3D shape and a multi-view statistical 2D shape model of its corresponding silhouettes. Our technique is automatic, requiring minimal user intervention, and does not need a calibrated camera. Each component of our proposed technique is extensively evaluated and validated.

Keywords: personalized avatars, 3D body shape modelling, multi-view ASM, image segmentation

1 Introduction

In today's world of rapid technological advancement, we find an increasing demand for low-cost systems that are capable of fast and easy generation of realistic avatars with minimal user intervention. Such systems have applicability in many Virtual Reality (VR) applications. For example, they can be used in cultural heritage visualizations and to populate virtual worlds in multi-player computer games, enhancing a user's immersion experience. In addition, they can facilitate communication and education when incorporated in chat applications and in virtual classrooms. They can even assist users to make shopping decisions by allowing them to dress their avatars accordingly. Existing technologies make it possible to create 3D avatars, e.g., using 3D scanning devices such as Microsoft's Kinect, using 3D modelling software, etc. However, this kind of technologies tend to require dedicated hardware, calibration of imaging equipment, creative skills and considerable amount of post-processing manual intervention.

Motivated by this multitude of applications and the limitations of existing technologies, we develop a model-based method for the generation of realistic personalized full-body 3D avatars [1]. The novelty of our work is that it is automatic and it works using photographs taken by any uncalibrated low cost camera. Our method extends the work of Hilton et. al. [2], however ours does not require the user to mark the location of 3D landmarks in the 2D image, in order to extract model-silhouette correspondences, for

the purpose of improving accuracy. Furthermore, the proposed method does not require the user to measure the camera's field of view and the subject's distance to the camera, which serve as camera calibration. Instead, and as our main contribution, we train and use a *multi-view* Active Shape Model (ASM) [3] of human silhouettes (as viewed from the front, left, right and back), in order to refine the extracted silhouettes obtained from the segmentation result and to register them to the projection of the 3D shape model in each of the orthogonal views. Our only assumption is that when taking the multi-view input images the user does not significantly change their depth relative to the camera. This means that the avatars can be generated quickly and easily, as no time is wasted for calibrating the camera nor for marking correspondences. Additionally, in order to increase the recognizability and hence the realism of the generated avatar, together with the full-body model we use a face-only 3D model, which has a significantly higher resolution aimed at capturing the more detailed facial characteristics.

An outline of our method is illustrated in Figure 1. First we train a statistical model of 3D human shape using a dataset of dense full-body 3D scans [4]. We project the 3D scans of the dataset to orthogonal views and extract the corresponding silhouettes, which we use to train our multi-view ASM [3], thus concluding the training phase. Once these two models are trained, we can deploy the proposed technique live as follows. Orthogonal input images are captured using an uncalibrated low cost camera and we extract 2D silhouettes using background subtraction and other image processing techniques. The extracted silhouettes are refined and registered to the 3D model's projection using the trained multi-view ASM. For improved accuracy around the face, we detect facial features (eyes, nose and mouth) using Viola/Jones detectors [5] and combine the detection result with the registered silhouettes, while performing an optimization over the space of permissible 3D shape parameter. Once the shape is reconstructed, the two 3D models (full-body and face-only) are aligned with each other, using a rigid transformation and then texture from input images is mapped to both models assuming a cylindrical model [2].

The remainder of our paper is organized as follows. Section 2 covers previous work on the problem of avatar generation. Section 3 discusses in detail each phase of our proposed method, such that Sect. 3.1 covers the components of the off-line model training phase and Sect. 3.2 describes the steps involved in the avatar generation phase. Section 4 presents experimentation for the evaluation of our work. We conclude with Sect. 5 where we also mention a few thoughts on possible future work.

