

3D Body Reconstruction from Photos Based on Range Scan

Hyewon Seo¹, Young In Yeo² and Kwangyun Wohn²

¹ CGAL, Chungnam National University
220 Gung-dong, Yuseong-gu, Daejeon, Korea
hseo@cnu.ac.kr

<http://cs.cnu.ac.kr/~hseo>

² VRLab., KAIST 373-1 Guseong-dong, Yuseong-gu Daejeon, Korea
yyiguy, wohn@vr.kaist.ac.kr
<http://vr.kaist.ac.kr/>

Abstract. We present a data-driven shape model for reconstructing human body models from one or more 2D photos. One of the key tasks in reconstructing the 3D model from image data is shape recovery, a task done until now in utterly geometric way, in the domain of human body modeling. In contrast, we adopt a data-driven, parameterized deformable model that is acquired from a collection of range scans of real human body. The key idea is to complement the image-based reconstruction method by leveraging the quality shape and statistic information accumulated from multiple shapes of range-scanned people. In the presence of ambiguity either from the noise or missing views, our technique has a bias towards representing as much as possible the previously acquired ‘knowledge’ on the shape geometry. Texture coordinates are then generated by projecting the modified deformable model onto the front and back images. Our technique has shown to reconstruct successfully human body models from minimum number images, even from a single image input.

1 Introduction

One of the oldest goals in computer graphics has been the problem of reconstructing human body to convincingly depict people in digital worlds. Indeed, digitally modeling human bodies from measurement is being actively and successfully addressed by image-based and hybrid techniques. During its formative years, researchers have focused on developing methods for modeling appearance and movements of real people observed from 2D photos or video sequences[6][7][9]. These efforts use silhouette information from multi-view images for determining the shape and optionally the texture of the model to be reconstructed. To simplify the problem of general reconstruction, a template or generic model of the class to be modeled – human body – is fit to observations of a particular subject.

Today, whole body range scanners are becoming more and more available and hence much of the focus of graphics research has been shifted to the acquisition

of human body models from 3D range scans [2][10]. The measurements acquired from such scanning devices provide rich set of shape information which otherwise requires considerable amount of time and effort by experienced CG software users. Range scanners however remain by far more expensive, difficult to use, and provide limited accessibility, compared to 2D imaging devices. Moreover, many whole body scanners today provide only geometry data without color or texture [5][13][14].

We note that combining 2D images and the range scanned measurement can lead to successful reconstruction results. The quality shape and collective knowledge from scanned dataset can efficiently be used to complement the shape recovery from image inputs. In the domain of human face reconstruction, for example, one of the most impressive 3D reconstructions of human faces was presented by Blanz and Vetter[3]. They described a face modeler in which a prior knowledge collected from a set of head scan data is exploited to find the optimizing surface and texture parameters that best fit the given image data. While their method suggest a powerful approach to the image-based human modeling, it has not been applied to model human body. Indeed, it would be difficult to extend that method to modeling of entire human body, due to large-scale occlusions and articulations.

In this paper, we propose a system for reconstructing an entire human body model from minimum number of multi-view images, exploiting the quality shape captured by range scans. Our specific goal is an on-line clothing store[4] where users can try on garment items on their 3D avatar models, hence we limit our focus to the reconstruction of lightly clothed subjects. The distinguishing aspect of our modeler in comparison to existing image-based body modeler is that it employs a data-driven shape model that has been constructed from range scans of real bodies. Subsequently, we exploit the quality shape as well as statistical information collected from a database of laser-scanned bodies, in the presence of ambiguity, noise, or even underconstraints caused by missing views.

1.1 Related work

While image-based model reconstruction has been at the center of digital human modeling across several research groups, the majority of research progress in this avenue falls into the category of facial modeling. This is perhaps primarily due to the complex articulated structure and high degree of self-occlusion exhibited in our bodies.

