Making Them Alive

Henry Braun, Humberto Souto Júnior, Júlio C. S. Jacques Júnior, Leandro L. Dihl,
Adriana Braun, Soraia Raupp Musse
Pontifícia Universidade Católica do Rio Grande do Sul
Cláudio Rosito Jung — Universidade Federal do Rio Grande do Sul
Marcelo R. Thielo — HP Brazil
Renato Keshet — HP Israel

Abstract

This paper presents a model to reconstruct 3D virtual humans based on a single and spontaneous image. The main goal is to use computer vision and pattern recognition techniques to build coherent virtual humans according to an input picture. To achieve this goal we provide a semi-automatic process that includes 3D posture detection, segmentation of human body parts, and silhouette processing. Such information is used to generate a 3D virtual human, which can be further animated. The approach proposed in this paper aims to speed up the creation of 3D articulated characters, providing avatars based on pictures. Experimental results indicate that our approach is a good option for generating virtual humans from images based on a few mouse clicks.

Keywords:: Human segmentation, virtual human reconstruction, Digital Games

Author's Contact:

1 Introduction

The reconstruction of virtual humans (VHs) is important for several applications, such as games and simulations. Lately, the game industry has achieved a large number of players around the world, pushing game developers to create new gameplay experiences. One of these new experiences is to allow the player to use her/his own appearance-like avatar while playing, which is already present in several Electronic Arts sports games (`http://www.ea.com/games`). However, in order to generate such virtual humans, artists usually spend a large amount of time modeling the characters.

Other applications for VH reconstruction include Collaborative Virtual Environments (CVEs) and Virtual Reality (VR) scenarios. In some cases, there are manual interfaces for character customization. One can imagine that such interfaces could be replaced by a semi-automatic approach, where users could just choose a picture for customizing their avatars instead of doing it manually. This approach could also be employed to avoid the work of generating several different characters in games and real-time applications. However, some challenges arise in VH reconstruction based on a single and spontaneous image. Firstly, 2D human segmentation in images is still an open research problem, mainly due to clutter background, varying illumination conditions, and a wide variety of human postures. Secondly, the 3D human pose in the picture should be known in order to build the virtual human coherently using spontaneous pictures. Finally, textures and animations can be applied to the generated VH.

In this paper we present a new model for the reconstruction of animatable virtual humans using information processed in pictures (human segmentation), 3D pose and silhouette data (extracted from the picture containing width values for the human parts). Our technique generates an articulated 3D model of a character that can be further animated. The remainder of this paper is organized as follows: related work are described in the next section, followed by details of our model in Section 3. Some results are shown in Section 4, while future work and final considerations are discussed in Section 5.

2 Related Work

Several techniques have been proposed to build virtual humans based on pictures (e.g. [Hasenfratz et al. 2003]) and/or video sequence (e.g. [Fua 1999]). One of the pioneer work concerning VH reconstruction based on pictures was proposed by Hilton et al. [Hilton et al. 1999]. The main idea of their work is to generate human models, aiming to populate virtual worlds. For that purpose, they used four human silhouettes (front, back and both sides) against four silhouettes obtained by a VRML2 (Virtual Reality Modeling Language) generic model. The comparison of these silhouettes in a 2D universe reveals pixels displacement in relation to each other, which are mapped to the 3D generic model. At last, the generic model is colorized, using an approach of cylindrical texture mapping [Thalmann N. 1997].

In 2000, Lee et al. [Lee et al. 2000] proposed a system that utilizes photos taken from the front, side and back of a person in any given imaging environment without requiring a special background or a controlled illuminating condition. The system is composed of two major blocks: face-cloning and body-cloning, using feature points on front and side images. The final integrated human model has photo-realistic animatable face, hands, feet and body.

In [Hasenfratz et al. 2003], the authors show how to capture an actor with no intrusive trackers and without any special background (such as blue set) to estimate his/her 3D-geometry and to insert this geometry into a virtual world in real-time. They use several cameras in conjunction with background subtraction to produce silhouettes of the actor as observed from the different camera viewpoints. These silhouettes allow the 3D-geometry of the actor to be estimated by a voxel based method. This geometry is rendered with a marching cube algorithm and inserted into a virtual world. A clear drawback of this approach is the necessity of different viewpoins, while the proposed approach aims to work on single images.

Recently, Daniel and Nadia Thalmann [Magnenat-Thalmann and Thalmann 2008] presented the latest techniques to model fast individualized animatable virtual humans for real-time applications. As a human is composed of a head and a body, they analyze how these two parts can be modeled and globally animated as in real-life. More precisely, they show how to model and deform human bodies and heads. In their presentation, however, they do not include segmentation and posture detection in single and spontaneous images.