2 Related Work

Previous efforts on the generation of realistic human-like avatars can be categorized into two groups. In one group are model-based methods (such as [2], [6], [7], [8], [9] and [10]) that rely on a model of human body shape to reconstruct the subject's geometry from a multi-view camera setup. In the other group are model-free methods (such as [11], [12], [13], [14] and [15]), which do not rely on a human body shape model and perform multi-view reconstruction, using for example, multi-view photometric stereo [12], [15] or even a setup of inexpensive range scanners like those found in the recently popularized Kinect device [11], [14], [16].

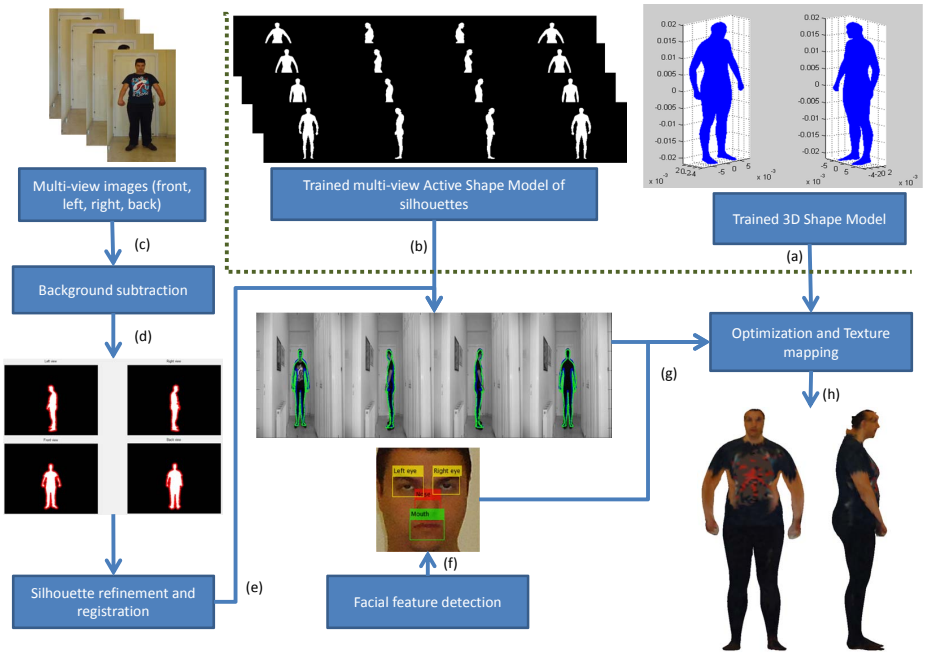

Fig. 1. Method overview (dotted line separates training and live phases): (a) A 3D shape model is trained by applying PCA on a dataset of dense full-body 3D scans, (b) A multi-view Active Shape Model is trained from the 2D silhouettes of the full-body 3D dataset, (c) Orthogonal input images are captured using an uncalibrated low cost camera, (d) Extract silhouettes using background subtraction, (e) Refine and register the silhouettes to the 3D model's projection using the multi-view ASM, (f) Detect facial features using Viola/Jones detectors, (g) Given the registered silhouettes and detected facial features optimize over the 3D shape parameters and map the texture from the input images, (h) Frontal and side views of generated 3D avatar

Fig. 2. (Left) Sample scans in used dataset [4]. Only 111 (in neutral pose) were used for training, (Right) The model can be used to generate random populations that follow the distribution of the training set by varying the shape parameters to vary e.g., height, weight, muscle/fat ratio, etc.

Model-based approaches use a prior model of the human body shape, which allows for consistent reconstruction across frames but the reconstructed shapes are limited by the utilized model, in that they may not accurately reconstruct a previously unseen human body shape that is not consistent with the training set. Model-free methods are more dynamic in nature and can reconstruct shape from any scene but because of their lack of a prior model, their reconstructed shapes may not be consistent even across neighboring frames and cannot handle ambiguities. An important advantage of model-based approaches is the uniformity of the avatars created in terms of the number and positioning of polygons, that facilitates the animation of the resulting avatars in virtual environments. In our approach we adopt a PCA model-based approach, since it naturally enforces consistency in the reconstructed shape across the four orthogonal views.