One approach that has been extensively investigated is model-based techniques. Hilton et al[6] and Lee et al[7] have gathered silhouette observations from multiview images such that they are used to transform a template humanoid model. Affine transformation has been followed by geometric deformation of the prior surface model. They use feature point locations along the silhouette to find the correspondence among different views and to generate consistent texture coordinates. More recently, Sand et al[9] use multi-view recordings to derive the skeleton configuration of a moving subject, which subsequently derives the skin surface shape. These works show how a prior knowledge can be used to avoid dif-

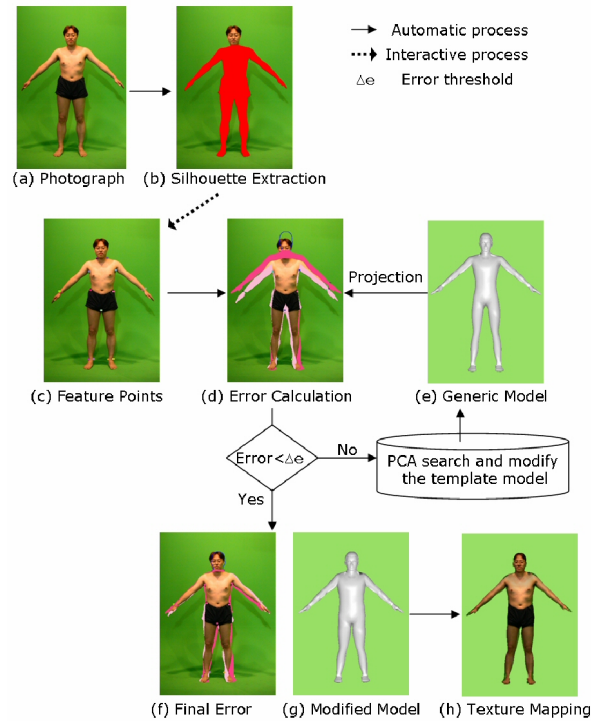


Fig. 1. Overview Of Our Approach

ficulties of general reconstruction. However, they do not accumulate observations which can efficiently be used to handle uncertainties.

The strength of gathering information from collective observation has been illustrated in face model acquisition by Blanz and Vetter[3]. In their modeler, a highly detail 3D face shape and texture spaces have been obtained, by transforming some two hundred laser-scanned faces into vector representation. Given a single photograph of a face, its 3D shape, orientation in space, and the illumination conditions are estimated. Starting from a rough initial estimate of shape, surface color, and lighting parameters, an optimization algorithm iteratively finds the best matching parameters to the input image. Shape and texture constraints derived from the statistics of our example faces are used to guide automated matching. While these methods are quite powerful, they have not been applied to image-based reconstruction of an entire human body. These considerations lead us to look for a more robust approach to image-based human body modeling. Based on our previous work [11], we adopt a model parameterization scheme based on a collection of observations of real bodies. Also adopted is a surface optimization framework, in order to match multiple camera images. As a

result, our technique handles complex articulated structure of the entire human body, and still runs at an arguably interactive speed.

1.2 Organization

An overview of our approach is illustrated in Figure 1. In Section 2, we describe the way we take photographs of a subject, and extract silhouettes and feature points on the images. We then use a deformable model that is based on the previous work of Seo and Magnenat-Thalmann[10] for the shape recovery, as detailed in Section 3. Using the silhouette data extracted from the input images, we explore the body space (a range of coefficient parameters that have been spanned by the database of the deformable model) and find the best fitting deformation parameters on a template model. Finally, Section 4 describes the method we use to generate texture coordinate data by projecting the deformed template model onto the input images. After demonstrating the result models in Section 5, we conclude the paper in Section 6.

2 Acquiring and processing input images

2.1 Taking photographs

We first take a minimum number of photographs of a subject, to acquire observations on a subject. Our system in principle does not require any special camera arrangements, nor does it require specific number of views. In actuality, however, at least two views – one from the front and the other from the back – are preferred, as we want to generate complete texture on the entire body. This is because our deformable model does not contain color data (See Section 5.1). In our experiments, we also found that adding side view makes considerable improvement on the characteristics of the body shape. Hence, we mostly take three photographs using a single camera, each from the front, the side, and the back sides of the subject, unless otherwise specified.