A competitive approach is the work proposed by Guan et al [Guan et al. 2010], where they describe a solution to the problem of estimating human body shape from a single photograph or painting. Their approach computes shape and pose parameters of a 3D human body model directly from monocular image. The model requires the estimative of the subject's height and a few clicked points on the body to find out the 2D and 3D postures. They also use a shape database containing several different human models based on SCAPE [Anguelov et al. 2005] to build a coherent 3D shape. Dilated and eroded versions of the projected 3D shape are used to generate a tri-map of regions inside, outside and on the boundary of the human, which is then used to segment the image using graph cuts. Finally, authors estimate the scene lighting to produce a synthesized body that robustly matches the image evidence. As stated by the authors, they focused on naked or minimally clothed people, since clothing may affect both the 2D silhouette segmentation and the lighting estimation proposed in their paper.

Zhou et al. [Zhou et al. 2010] use a model-based approach for reshaping human bodies in a single image, not really focusing on VH

creation. They rely on a 3D morphable model of human shapes to achieve globally consistent editing of body parts. They pose the body reshaping problem as the design of a 2D warping of a human body image, and introduce a novel body-aware image warping approach incorporating changes of the morphable model.

In the next section we describe our model for the reconstruction of 3D virtual humans based on single and spontaneous images.

3 The Proposed Model

We propose a semi-automatic pipeline to generate a VH based on single and spontaneous images. By spontaneous images we mean pictures of human subjects taken under usual circunstamces, which may contain a variety of poses of the person to be reconstructed, more then one people in the picture as well as heterogeneous background. Our approach is semi-automatic, since the user must provide a few clicks locating joints of the human structure (as later explained in Section 3.1). All the rest of process is automatic, except for the gender selection, which happens in the graphical interface to generate the VH.

The pipeline is formed by four main processes, as illustrated in Figure 1. Human Segmentation, Silhouette Processing and 3D Pose identification are described in Sections 3.2, 3.2.3 and 3.3, respectively. User intervention is performed through a few clicks, as explained in Section 3.1. These processes together generate all information required to build the VH, namely: XML files containing the 3D pose, the sizes of the body parts computed trough the silhouette, and files containing the original, segmented and binary images.

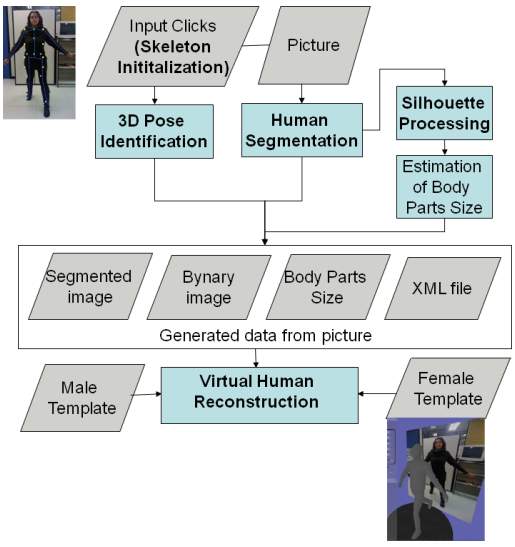

Figure 1: *Overview of our model for virtual humans reconstruction.*

The Virtual Human Reconstruction step is responsible for the VH generation. Firstly a template (female or male) is chosen by the user. Then, the 3D pose in the picture is taken into account to provide the VH posture, coherently with the original image. This step is followed by the reconstruction process, explained in Section 3.4. Finally, pieces of the segmented image, which are automatically extracted based on body parts clicked by the user, are processed to generate textures that are applied to the VH. This last phase can present better performance if post-processed by artists, and it should be noticed that texture processing is not included in the main scope of this paper. The next sections describe details of the main parts of our model.

3.1 Skeleton Initialization

Detecting humans in images is a challenging task, due to their variable appearance and the wide range of poses that they can adopt [Dalal and Triggs 2005]. As related by Hornung et al. [HORNUNG et al. 2007], interactive 2D human posture acquisition presents some advantages when compared to automatic procedures, since manual intervention usually takes just a few minutes to complete and leads to superior results in poses that are ambiguous for automatic human-pose estimators, or that are difficult to estimate due to occlusions, for example. For these reasons we also initialize our skeleton model of a human being using manual interventions – the user informs the height of the person (in pixels) and the positions of the joints (in image coordinates).

In our work, the skeleton model is composed by nineteen bones and twenty joints, as illustrated in Figure 2. All these bones have initial 3D lengths and widths, both parametrized as a function of the height h of an average person based on anthropometric values [Tilley 2002]. More precisely, for a certain body part with label i, the corresponding length l_i and width w_i are given by

$$l_i = hf_{li}, \;\; w_i = hf_{wi}, \tag{1}$$

where the proportionality factors f_{li} and f_{wi} are are derived from [Tilley 2002]. Table 1 presents all body parts used in this work, along with the corresponding values for f_{li} and f_{wi}. The initial bone lengths are used mainly in the image segmentation stage and in the 3D pose estimation procedure, which will be better detailed in Section 3.2.1.The estimated width of each body part is used in the segmentation stage, as discussed in Sec. 3.2.2.