3 Methodology

The proposed avatar generation method involves two main phases: (i) the off-line model training phase and (ii) the live avatar generation phase. Figure 1 illustrates the overall work flow of our proposed technique. In the following subsections we describe the various steps in detail.

3.1 Off-line Model Training

In the off-line phase, we use a dataset of full-body range scans [4] (selecting only the 111 samples with a neutral pose – see Fig. 2 (right)) to train a 3D full-body shape model. Then the training samples are projected to four orthogonal views (front, left, right, back) to generate a training set of corresponding multi-view silhouettes, which we use to train a multi-view ASM model (see Fig. 1 (a-b)).

3D Model In the training phase we utilize a Principal Component Analysis (PCA) model trained on range scans from a 3D body scan dataset [4] to extend the work of [2], [9] and [13]. The purpose of the PCA model is to learn the permissible modes of human body 3D shape variations reflected in the training set, so that the avatars we generate will have a realistic human shape. Each range scan is represented as a column vector, $\boldsymbol{x}$, of vertex coordinates such that $\boldsymbol{x} = [x_1, ..., x_N, y_1, ..., y_N, z_1, ..., z_N]^\top$, where N represents the number of vertices in the 3D mesh (in our case $N = 6449$). The range scans are aligned using Procrustes Alignment [17].

We apply eigen-decomposition on the covariance matrix of the aligned shapes, while keeping only the first m out of the resulting n eigenvectors, sorted in order of decreasing eigenvalue, λ_i, such that:

$$\arg \max_{1 \leq m \leq n} \left\{ \left(\sum_{i=1}^{m} \lambda_i \right) / \left(\sum_{i=1}^{n} \lambda_i \right) < v \right\} , \tag{1}$$

where $v \in [0, 1]$ is the amount of variance that we want the PCA model to capture (we have set $v = 0.98$ in our experiments, which resulted in $m = 13$). In this way, any new

shape, $\boldsymbol{x}_i$, can be represented as:

$$\boldsymbol{x}_i \approx \bar{\boldsymbol{x}} + \boldsymbol{C}\boldsymbol{b}_i \ , \tag{2}$$

where $\bar{\boldsymbol{x}}$ is the mean shape of the model, $\boldsymbol{C}$ are the principal m eigenvectors of the learned shape manifold and $\boldsymbol{b}_i$ is a column vector of shape parameters, also known as the encoding of shape $\boldsymbol{x}_i$. This encoding vector $\boldsymbol{b}$ is typically truncated, so that each element satisfies $b_i \in [-2\sqrt{\lambda_i}, +2\sqrt{\lambda_i}]$, where usually $k = 2$. This ensures that any shapes generated by (2) remain plausible with respect to the training set.

Silhouette Multi-view ASM We project the 3D range scans of the training set to four orthogonal views (front, left, right and back) extracting the corresponding 2D silhouette contour in each view. Each silhouette contour is represented as a column vector, $\boldsymbol{y}_i$, of 2D coordinates such that $\boldsymbol{y}_i = [x_1, ..., x_C, y_1, ..., y_C]^\top$, where C represents the number of points in the 2D contour and $i \in \{\text{front}, \text{left}, \text{right}, \text{back}\}$. For each training range scan, the set of four silhouette vectors, representing its multi-view orthogonal 2D projection, is stacked vertically, resulting in the augmented column vector: $\boldsymbol{y} = [\boldsymbol{y}_{\text{front}}, \boldsymbol{y}_{\text{left}}, \boldsymbol{y}_{\text{right}}, \boldsymbol{y}_{\text{back}}]^\top$.