Throughout this paper, we assume that the subjects are lightly clothed. To simplify the combinatorial complexity of the human shape and posture, we require the subject to stand in the specific posture; the limbs are straight and apart from the torso as shown in Figure 1. We also require that photos are taken in front of a blue backdrop, to facilitate automatic silhouette extraction. In this paper, images have been captured by the Nikon coolpix 5000 camera and stored in 1920 by 2560 resolutions for the texture mapping. They have been reduced to 480 by 640 resolutions for speedup during the silhouette comparison.

2.2 Virtual camera setup for the template model projection

We now describe the camera arrangements and projection matrix we use for projecting the template model onto the image space. The main idea is to simulate virtual camera as closely as possible to the physical setup. This allows us to use

input images directly for the silhouette comparison without additional process such as image size normalization.

In this paper, we adopted Tsai’s Pinhole camera model[12], which basically is a pinhole model taking the 1st order radial lens distortion into account. A camera has 5 intrinsic parameters (focal length f , 1st order radial lens distortion value $Kappa$, center of lens distortion C_x, C_y , scale factor S_x) and 6 extrinsic parameters 6 ($R_x, R_y, R_z, T_x, T_y, R_z$). To calculate these intrinsic and extrinsic parameters, we have taken an image of a calibration frame, similarly to the approach presented by Ahn et al[1]. The intrinsic parameters we found are illustrated in Table 1. The intrinsic parameters like $Kappa, C_x$, and C_y are used to calculate the degree of distortion for each pixel. In our experiments, the degrees of distortion were less than one pixel, and thus were simply discarded from the projection matrix.

Table 1. Intrinsic parameters calculated for our camera

f	$Kappa$	C_x	C_y	S_x
21.7mm	0.0000314	247.0	308.8	0.99

Next, we setup our virtual camera from the measured parameters of the physical camera; the camera is approximately 5 meters distant from the template model and 1.2 meters from the ground. From the basic trigonometry we have obtained the FOVy (Field of View y) angle of 25.4 degrees. FOVy determines the perspective projection matrix we use for projecting the template model onto the image space.

2.3 Silhouette extraction and feature point identification

We take photos in front of a uniformly colored screen so that simple methods such as using color key can be used for the automatic silhouette detection. The method we use is a standard background subtraction to isolate silhouettes from images using a color key. Among several color models, we use the Hue-Saturation-Value (HSV) color model. HSV model is commonly used in interactive color selection tools, be-cause of its similarity to the way how humans perceive colors. We first map each pixel in the image to the color space defined by the HSV hexagonal cone. Pixels in the background region form a cluster in the HSV space, as illustrated in Figure 2. In our experiments, we first project an image of the background drape to obtain the range of background pixel clusters. The cluster is defined by the H value of $180^\circ \sim 240^\circ$, and S value larger than a threshold, say, ‘0.3’. As a subject stands in front of the background, shadows appear and they contribute to the background clouds elongated downwards along the V axis of the hexagon. Thus, we use color keys in H and S to determine the background

pixel cluster. As illustrated in Figure 2, shadows have been successfully labeled as background.