There are two different ways to obtain the height of the person through manual intervention. When the person is standing in the photograph and the full body is visible, the user simply clicks on the top of the head and on the bottom of the feet, obtaining the height directly. In any other situation (e.g. if the person is sitting down), his/her height can be estimated based on any bone the user selects to be used as reference (including the face), and the height can be estimated based on Table 1. Since the camera parameters are not known, it is advisable to select a bone that is parallel to the image plane, to reduce the influence of perspective issues. For instance, if the user choose to use the face size as reference, the user clicks on the top of the head and on the tip of the chin, to compute the height of the face h_f. The height of the person is then estimated by $h = h_f/0.125$, where 0.125 is a weight derived from anthropometric values [Tilley 2002].

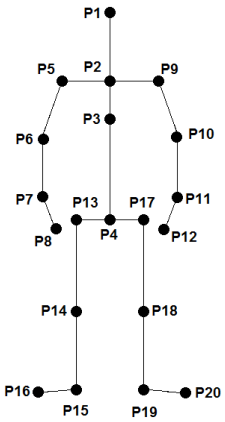

Figure 2: *The adopted skeleton model.*

3.2 Image Segmentation

This section describes the proposed approach to segment human subjects in a semi-automatic way, similarly to the automatic method described in [Jacques Jr. et al. 2010]. Although the approach in [Jacques Jr. et al. 2010] estimates each body part (from the upper body only) automatically, it is prone to errors in more complex poses, as most automatic models. In this paper, we explore manually informed data about the joints (and, consequently, body parts) to achieve better accuracy, and also to handle with lower body parts. Our approach can be divided in two main steps: (i) skeleton initialization and (ii) object segmentation. The skeleton initialization is done manually, as described in Section 3.1. The object segmentation is done automatically and it is briefly described next. Yet, it

Table 1: *In the first column: the body part index; in the second column: the body part (bone); in the third column: the two joints that form each bone; in the fourth column: the weights used to compute each bone length; and in the fifth column: the weights used to compute each bone width.*

i	Bone	Joints	f_{li}	f_{wi}
0	Head	(P1 - P2)	*0.20*	*0.0883*
1	Chest	(P2 - P3)	*0.098*	*0.1751*
2	Abdomen	(P3 - P4)	*0.172*	*0.1751*
3	Right Shoulder	(P2 - P5)	*0.102*	*not used*
4	Right Arm	(P5 - P6)	*0.159*	*0.0608*
5	Right Foreman	(P6 - P7)	*0.146*	*0.0492*
6	Right Hand	(P7 - P8)	*0.108*	*0.0593*
7	Left Shoulder	(P2 - P9)	*0.102*	*not used*
8	Left Arm	(P9 - P10)	*0.159*	*0.0608*
9	Left Foreman	(P10 - P11)	*0.146*	*0.0492*
10	Left Hand	(P11 - P12)	*0.108*	*0.0593*
11	Right Hip	(P4 - P13)	*0.050*	*not used*
12	Right Thigh	(P13 - P14)	*0.241*	*0.0912*
13	Right Calf	(P14 - P15)	*0.240*	*0.0608*
14	Right Feet	(P15 - P16)	*0.123*	*0.0564*
15	LeftHip	(P4 - P17)	*0.050*	*not used*
16	LeftThigh	(P17 - P18)	*0.241*	*0.0912*
17	LeftCalf	(P18 - P19)	*0.240*	*0.0608*
18	LeftFeet	(P19 - P20)	*0.123*	*0.0564*

is important to mention that any other method of people segmentation can be used (e.g. based on graph or grab cuts [Rother et al. 2004], [Boykov and Jolly 2001]). We develop a method that could be used with the same input that posture estimation instead of having any other intervention in the image.

3.2.1 Learning the color model

In this paper we propose a method to segment human body parts in images, based on dominant colors. To do this, we firstly create a color model for each body part, and estimate a search region. Basically fourteen body parts are segmented given their joints (head, chest & abdomen as only one body part, right and left arms, right and left forearms, right and left hands, right and left thighs, right and left calf, and finally, right and left foot). It is important to mention that this method is mainly focused on dominant colors, consequently textured parts of the body may not be segmented correctly.

