

3D Reconstruction of Human Skeleton from Single Images or Monocular Video Sequences

Fabio Remondino, Andreas Roditakis

Institute for Geodesy and Photogrammetry - ETH Zurich, Switzerland
E-mail: <fabio>, <roditak>@geod.baug.ethz.ch

Abstract. In this paper, we first review the approaches to recover 3D shape and related movements of a human and then we present an easy and reliable approach to recover a 3D model using just one image or monocular video sequence. A simplification of the perspective camera model is required, due to the absence of stereo view. The human figure is reconstructed in a skeleton form and to improve the visual quality, a pre-defined human model is also fitted to the recovered 3D data.

Introduction

In the last years the generation of 3D models of man made objects has become a topic of interest for several researchers. Particular attention has also been paid on the reconstruction of realistic human models, which could be employed in a wide range of applications such as movies, medicine, surveillance, video games, virtual reality environments or ergonomics applications. A complete human model usually consists of the shape and the movement of the body. Some available systems consider the two modeling processes as separate even if they are very close. Considering the techniques that recover the *shape of static humans*, nowadays a classical approach commonly used relies on *3D scanners* [5, 7, 27]: these sensors (Figure 1, A) are quite expensive but simple to use and various software is available to model the 3D measurements. They work according to different technologies providing for millions of points, often with related color information. Other techniques try to recover the shape of human figures with *image-based* approaches. They can use single camera stereo-view geometry (Figure 1, B) [20], silhouette extraction [11] or single image measurements [3, 16, 23]. *Computer animation software* [1, 18, 21] can instead produce realistic 3D human model subdividing and smoothing polygonal elements, without any measurements (Figure 1, C). These spline-based systems are mainly used for movies or games and the created virtual human is animated using similar animation packages or with motion capture data. Concerning the *motion of the human*, the main problem is the great number of degrees of freedom to be recovered. Existing and reliable commercial systems for capturing human motion typically involve the tracking of human's movements using *sensor-based hardware* [2, 19]. Other approaches instead rely on *2D monocular videos* of human as primary input [12, 22]. They use computer vision techniques, image cues, background segmentation,

blob statistics, prior knowledge about human motion, probabilistic approaches and pre-defined articulated body models to recover motions and 3D information. Finally *multi-cameras approaches* [8, 10, 25] are employed to increase reliability, accuracy and avoid problems with self-occlusions.

Many research activities in this area has focused on the problem of tracking a moving human (human motion analysis) through an image sequence acquired with single/multiple camera(s) and often using pre-defined 3D models. But little attention has been directed to the determination of 3D information of a human directly from a single image or monocular sequence using a camera model (deterministic approach). In this contribution we present a simple and efficient method to find the poses and the 3D model of a human imaged in a single image or in a monocular sequence. Our work is similar to [23], but additional changes and improvements are presented and discussed. The 3D human model is recovered with a deterministic approach and only for visualization purposes a laser scanner 3D model is fitted to the recovered data.

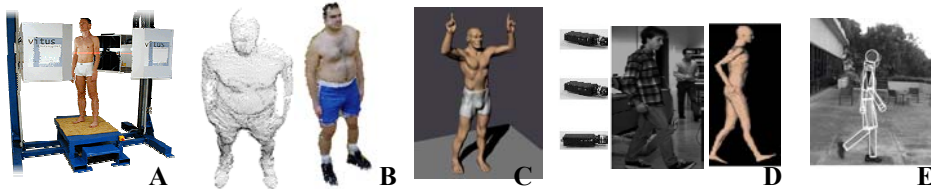


Figure 1: Different approaches to recover a human body model. A: laser scanner system [27].

B: single-camera stereo-view image sequence [20]. C: computer animation software [21].

D: Multi-camera system [8]. E: Probabilistic approach on monocular sequence [22].

The reconstruction algorithm

Usually the algorithms that want to recover accurate 3D models from images are based on the collinearity equations [20]. They state that a point in object space, its corresponding point in an image and the projective center of the camera lie on a straight line. If a point is stereo-imaged in more than one frame, its 3D coordinates in the world coordinate system can be recovered e.g. with the bundle method, as well as the camera parameters. Although this method is very accurate, it requires a point to be imaged in at least two images and a good baseline between consecutive frames; therefore it is not possible to use it when a (rotating) monocular sequence or a single image is used. A simplification of collinearity equations leads to the perspective projection that relates the image measurements to the world coordinate system just through the camera constant c :

$$\begin{aligned} x &= -c \cdot \frac{X}{Z} \\ y &= -c \cdot \frac{Y}{Z} \end{aligned} \quad (1)$$