The set of augmented column vectors is aligned using Procrustes Alignment and subsequently used to train by application of PCA a *multi-view* ASM model (we set $v = 0.98$, yielding $m = 33$), instead of training a separate ASM model per view. In this way, during the ASM search algorithm [3] the shape parameters are optimized across all four views simultaneously, instead of independently, providing robustness to outliers, as it naturally enforces shape consistency across all views, yielding a more accurate result. Figure 3 illustrates the modes of variation learned by the first few eigenvectors of the trained multi-view ASM.

3.2 Avatar Generation

In the live phase, we extract the subject's silhouettes and then we use the trained models to reconstruct the 3D shape of the subject, register the face model and perform texture mapping (see Fig. 1 (c-h)).

Image Acquisition Image acquisition is done using a low cost camera. First we photograph the background and then the subject is photographed from the front, left, right and back, using a tripod to keep the camera stationary. We assume that the subject maintains a constant distance from the camera during the acquisition.

Segmentation The segmentation step is needed to extract the silhouettes in each view. First the input images are converted to the CIELAB colorspace and we obtain their difference image with respect to the background. The difference image is then thresholded and we apply erosion and dilation to merge foreground regions, selecting the largest connected component as the foreground. We extract the contour by following 8-connected adjacent pixels on the silhouette boundary. Figure 5 illustrates the procedure.

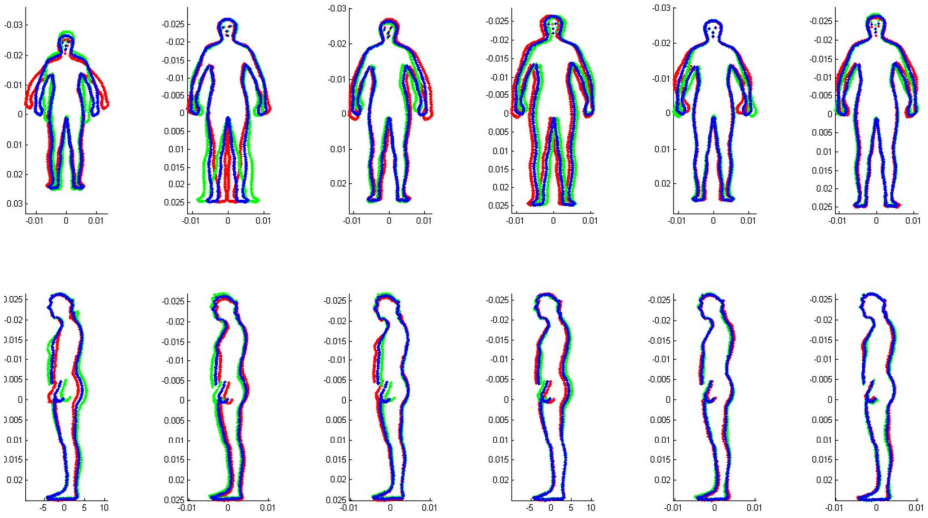

Fig. 3. Illustration of the modes of variation in the first 6 eigenvectors of our trained silhouette multi-view ASM for the frontal (top) and left-side (bottom) views. Blue plots represent the mean shape, while the green and red plots represent a variation from this mean by an amount of -2$\sqrt{\lambda_i}$ and +2$\sqrt{\lambda_i}$, respectively, where λ_i represents the i^{th} eigenvalue and $i \in \{1, ...6\}$