We initially define a region Tr_i around the corresponding bone that will be used to learn the dominant color(s) of each body part. The selected region is a rectangle, which central axis coincides with the corresponding bone, and that should ideally contain only pixels related to the corresponding body part. In the proposed approach, the length of the rectangle is exactly the length of the bone as clicked by the user, and the width is a fraction s_1 (set experimentally to 0.4) of the expected width of the corresponding body part w_i, given in Table 1. It is important to note that the 3D length l_i of each body part is used instead of the 2D distance (that could be computed directly from the corresponding joints in image coordinates). In fact, assuming that each body part is approximately cylindrical, the width of the 2D projection is not affected significantly by perspective issues. On the other hand, the 2D-3D correspondence of the bone length may change significantly depending on the pose of the body part (e.g. arms parallel to the ground).

To obtain the dominant color(s) of each body part,the unsupervised color-based segmentation algorithm [Jung 2007] is initially applied to obtain the main regions within Tr_i, as illustrated in Figure 4. In most cases, the largest of these regions is related to the dominant color. However, there are some common situations (shirts with writings, illustrations, shadows, etc.) in which the largest segmented region does not correspond to the dominant color. To cope with this issue, the N largest segmented regions within Tr_i, with area larger than a threshold T_a are retrieved (we experimentally set $N = 3$ and $T_a = 0.1\#Tr_i$, where $\#Tr_i$ is the area of Tr_i).

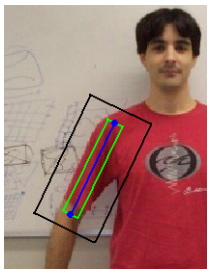

Figure 3: *Illustration of the region used to learn and search the dominant colors. The blue line is the informed bone; the green rectangle is the estimated region for learning; and the black rectangle is the estimated region for searching.*

Figure 4: *Illustration of the initial segmentation adopted in the color model learning stage. From left to right: (i) input image, (ii) initial segmentation, (iii) segments, and (iv) retrieved segments.*

For a given body part i, let us consider the $N_i \leq N$ largest segmented regions that satisfy the minimum area criterion. The color distribution within each region is represented as a multivariate Gaussian model, which requires the computation of the mean vector ($\boldsymbol{\mu}_{ij}$) and covariance matrix ($\boldsymbol{C}_{ij}$), where $1 \leq j \leq N_i$ relates to a different color model for the body part.

3.2.2 Finding the Silhouette

To find the pixels that are related to body part i, search region Te_i (also rectangular) is defined. Unlike the training search region Tr_i, this test region should be large enough to comprise all pixels related to the body part. The length of the search region is the length of Tr_i increased by a multiplicative factor (set experimentally to 1.15), and the width of Te_i is based on the estimated value of the corresponding (body part given in Table 1) increased by another multiplicative factor (set experimentally to 2). This means that body parts as wide as twice the average antrhopometrical width may be detected using the proposed approach.

Figure 3 illustrates a clicked bone (the blue line), the training region Tr_i used to learn the dominant color models (green rectangle), and the test region Te_i used to find pixels coherent with the N_i learned models. To compute such coherence within Te_i, the squared Mahalanobis distance $D_{ij}(\boldsymbol{c})$ for each pixel with color $\boldsymbol{c}$ and retrieved segment j is obtained through

$$D_{ij}^2(\boldsymbol{c}) = (\boldsymbol{c} - \boldsymbol{\mu}_{ij})^T \boldsymbol{C}_{ij}^{-1} (\boldsymbol{c} - \boldsymbol{\mu}_{ij}), \;\; 1 \leq j \leq N_i, \qquad (2)$$

and a threshold T_{ij} is computed for each dominant color j automatically, based on the peaks and valleys of the histogram of $D_{ij}^2(\boldsymbol{c})$ (see [Jacques Jr. et al. 2010] for more details). Then, a given pixel with color $\boldsymbol{c}$ is aggregated to body part i if for at least one dominant color j the relationship $D_{ij}^2(\boldsymbol{c}) \leq T_{ij}$ is satisfied.

The color-based approach described so far provides an initial estimate of each body part. However, noise, varying illumination, texture, and non-uniform regions may generate spurious responses and/or holes in the segmented regions. Morphological operators are then used to remove residual noise and fill small holes. More precisely, a sequence of an opening and a closing operator with a 3×3 cross-shaped structuring element is applied. The opening removes isolated responses, but may separate regions that are connected by narrow bridges. The subsequent closing operator intends to connect

disjoint regions that are sufficiently close to each other (including those separated by the initial opening). Then, a hold filling operator is used to complete possible holes in the interior of the binary images (particularly in the chest regions, due to possible text and/or images in the shirt).

Figure 5 illustrates the final segmentation procedure, were each body part is shown with a different color. Interceptions of different body parts are also shown in a different color (i.e., the thigh and calf are painted using two different colors, as they interception - probably the knee). The union of all detected body parts compose a binary silhouette of the person, as shown in Figure 6-a). However, since the detected individual body parts overlap, a separate procedure is used to estimate the width of body part, which is required to reconstruct the VH.