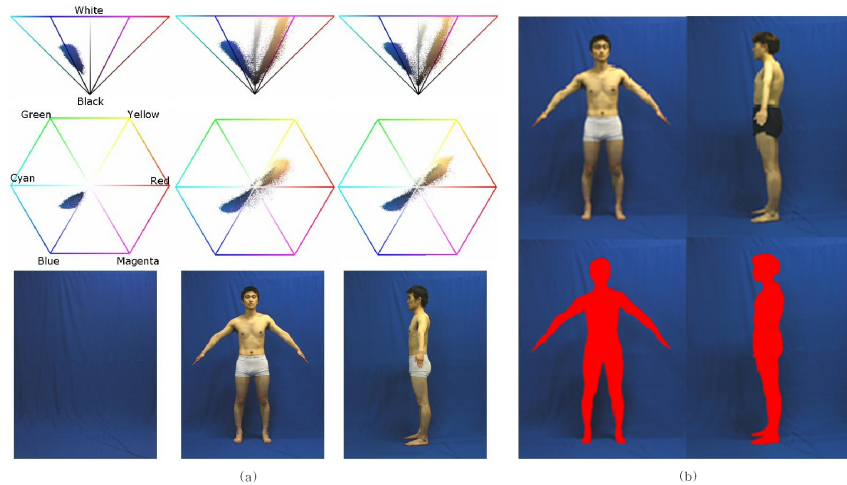


Fig. 2. (a) Projection of images onto the HSV color space: Empty background (left), front(middle) and side (right) views. (b) Silhouette extraction results of two subjects with white (left) and black (right) underwears

Next, we label 12~15 feature points on the silhouettes. In addition to the silhouette data as described earlier, we make use of a number of feature points when matching the template model to the target subject in the image. Using features points allows to deform the template not only to match the silhouette but also to ensure the correspondence. Unfortunately, only a limited set of feature points can be found automatically, such as those on the top of the head, the bottom of the feet and the tip of hands. For the rest of the feature points, the user manually places them on the images. Feature points on the template model are identified in a similar way on the 3D mesh(see Figure 3).

3 Shape recovery by searching deformation space

3.1 Parameterized body space construction

For the shape recovery we use a parameterized shape model that is based on our previous work[10]. In that modeler, each of the scanned body shape is represented by combining two shape vectors that are used to deform the template. These vectors respectively encode the skeleton-driven deformation (joint vector) and vertex displacement of a template model (displacement vector) that is necessary

to reproduce its shape. By collecting these shape vectors from multiple subjects, we obtain what we call the body space.

Given a set of example body shapes represented as vectors, we apply principal component analysis (PCA) to both joint and displacement vectors. The result is two linear models for the two components:

$$\begin{aligned} \mathbf{j} &= \bar{\mathbf{j}} + \mathbf{P}_j \mathbf{b}_j \\ \mathbf{d} &= \bar{\mathbf{d}} + \mathbf{P}_d \mathbf{b}_d \end{aligned}$$

where $\bar{\mathbf{j}}$ and $\bar{\mathbf{d}}$ are the mean vector, \mathbf{P}_j and \mathbf{P}_d are sets of orthogonal modes of variation, and \mathbf{b}_j and \mathbf{b}_d are the sets of parameters. The appearance of any body models can thus be represented by the set of PC coefficients of joint vector \mathbf{b}_j and that of displacement vector \mathbf{b}_d . A new model can be synthesized for a given pair \mathbf{b}_j , \mathbf{b}_d by deforming the template from vector \mathbf{j} and adding the vertex displacement using the map described by \mathbf{d} .

Note that the PCA has the additional benefit that the dimension of the vectors can drastically be reduced without losing the quality of shape. Upon finding the orthogonal basis, the original data vector \mathbf{v} of dimension n can be represented by the projection of itself onto the first $M (\ll n)$ eigen vectors that correspond to the M largest eigen values. In this paper, we have used 30 bases both for the \mathbf{b}_d and \mathbf{b}_j . Thus, each body is represented as a set of parameter vector consisting of 30 PC's for the joints and 30 for displacement, giving a total of 60 parameters for the body shape space.

3.2 Search-based deformation

Building such parameterization of the body space as described in the previous section permits us to easily and efficiently explore the coefficients, thus simplifying the complex process of acquiring geometry of full body. Given a set of images, we use the extracted silhouette information to reconstruct the geometry by searching the body space (a range of coefficient parameters that have been spanned by the database of the deformable model) and finding an optimum parameter set. A set of coefficient parameters comprises an optimum solution if, when collectively applied to the template model, it produces silhouettes that best fit the given image silhouettes. The key point is that instances of the models are deformed in ways that are found in the example set, guaranteeing a realistic, robust shape acquisition.