If we want to recover 3D information from a single uncalibrated view, we have the so called ‘ill-posed’ problem: for each point, we have two equations and three unknown coordinates, plus the camera constant. Therefore the system is underdetermined and some more assumptions need to be introduced. For man made

objects (e.g. buildings), geometric constraints on the object (perpendicularity and orthogonality) and image invariant can be used to solve an ill-posed problem [24]. But in case of free form objects (e.g. the human body) these assumptions are not valid. Therefore equation (1) can be furthermore simplified, describing the relationship between the 3D object coordinates and 2D image measurements with an orthographic projection scaled with a factor $s = -c/Z$:

$$x = s \cdot X \quad (2)$$

$$y = s \cdot Y$$

The effect of orthographic projection is a simple scaling of the object coordinates. The scaled-orthographic model amounts to parallel projection, with a scaling added to mimic the effect that the image of an object shrinks with the distance. This camera model can be used if we assume the Z coordinate almost constant in the image or when the range of Z values of the object (object's depth) is small compared to the distance between the camera and the object. In those cases the scale factor c/Z will remain almost constant and it is possible to find a value of s that best fit in equation (2) for all points involved. Moreover it is not necessary to recover the absolute depth of the points with respect to the object coordinate system. Furthermore the camera constant is not required and this makes the algorithm suitable for all applications that deal with uncalibrated images or video. But, as it is generally an ill-posed problem, we still have an undetermined system, as the scale factor s cannot be determined only by means of equation (2) and a single frame. Therefore, supposing that the length L of a straight segment between two object points is known, it can be expressed as $L_{12}^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2$. By combining this equations with (2) we end up with an expression for the relative depth between two points:

$$(Z_1 - Z_2)^2 = L_{12}^2 - [(x_1 - x_2)^2 + (y_1 - y_2)^2] / s^2 \quad (3)$$

So, if the scale parameter s is known, we can compute the relative depth between two points as a function of their distance L and image coordinates. Therefore the whole reconstruction problem can be reduced to the problem of finding the best scale factor for a particular configuration of image points. Equation (3) also shows that, for a given scale parameter s , there are two possible solutions for the relative depth of the endpoints of each segment (because of the square root). This is caused by the fact that even if we select point 1 or point 2 to have the smaller Z coordinate, their (orthographic) projection on the image plane will have exactly the same coordinate. In order to have a real solution, we have to impose that:

$$s \geq \frac{\sqrt{[(x_1 - x_2)^2 + (y_1 - y_2)^2]}}{L_{12}} \quad (4)$$

By applying inequality (4) to each segment with known length one can find the scale parameter that can be used in equation (3) to calculate the relative depth between any two segments endpoints. Because of the orthographic projection assumed, we have to decide an arbitrary depth for the first point and then compute the second point depth relative to the first one. For the next point we use a previous calculated depth and equation (3) to compute its Z coordinate and so on in a segment-by-segment way. Due to the difference in the left side of equation (3), we have also to decide, for each segment, which one is closer to the camera. Then, knowing the scale factor, equation (2) can be used to calculate the X and Y coordinates of the image points. In [23] is mentioned that images with significant perspective effect could not

be modeled with this approach. In fact, in some results, because of measurement or assumption errors, a segment that seems to be almost parallel to the image plane can get foreshortened or warped along one axis. But if a segment is almost parallel to the image plane or can be assumed to lie on a plane, then the two points can be treated as being at the same depth. And imposing additional constraints, such as requiring that 2 points must have the same depth, this mistake can be avoided and the resulting 3D model is more accurate and reliable. An example is presented in Figure 2-A where the 3D skeleton recovered with simple orthographic projection is improved using some depth constraints. Other constraints could be the perpendicularity of two segments or a closure constraint, imposing that the two points must coincide (Figure 2-C).

The human body model and its representation

The human skeleton system is treated as a series of jointed links (segments), which can be modeled as rigid bodies. For the specific problem of recovering the pose of a human figure, we describe the body as a stick model consisting of a set of thirteen joints (plus the head) connected by thirteen segments (we consider the shoulder girdle as unique segment), as shown in Table 2. The head joint is used to model some figure where it is inclined, as shown in, Figure 2-C and D. The algorithm needs the knowledge of the relative lengths of the segments, which can be obtained from anthropometric data (motion capture databases or literature). Two sets of length values are used in our tests, leading almost to the same 3D models. The first set of relative distances between human joints is derived from a motion capture database [4] (Table 2, central). The second set is more general and follows the studies performed by Leonardo Da Vinci and Michelangelo on the human figure [13, 26] (Table 2, right). It represents the human figure as an average of eight heads high. Once the program has computed the 3D coordinates of the human joints, they are given to a procedure that uses VRML language to visualize the recovered model. All the joints are represented with spheres and they are joined together with cylinders or tapered ellipsoids (to model the shape of the muscles).