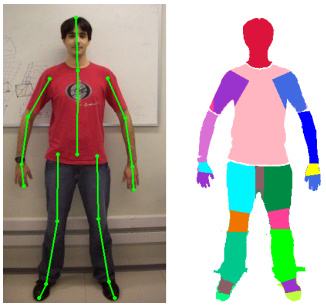

Figure 5: *Illustration of the final segmentation result.*

3.2.3 Silhouette Processing

Given the binary silhouette of the person and the 2D clicked joints, we estimate the width of each body part. The main idea is to compute the length of line segments connecting a bone and the boundary of the silhouette, and then combining these measurements robustly to estimate the corresponding width.

For each body part i, the central part of the bone (clicked by the user) is retrieved, as shown in Figure 6-c. Along this portion of the bone, line segments are traced perpendicularly to its corresponding bone to both sides, until a silhouette contour point is reached. Ideally, the lengths ls_{ik} of those segments should be close to $w_i/2$ (half the standard width of the body part, computed through Equation (1)). However, due to segmentation errors, some of these segments may be significantly smaller or larger than w_i. To cope with this problem, a range of valid possible lengths is created, and modified lengths ls'_{ik} are computed through:

$$ls'_{ik} = \min\{\max\{L_{low}, ls_{ik}\}, L_{high}\}, \quad (3)$$

where $L_{low} = 0.5w_i$ and $L_{high} = 2w_i$ define the limits of valid lengths, so that body parts between half and twice the average anthropometrical value can be detected.

Given the set of modified length values ls'_{ik}, the the estimated width of body segment i is given by

$$ew_i = 2 \underset{k}{\text{median}}\{ls'_{ik}\}, \quad (4)$$

so that outlier estimates are removed by the median operator.

3.3 3D Pose Identification

The problem of estimating the 3D pose of a person from image data has received a special attention in the computer vision literature. This is, in part, due to the fact that solutions to this problem could be employed in a wide range of applications. According to Taylor [Taylor 2000], most of the research in this area has focused on the problem of tracking a human actor through an image sequence, and less attention has been directed to the problem of determining an individuals posture based on a single picture. Indeed, this problem is challenging because 2D image constraints are often not sufficient to determine 3D poses of an articulated object. Our solution to this problem is based on Taylor's work [Taylor 2000], which

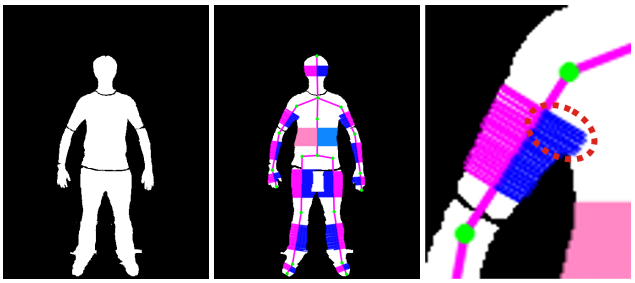

Figure 6: *(a) Silhouette obtained from the segmentation process using Figure 5. (b) Found widths for each interested body part. (c) Zoom given at right arm. The dotted area shows an place were a left edge was not found, so the measurement of the right side of the bone was replicated to the left side.*

presents a method for recovering information about the configuration of articulated objects from a single image.

According to Taylor, if we have a line segment of known length l in the image under scaled orthographic projection, the two 3D end points (x_1, y_1, z_1) and (x_2, y_2, z_2) are projected to (u_1, v_1) and (u_2, v_2), respectively. If the scale factor s of the projection model is known, it would be a simple matter to compute the relative depth of the two endpoints, denoted by $\Delta z = z_1 - z_2$, using the following equation [Taylor 2000]:

$$\Delta z^2 = l^2 - \frac{(u_1 - u_2)^2 + (v_1 - v_2)^2}{s^2}. \quad (5)$$

Such formulation generates ambiguities for each segment, since the sign of Δz can not be determined (i.e., we may have $z_1 > z_2$ or $z_2 > z_1$). If the skeleton has twenty joints, there are 2^{20} possible postures in the worst scenario. Despite the generation of ambiguities, this approach is very simple to implement, requiring only a straightforward sequence of computations. To minimize the problem of ambiguity, we firstly consider z1¿z2, and then the user can change it by using a tool we implemented, where the user can easily improve the generated pose, if necessary. Figure 7 shows an example of the manual skeleton initialization and three views of one 3D pose obtained with the model.