We find the solution in a coarse-to-fine manner. Since the deformation is parameterized with PCA space for each of the vector components, we first find the optimizing joint parameter \mathbf{b}_j , followed by the subsequent search for the \mathbf{b}_d in the displacement vector space. Our optimization technique is based on a direction set method [8]. The algorithm repeats ‘search-deform-compare’ loop until we obtain sufficient degree of matching between the silhouette of the deformed model and that of the input image – It generates a body shape from the current coefficients, projects the body model onto 2D space, and updates the coefficients according to the silhouette difference. The first set of iterations is performed by

optimizing only the first coefficients controlling the first few PCs. In subsequent iterations, more PCs are added to further deform the template.

3.3 Error metric

One important step in our modeler is to measure the silhouette matching error between the segmented images to projections of the template under deformation. We consider two error terms: The first one is the distance between corresponding feature points. The second one is the silhouette error.

Distance between corresponding feature points The first criterion of a good match is the distance between corresponding feature points in the image space. We define the distance error term E_d as the sum of the squared distances between each corresponding feature point’s location in the data image and its location on the projected image of the template mesh:

$$E_d = \sum dist^2(P(F_{T,i}), F_{D,i}), (i=1...n),$$

where n is the number of feature points and $dist$ is the Euclidean distance among two pixels in the image, $F_{D,i}$ is the i -th feature pixel in the image, $F_{T,i}$ the corresponding i -th feature point on the template model, and $P:R^3 \rightarrow R^2$ describes the perspective projection of the template mesh to the 2D images.

We consider a sparse set of feature points that are important for calculating joint configurations (scale, and rotation of each joint except for the root that has translation) of the template model. We have found that feature points around the neck, arm-pits, wrists, crotch and ankles are particularly important, as they undergo relatively high degree of transformation for a matching. Note that they overlap pretty much with anthropometric landmarks as well. 27 points were manually placed on the template mesh prior to projection to the images. On the images, 15 and 12 feature points were defined on the front and side views, respectively. In addition to approximating distance error, this term also insures the correspondence at the feature locations. In Figure 3, corresponding feature points are represented with the same color.

Silhouette error Using the distances among each feature point location will not result in a successful matching, because even though corresponding feature points are in the same position, actual body shapes can be different from each other. To acquire detailed match of the template model to the image, we define a silhouette error and denote as E_a . By silhouette error we refer to the fraction of pixels for which the projected and observed silhouettes do not match, as shown in Figure 3. The number of background pixels that lie inside the projected template model is summed up with that of foreground pixels that lie outside of it:

$$E_a = \frac{\sum(T(i,j) \cdot \bar{D}(i,j))}{\sum T(i,j)} + \frac{\sum(\bar{T}(i,j) \cdot D(i,j))}{\sum D(i,j)}$$

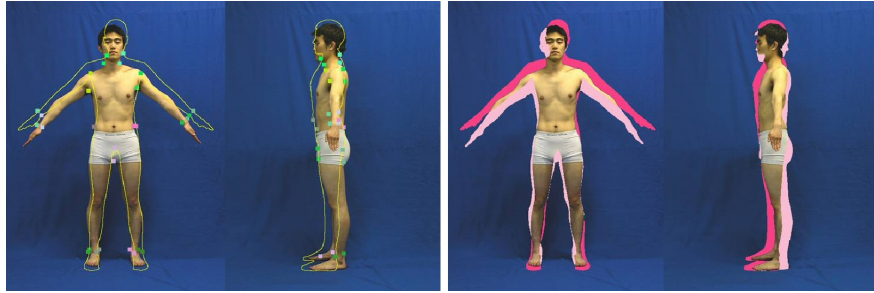


Fig. 3. Distance between corresponding feature points (left) and non-overlapping area error calculated(right) on the front and side photos of a male subject (right)

$T(i,j)$, and $\bar{T}(i,j)$ are the boolean values indicating if the pixel at location (i,j) is inside and outside of the template model, respectively. $D(i,j)$ and $\bar{D}(i,j)$ are 1 if the pixel located at (i,j) is foreground and background, respectively. This notion of non-overlapping area is effectively equivalent to the silhouette error used by Sand et al[9].