Fig. 4. Generated avatar samples

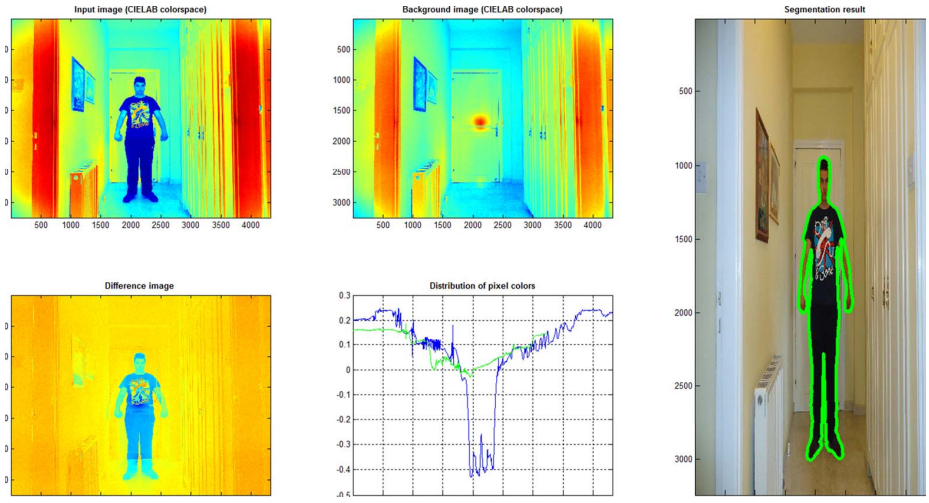

Fig. 5. Segmentation illustration: (Top left) Input image of subject converted to CIELAB colorspace, (Top middle) Input image of background converted to CIELAB colorspace, (Bottom left) Raw difference image, (Bottom middle) Average pixel values along y-axis (green graph) and average pixel values along x-axis (blue graph), which can be used to guide the selection of an appropriate segmentation threshold, (Right) Extracted silhouette

Multi-view ASM Fitting The procedure for fitting the trained multi-view ASM to a set of four silhouette images follows the ASM search algorithm presented in [3]. For simplicity, we chose the search function in such a way that the model drives itself towards regions of high gradient, i.e. strong edges, which tend to coincide with foreground-background boundaries. In each iteration, each landmark searches a window in its neighborhood for the point of strongest gradient and moves towards it. Once all points have moved to new locations, the new shape vector is projected to the ASM shape manifold to regularize it and the resulting shape parameter vector is truncated to maintain shape plausibility. The ASM search terminates once the 2-norm of the shape parameter vector stabilizes within an ϵ-value between successive iterations (see Fig. 6).

3D Model Fitting Silhouette points on the reconstructed 3D shape when projected to each of the four views, should minimize the RMS error with the corresponding silhouette points in the input images. Therefore, we seek to find the shape parameter vector, $\boldsymbol{b}^*$, which minimizes the constrained objective given below:

$$\boldsymbol{b}^* = \arg\min_{\boldsymbol{b}} \sum_i \left(\frac{1}{N_i} \| \boldsymbol{P}_i \left(\bar{\boldsymbol{x}} + \boldsymbol{C}\boldsymbol{b} \right) - \boldsymbol{Q}_i \boldsymbol{Y}_i \|_2 \right) \ , \tag{3}$$

$$\text{such that. } b_j \in [-2\sqrt{\lambda_j}, +2\sqrt{\lambda_j}] \ , \tag{4}$$

where $i \in \{\text{front}, \text{left}, \text{right}, \text{back}\}$, $\boldsymbol{P}_i$ is the projection matrix, $\boldsymbol{Y}_i$ is the matrix of 2D coordinates of the silhouette contour in the input image, $\boldsymbol{Q}_i$ is the alignment transformation matrix and N_i is the number of silhouette contour points in the i^{th} view.

We solve the problem in (3) using CVX, a package for specifying and solving convex programs [18], [19].

Texture Mapping Once we have the shape of the avatar we fill-in the realistic appearance by mapping texture information from the input images to the projected model vertices using the cylindrical model described in [2]. This involves converting the coordinates of each reconstructed 3D vertex to polar coordinates to determine which image needs to be looked up to get the texture. Then the 3D coordinates are projected to 2D image coordinates (using $\boldsymbol{P}_i$ from (3) above) to get the color value at the projected image pixel.