| Segment | Relative Length (MC) [cm] | Relative Length (L) [unit] |
|-----------------|---------------------------|----------------------------|
| Height | 175 | 8 i |
| Lower arm | 35 | 2 i |
| Upper Arm | 25 | 1 ½ i |
| Neck-Head | 25 | 1 ¼ i |
| Shoulder Girdle | 44 | 2 i |
| Torso | 53 | 2 ½ i |
| Pelvic Girdle | 30 | 1 ½ i |
| Upper leg | 46 | 2 i |
| Lower leg | 52 | 2 i |
| Foot | 22 | 1 i |

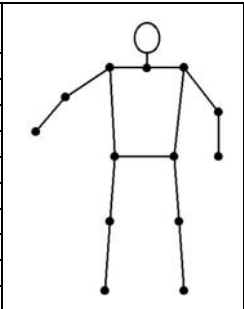


Table 2: Two different sets of relative lengths of the segments used in the computation of the human skeleton. MC = Motion Capture. L = Literature. The coefficient i has been added in the second one to consider the variation of the human size from the average size ($i=1$).
On the right the human model skeleton used in the reconstruction.

Implicit/Explicit surface fitting

Fitting a surface on a given set of points is a problem that has been approached in several ways in Computer Graphics literature. The main classification considers the type of the final representation, dividing the fitting methods into explicit and implicit. *Triangular meshes*, *volume grids* [6] and *parametric piecewise functions (NURBS)* [15] are explicit descriptions of the surface, while *soft* or *blobby objects*, also known as *metaballs* [8, 14] describe the surface as the isosurface of a distance function. On one hand the explicit functions appear to be a popular representation in modeling software and are hardware supported, achieving rendering performances of millions of texture mapped polygons per second. Fitting such surfaces on a set of measurements presents, though, the difficulty of finding the faces that are closest to a 3D point, and the disadvantage of non-differentiability of the distance function [14]. Implicit surfaces are more suitable for modeling soft objects as they have been used in modeling clouds [9] or soft tissue objects [14], but present difficulties in deformations and rendering.

In our application, to improve the visual quality of the results, we chose to fit the recovered 3D skeleton with a polygonal mesh [6], using the modeling and animation software Maya [17]. Our skeleton has a hierarchical structure, which is used to calculate the number of joints that influence every point (called influence depth). Large influence depths result in ‘softer’ objects. We kept a low depth to avoid too soft deformations and after the automatic fitting we adjusted manually the influence weight of some skeleton joints on the polygons to eliminate hard edges. For the movement of the skeleton there are two solutions available in Maya, called *Forward* and *Inverse Kinematics*. The first method requires the rotation and translation of all the joints, starting from the parent and ending to the last child joint, to achieve the final pose. The latter method requires that only the position and rotation of the desired pose, or *target locator*, is given from the user and then the position of the intermediate joints is calculated automatically. In this case, the use of joint rotation constrains is essential in order to achieve a correct solution. In the present paper, we use inverse kinematics, because of the simplicity and automation of the procedure. In Figures 2-E and 3 we present the results of fitting a polygonal model acquired with a Cyberware body scanner [6] to the poses recovered from single image and monocular video sequence.

Results on single uncalibrated images

In order to determine the limitations, advantages and accuracy of the method, a series of experiments were performed at the beginning on single images taken from the Internet or extracted from videos. Figure 2 shows some images and the associated 3D models looked from different viewpoints. In particular, column A shows a 3D model obtained with the simple orthographic projection (central image: 3D model warped and distorted) and after the applied constraints (lower result). Furthermore, measurements occlusions are handled selecting the most adequate point in the image and computing its 3D coordinate using a depth constraint (e.g. right knee in Figure 2-

A or right shoulder in Figure 2-B). The accuracy of the reconstruction is demonstrated by the fitting results (Figure 2-E): in fact the laser model (that is precise for definition) is just scaled and its segments are rotated to match our skeleton.