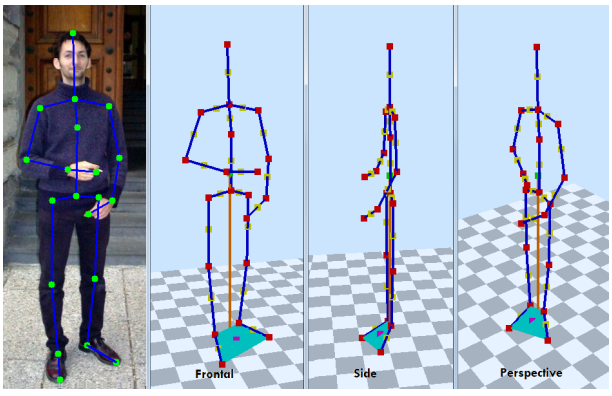

Figure 7: *Illustration shows three different views of a 3D pose obtained with our model.*

3.4 Virtual Human Reconstruction

This section describes our approach proposed to reconstruct the VH based on two templates (male and female), but any other human-like templates are possible (e.g. children). Given the gender of the person (manually informed), the corresponding template is deformed to match the silhouette processing information saved in the XML file. There are three steps required to transform the initial template into the final VH, briefly described next.

The first step is to adequate the posture of the template to match the posture estimated from the image. The XML file generated by the posture detector contains a set of labeled joints, as well as the corresponding 3D positions. From this set of points, it is easy to obtain the orientation of each bone, or the rotation angle at each joint. The rotation angles at each joint are then used to modify the posture of a generic VH, which assumes the posture of the person being analyzed in the photograph.

Once we put the VH in the same pose of the character in the picture, it is necessary to match dimensions of each individual body part of the template (e.g. arms, thorax, thighs, forearms and etc) to the dimensions (namely, length and width) computed from the image. Since each body part of the generic VH (template) presents predefined length tl_i and width tw_i, a simple linear scaling applied to both dimensions (length and width) can be used to obtain the desired dimensions of each body part in the final VH model. It is important to emphasize that the simple linear scale has been chosen in order to reconstruct the VH in a easy and fast way, making it applicable to games and mobile applications.

Finally, the geometric 3D model of the VH must be filled with color and texture. In our work, small pieces of textures are generated automatically (in regions defined as bones when clicked by the user) during the segmentation process, and in this phase they are used to provide textured body shapes. Since this process is automatic, we avoid using textures of faces and hands, which could include problems (e.g. when face picture is not frontal). It is also important to note that post-processed textures by artists (mainly for faces and hands) are recommended in order to improve the quality of mapped textures, but the scope of this paper is mostly focused on posture and geometry.

4 Results

Our pipeline is organized in two prototypes. Firstly, the prototype responsible for manual 2D clicks, segmentation, pose estimation and silhouette processing, which generates an XML file containing all information required to model the virtual human. For VH rendering and animation, our prototype uses Irrlicht Engine (`http://irrlicht.sourceforge.net/`) and Cal3D (`http://gna.org/projects/cal3d/`), respectively. Results discussed in this section were obtained using an Intel Xeon E405 equipped with a NVidia Quadro FX 4800 graphic card. The creation of the two generic templates (male and female, containing 4825 and 4872 vertices, respectively) were made with Autodesk 3D Studio MAX 9 (`http://usa.autodesk.com`).

All examples presented in this section have been automatically processed (except the manual intervention to inform the joints, the height of the person and the gender), as described previously in the paper. Occasionally, another manual intervention can be to improve the generated 3D posture. Also, when the face is completely frontal in the picture, the texture mapping is more likely to work, since the 3D model has always frontal orientation for the face. This is one of aspects that should be improved in future works. The images shown in Figure 8 present results obtained by using the proposed approach. The same VH model is shown in Figure 9 including the face texture. Also, it is possible to export (in CAL3D format) the VH generated in our model, and subsequently import it in another tool or animation Engine. In case of Figure 9, we animated the VH in the Irrlicht engine, using a pre-defined animation file.

Figures 10 and 11 illustrate the whole pipeline proposed in this work. From left to right, top to bottom: the estimated posture, bone boxes used in the segmentation, segmented image and silhouette processing, as well as two points of view of virtual human generation are illustrated. In this case, it is possible to observe problems that happen in segmentation (see third image in the top of Figure 10). The segmentation result impacts in errors in the computed width of the legs, which can be solved based on the simple rules specified in Section 3.4.

Figures 12 and 13 show visually adequate results, but it is possible to observe that the hands are not touching the body or in the pockets, as observed in the pictures. They illustrate the result of ambiguities in the pose estimation.

(a) Generated virtual human. (b) Another point of view from the same scene.