In order to obtain an avatar with a more realistic appearance, we then register the facial landmarks (nose, eyes, mouth) of the higher resolution face model (11655 vertices) to corresponding landmarks on the full-body 3D model. This allows us to determine the rigid transformation that aligns the two models. Hence, we transform the higher resolution face model and apply the texture mapping process again, transferring the realistic appearance to the high resolution face model (sample results shown in Fig. 4).

4 Evaluation

In order to evaluate the ability of the trained model to reconstruct a subject's 3D body shape when given four orthogonal images of the body's 2D projection, we conducted the following experiment. Using the trained PCA model of 3D body shape, we trained regression functions [4] for various physical shape-controlling parameters e.g., height. In this way, we were able to synthesize 100 new shape models (50 male, 50 female) having desired physical measurements:

- Weight: varied uniformly in the range [50kgs, 120kgs]
- Height: varied uniformly in the range [150cm, 200cm]
- % Muscle: varied uniformly in the range [0%, 100%]

From each synthetic shape we constructed four orthogonal images of its 2D projection (front, back, left, right). Landmark correspondences between image and model points were manually marked. The goal was then to recover the 3D synthetic shape when presented with only its four projection images and the manually marked landmark correspondences, and assess the 3D shape reconstruction accuracy. The shape parameters were estimated by minimizing a cost function of the average error between the image landmarks and the corresponding projected landmarks of the model. We performed the experiment twice, once using on average only 17 landmarks per view (selected similarly to the feature points in [2]) and once using 144 landmarks per view (sampled regularly around the silhouette outline). For each shape we then calculated the average relative point-to-point reconstruction error (calculated as the absolute difference between the reconstructed projection and the ground truth projection of each landmark, averaged over all 6449 landmarks; this was then divided by the maximum point to point distance between model vertices to yield the average relative point-to-point reconstruction error). The results were as follows: (i) 144 landmarks per view: $\mu = 0,81\%$, $\sigma = 0,1803$, (ii) 17 landmarks per view: $\mu = 0,90\%$, $\sigma = 0,2167$.

Finally, we calculated the sample correlations and p-values to determine correlations between the reconstruction error and the physical parameters. We found that the most significant correlations of reconstruction error were between weight and % muscle. More specifically, for weight we got $\rho = 0.2685$ and p-value=0.0069, indicating that error increases with increasing weight. For % muscle we got $\rho = 0.3503$ and p-value=0.0004, indicating that error increases with increasing % muscle because higher % muscle cause more bulging of the body thus greater deviation from the mean shape.

In another experiment, we evaluated the accuracy of the multi-view ASM against the accuracy of four separate single-view ASM models. The results, which are shown in Fig. 6 illustrate the superiority of the multi-view ASM. We are in the process of performing evaluation simulations of the proposed method against a large synthetic dataset as well as quantitative evaluation on real data. See Fig. 4 for a visual evaluation.

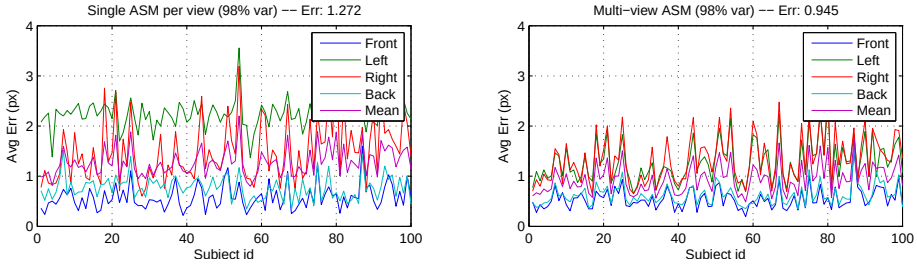

Fig. 6. Comparative evaluation of our silhouette registration method by a multi-view ASM on a synthetic dataset of 100 subjects. (Left) Plot of average relative silhouette registration error for each view, using single-view ASM models (98% variance). (Right) Corresponding plot evaluating silhouette registration error using a multi-view ASM model.