Combining the error We use weighted sum of the two error terms, as denoted by $E = \alpha E_d + (1-\alpha)E_\alpha$. At the first iterations we need to quickly search for joint parameters, hence we set $\alpha=1$. Feature points from both the frontal and the side images are measured. Next, we further improve the fitting accuracy by setting $\alpha=0.3$. The deformable model is first fit to the frontal image and then side image error is added. Finally, the displacement map is explored with the same setting. At each iteration, we combine the errors from the frontal image and the side image, so that the fitting of the template to frontal and side images can simultaneously be handled.

4 Mapping textures

One of our goals is to generate color data of the model, to maximize visual effects. Photo images are then used to generate texture on the geometry obtained from the above phase. Although we require the subject to keep consistent poses among different views, they may be slightly different from one view to another, as they are temporally distinct views. To handle such inconsistency, we use only the front and the side images for the shape recovery, and we handle the texture coordinate creation process for the front and the back parts separately.

Two separate texture coordinates are obtained by projecting the deformed template model onto the front and the back images: If the angle between vertex normal and the view direction is between $-\pi/2$ and $\pi/2$, we project the vertex on the deformed template surface onto the front image. The other vertices are projected on to the back image. Prior to the second projection, we must adjust

the posture of the model by matching the template with the silhouette data on the back image. This is due to the slight difference among postures seen from one view to another.

5 Results

5.1 Dataset

Whole body scans of 40 male bodies of European adults, recorded with TecmathTM laser scanners[13] were used in our experiments. The scanner does not provide color data, and thus we have used input photos to generate texture maps. The dataset was originally captured for made-to-measure garment retails. Analogous to those subjects for the image-based reconstruction, these subjects for the three dimensional scan were lightly clothed without accessories, and were in a standing posture with arms and legs slightly apart. Additionally, the face was removed digitally via a vertical cut anterior to forehead, with manual editing on the point cloud.

5.2 Model reconstruction

We have applied our modeler to a number of example images. Some of the result models are illustrated in our video demo. In all examples, we matched the template model built from the first 30 joint and 30 displacement principal components that were derived from the whole body scan dataset, as described in the previous section. Once the silhouettes have been extracted, the whole matching procedure was performed in less than 1000 iterations for each principal component (PC), taking a total of 6~7 minutes on a Pentium 4 processor when 3 PCs of the joint vector and 3 PCs of the displacement vector are optimized.

5.3 Single image input

Because our modeler is based on optimization-based searching and not on an entirely geometric technique, we can handle some ambiguous situations robustly. To demonstrate such robust behavior of our modeler, we have reconstructed a 3D geometry using only a single image input. In Figure 4(a), we used only the front image of the subject to reconstruct the shape of the template model. Analogously, only the side image of the subject was used to reconstruct the shape shown in Figure 4(b). In both cases, a back view image was used to complete the texture map.

6 Conclusion

We have presented a technique for reconstructing human body model from a limited number of 2D images. Using the three-dimensional body space that has been generated from processing range scans, we propose reconstructing the 3D

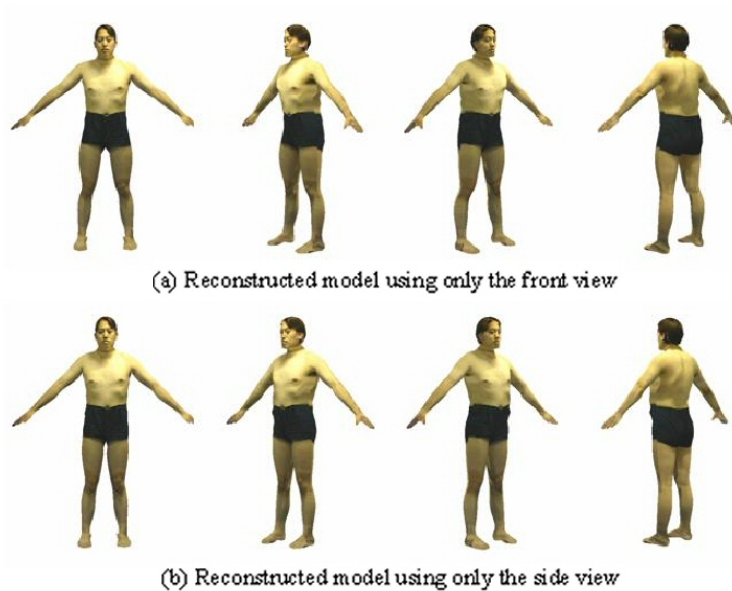


Fig. 4. A reconstructed model by using a single image input. (a) front image (b) side image