Figure 8: *Results from our model illustrating the VH and the picture used as input.*

Figure 9: *Results from VH model imported and animated using Irrlicht engine.*

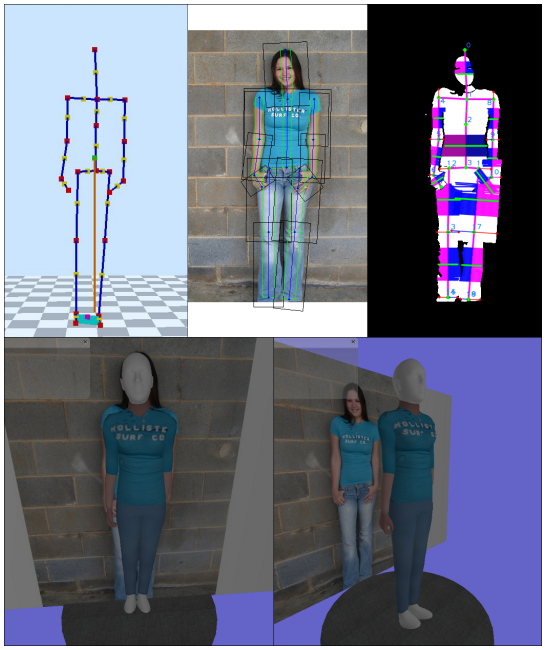

Figure 10: *From the left to the right and top to the bottom: the estimated posture, bone boxes used in the segmentation, segmented image and silhouette processing, as well as two points of view of virtual human generation are illustrated.*

The reconstruction of the virtual human is not performed in real time, and it is impacted negatively as the number of vertices in the models increases. In average, the whole process to generate a new virtual human based on the templates described in this paper takes around 3 minutes. Other results can be found in accompanying video file.

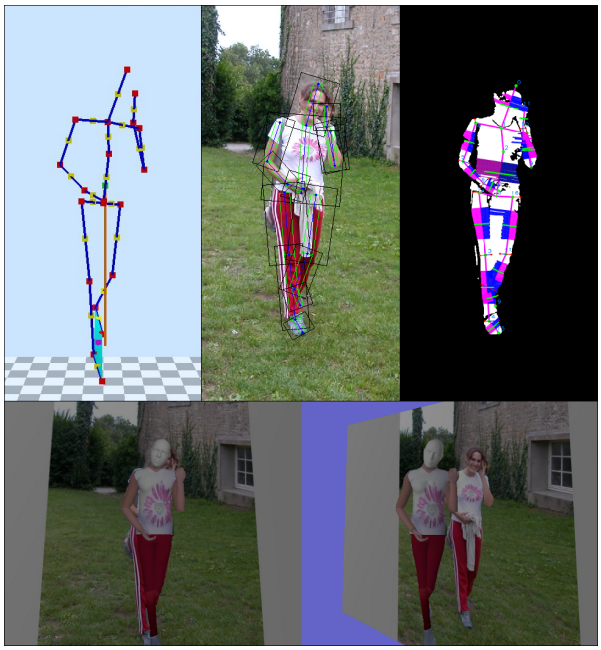

Figure 11: *Another Result of our model, illustrating pipeline steps.*

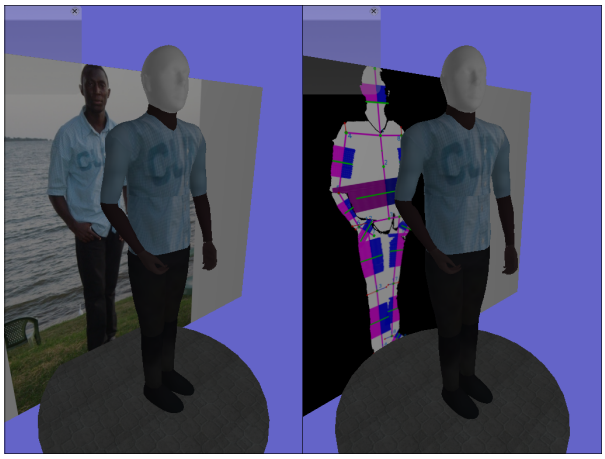

Figure 12: *Left: the original image and the generated VH. Right: the processed silhouette and the VH.*

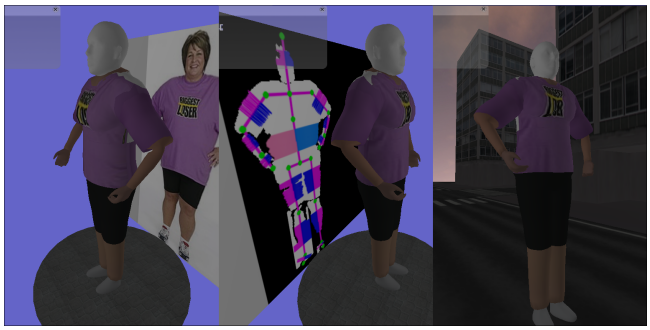

Figure 13: *From left to right: the original image and the generated VH, another point of view of 3D model and processed silhouette, and the VH inserted in a virtual world.*

5 Final Considerations

This paper presented a model to reconstruct 3D VHs based on a single and spontaneous image. Although the reconstruction may not be exactly accurate, the whole creation process is fast, and it provides good visual results maintaining its coherence with the original picture. The pipeline needed to generate VH is basically organized in four parts: human segmentation and silhouette processing, pose estimation, all of them in image domain; the last phase is responsible by the VH generation and is performed in the graphic domain.