5 Conclusions and Future Work

We presented our ongoing work towards the generation of realistic animation ready avatars that will allow users and their friends to participate and collaborate in VR game adaptations. While this is work in progress and additional evaluation experiments are under way, early results obtained prove the potential of our overall approach. In the future we plan to increase the accuracy of model fitting and texture mapping so that the end result is cleaner and more realistic. Furthermore, we plan to add automatic rigging and animation on the personalized avatars, so that they can be easily incorporated in collaborative VR applications.

Acknowledgements

This work was partially supported by the Cyprus Research Promotion Foundation and the European Union Structural Funds (project VR-CAVE: IPE/NEKYP/0311/02). We would like to thank Dr Nils Hasler for providing the 3D human body database [4].

References

1. Michael, N., Kater, A.E., Lanitis., A.: Increasing user engagement in re-designed classic video games. In: Procs. of Joint Conference on Virtual Reality. (2013)
2. Hilton, A., Beresford, D., Gentils, T., Smith, R., Sun, W., Illingworth, J.: Whole-body modelling of people from multiview images to populate virtual worlds. The Visual Computer **16**(7) (2000) 411–436
3. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding **61**(1) (1995) 38–59
4. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. Comput. Graph. Forum **28**(2) (2009) 337–346
5. Viola, P., Jones, M.: Robust real-time object detection. In: International Journal of Computer Vision. (2001)
6. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. ACM Trans. Graph. **24**(3) (July 2005) 408–416
7. Lee, W., Gu, J., Magnenat-thalmann, N.: Generating animatable 3d virtual humans from photographs. In: Computer Graphics Forum. (2000) 1–10
8. Ahmed, N., de Aguiar, E., Theobalt, C., Magnor, M., Seidel, H.P.: Automatic generation of personalized human avatars from multi-view video. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology. VRST '05, New York, NY, USA, ACM (2005) 257–260
9. Starck, J., Hilton, A.: Model-based multiple view reconstruction of people. In: ICCV, IEEE Computer Society (2003) 915–922
10. Weiss, A., Hirshberg, D.A., Black, M.J.: Home 3d body scans from noisy image and range data. In Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V., eds.: ICCV, IEEE (2011) 1951–1958
11. Cui, Y., Chang, W., Nöll, T., Stricker, D.: Kinectavatar: Fully automatic body capture using a single kinect. In: Proceedings of the 11th International Conference on Computer Vision - Volume 2. ACCV'12, Berlin, Heidelberg, Springer-Verlag (2013) 133–147
12. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(8) (2010) 1362–1376
13. Starck, J., Miller, G., Hilton, A.: Video-based character animation. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. SCA '05, New York, NY, USA, ACM (2005) 49–58
14. Tong, J., Zhou, J., Liu, L., Pan, Z., Yan, H.: Scanning 3d full human bodies using kinects. IEEE Trans. Vis. Comput. Graph. **18**(4) (2012) 643–650
15. Vlasic, D., Peers, P., Baran, I., Debevec, P.E., Popovic, J., Rusinkiewicz, S., Matusik, W.: Dynamic shape capture using multi-view photometric stereo. ACM Trans. Graph. **28**(5) (2009)
16. Mashalkar, J., Bagwe, N., Chaudhuri, P.: Personalized animatable avatars from depth data. In: Proceedings of the 5th Joint Virtual Reality Conference. JVRC '13, Aire-la-Ville, Switzerland, Switzerland, Eurographics Association (2013) 25–32
17. Kendall, D.G.: A survey of the statistical theory of shape. Statistical Science **4**(2) (05 1989) 87–99
18. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., Kimura, H., eds.: Recent Advances in Learning and Control. Lecture Notes in Control and Information Sciences. Springer-Verlag Limited (2008) 95–110 `http://stanford.edu/~boyd/graph_dcp.html`.
19. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx` (March 2014)