surface in a different manner than existing modelers. For the shape recovery we start with a deformable template model whose deformation is parameterized with PCA of the scanned body shapes. Given a set of images, the optimizing shape is found by searching the shape space, such that it minimizes the matching error measured between silhouettes. The idea is to start from a space consisting of a few PCs and to increase its size by progressively adding new PCs. This provides us powerful means of matching the template model to the image in a coarse-to-fine manner. In addition, a high level of detail and accuracy is acquired, since our modeler essentially blends multiple shapes of the human body acquired from 3D laser scanners. This constitutes a good complement to geometric methods, which cannot capture detailed shape solely from image input.

Our system is intended for off-line reconstruction of geometry, but it is reasonably efficient and can also be adopted for on-line applications. When we reduce the number of PCs to be optimized during searching, say 4~5 for each of the shape vector, we can obtain fairly good results without losing much of the accuracy. Additionally, we have successfully integrated our models in animation systems. We are currently exploring an extension of this technique with which we can reconstruct the body shape even when given images of casually dressed subjects.

Acknowledgements

The authors are grateful to Gun-Woo Kim, who has helped us with the formatting of the paper. This work has been supported in part by Chungnam National University, BK21 project of EECS Department at KAIST, and MMRC project funded by SK Telecom.

References

1. Ahn J.H., Sul C.H., Park E.K., Wohn K.W., "Lan-based Optical Motion Capture System", pp.429-432, Proc. Korea Conference on Human Computer Interaction 2000.
2. Allen B., Curless B., Popovic Z., "The space of human body shapes: reconstruction and parameterization from range scans", Proc. SIGGRAPH '03, pp.587-594, Addison-Wesley, 2003.
3. Blanz B. and Vetter T., "A morphable model for the synthesis of 3D faces", Proc. SIGGRAPH '99, Addison-Wesley, pp. 187-194, 1999.
4. Cordier F., Seo H., Magnenat-Thalmann N., "Made-to-Measure Technologies for Online Clothing Store", pp. 38-48, IEEE CG&A special issue on Web Graphics, January 2003.
5. Hamamatsu BL scanner, <http://www.hpk.co.jp>
6. Hilton A., Beresford D., Gentils T., Smith R.J., Sun W. and Illingworth J., "Whole-body modelling of people from multi-view images to populate virtual worlds", Visual Computer: International Journal of Computer Graphics, 16(7), pp. 411-436, 2000.
7. Lee W., Gu J., and Magnenat-Thalmann N., "Generating Animatable 3D Virtual Humans from Photographs", Computer Graphics Forum, vol. 19, no. 3, Proc. Eurographics 2000 In-terlaken, Switzerland, August, pp. 1-10, 2000.
8. Press W. H., Flannery B. P., Teukolsky S. A., and Vetterling W. T., "Numerical Recipes in C", The art of scientific computing. Cambridge University Press, 1988.
9. Sand P., McMillan L., Popovic J., "Continuous Capture of Skin Deformation", Proc. ACM SIGGRAPH 2003, pp.578-586.
10. Seo H. and Magnenat-Thalmann N., "An Automatic Modeling of Human Bodies from Sizing Parameters", ACM SIGGRAPH Symposium on Interactive 3D Graphics, ACM Press, pp. 19-26, 2003.
11. Seo H., "Parameterized Human Body Modeling", PhD thesis, Departement d'informatique, University of Geneva, 2004.
12. Tsai R. Y., "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 1986, pp. 364-374.
13. Tecmath AG, <http://www.tecmath.com>.
14. Telmat Industrie SA, <http://www.telmat.com>.