The main contribution of this work is the possibility of fast and coherent VH creation without requiring training databases, but deforming templates. Moreover, our model has some advantages since it does not rely on any background subtraction or lighting setting. In the results of this paper, we used only 2 templates - 1 male and 1 female. However, problems can arise mainly due to the segmentation process, which can generate erroneous silhouettes that impact the VH deformation. Depending on the characteristics and occlusions existent in the picture, the segmentation process can generate inconsistent silhouettes that are passed on to the rest of the pipeline. Concerning pose estimation, the positions of clicks are very relevant in order to compute the final postures, but our graphical interface can deal with these problems. The linear scale is certainly a current limitation, however it is a good solution for purposes of games and real-time applications in mobiles, for instance. It is important to emphasize that it is very hard to perform a quantitative evaluation of the obtained 3D models, and even more if we want to compare with other methods presented in literature.

Future work should include the perspective estimation in the picture and also propose a way to manage with clothes as other few approaches in the literature. Another future work could be to improve the segmentation method by using textures and not only dominant colors, as well as to provide not only linear scale for template deformation. In the latter case a morphable model can be included based on many templates or still in SCAPE database, as other related work [Zhou et al. 2010].

References

ANGUELOV, D., SRINIVASAN, P., KOLLER, D., THRUN, S., RODGERS, J., AND DAVIS, J. 2005. Scape: shape completion and animation of people. *ACM Trans. Graph 24*, 408–416.

BOYKOV, Y. Y., AND JOLLY, M. P. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on 1*, 105–112 vol.1.

DALAL, N., AND TRIGGS, B. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, 886 – 893.

FUA, P. 1999. Human modeling from video sequence. *Geomatics Info Magazine 13*, 7 (July), 63–65.

GUAN, P., WEISS, A., BALAN, A., AND BLACK, M. 2010. Estimating human shape and pose from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, IEEE, 1381–1388.

HASENFRATZ, J.-M., LAPIERRE, M., GASCUEL, J.-D., AND BOYER, E. 2003. Real-time capture, reconstruction and insertion into virtual world of human actors. In *Vision, Video and Graphics*, Elsevier, Eurographics, 49–56.

HILTON, A., BERESFORD, D., GENTILS, T., SMITH, R., AND SUN, W. 1999. Virtual people: Capturing human models to populate virtual worlds. *Computer Animation 0*, 174.

HORNUNG, A., DEKKERS, E., AND KOBBELT, L. 2007. Character animation from 2d pictures and 3d motion data. In *ACM Transactions on Graphics*, vol. 26, 1 – 9.

JACQUES JR., J. C. S., DIHL, L., JUNG, C. R., THIELO, M. R., KESHET, R., AND MUSSE, S. R. 2010. Human upper body identification from images. In *International Conference on Image Processing, 2010. ICIP'05. IEEE Computer Society Conference on*, 1 – 4.

JUNG, C. R. 2007. Unsupervised multiscale segmentation of color images. *Pattern Recognition Letters 28*, 4 (March), 523–533.

LEE, W., J.GU, AND MAGNENAT-THALMANN, N. 2000. Generating animatable 3d virtual humans from photographs. *Computer Graphics Forum 19*, 3 (August), 1–10.

MAGNENAT-THALMANN, N., AND THALMANN, D. 2008. Real-time individualized virtual humans. In *SIGGRAPH Asia '08: ACM SIGGRAPH ASIA 2008 courses*, ACM, New York, NY, USA, 1–64.

ROTHER, C., KOLMOGOROV, V., AND BLAKE, A. 2004. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. 23* (August), 309–314.

TAYLOR, C. J. 2000. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Comput. Vis. Image Underst. 80*, 3, 349–363.

THALMANN N., S. G. 1997. A user-friendly texture-fitting methodology for virtual humans.

TILLEY, A. R. 2002. *The measure of man and woman - Human factors in design*. John Wiley & Sons, inc.

ZHOU, S., FU, H., LIU, L., COHEN-OR, D., AND HAN, X. 2010. Parametric reshaping of human bodies in images. In *SIGGRAPH '10: ACM SIGGRAPH 2010 papers*, ACM, New York, NY, USA, 1–10.