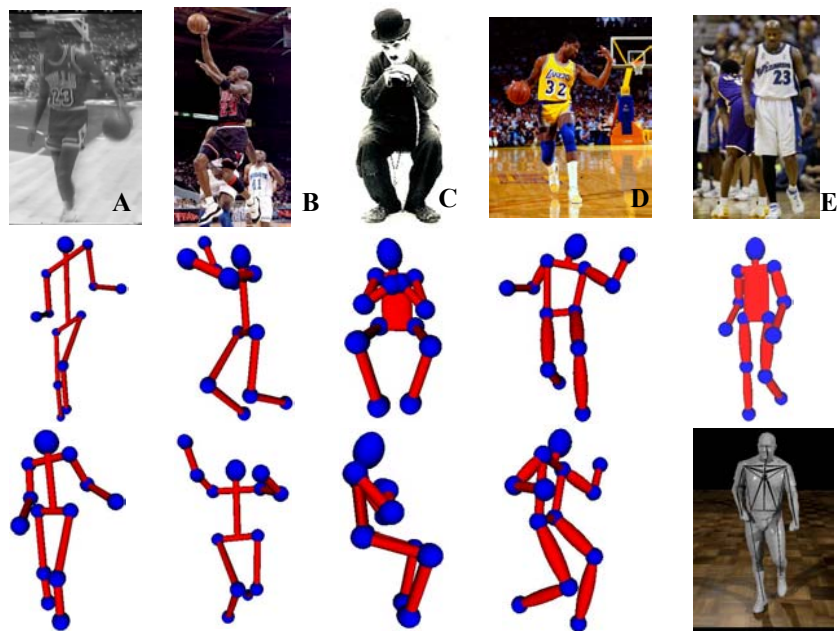


Figure 2: 3D model of human figure recovered from single images. Different representations of the human skeleton are presented: simple cylinders, scaled cylindrical torso and tapered ellipsoids to model the muscles. Note also the head of the figure (C and D), that is represented in the recovered model inclined according to the image data. Column E shows a 3D laser model fitted to our skeleton.

Application to monocular video sequences

The reconstruction algorithm has been also extended to solve the reconstruction problem in consecutive frames obtained from videos. The poses and movements of a figure over a short interval of time are recovered and then, by playing the reconstructed poses at an adequate frame rate, we can get a very realistic and observer independent reconstruction of the original motion. However, because the algorithm finds only relative distance in each frame, subsequent poses normally have no object coordinates in common. This is solved assuming that a joint in one frame has the same coordinates (plus a small translation vector) of the corresponding joint in the previous frame. In the example presented in Figure 3, we digitized 20 frames from an old videotape. The image points were measured semi-automatically with a Least Squares Matching Algorithm. The recovered models show the reliability of the algorithm and its possible application also in case of small perspective effect. In case

of good image quality, the corresponding joints could be tracked automatically over the video sequence, as described in [8].

Conclusion

In this work we presented the problem of recovering 3D models of humans from single images and monocular video sequences. The problem was solved with a scaled orthographic projection and some additional constraints. The reconstruction and visualization algorithm was tested with several images and sequences and the presented results show the reliability of our extended algorithm, also when some perspective effects are present. As future work we want to add the foot and the hands to the recovered skeleton and we will try to model muscles and shape of some segments using tapered cones. Moreover a perspective camera model will also be tested on human figures imaged in monocular sequences.

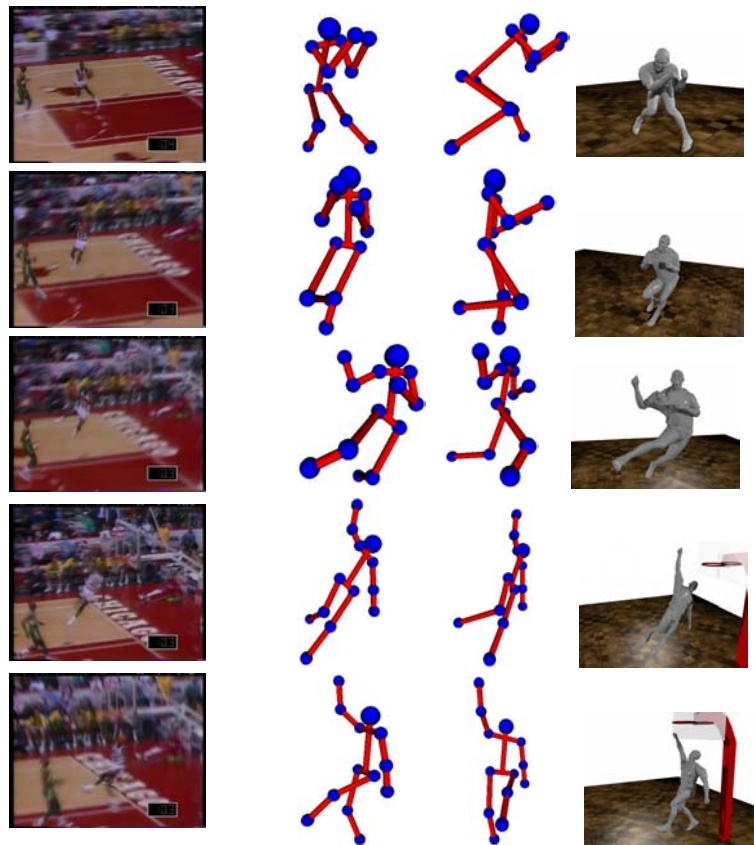


Figure 3: Reconstructed 3D model of the moving human in the monocular video sequence. The second and third columns show the recovered skeleton viewed from two different viewpoints. The last column shows the fitted laser model to our 3D data.

References

1. 3D Studio Max: <http://www.3dmax.com> [June 2003]
2. Ascension: <http://www.ascension-tech.com/> [June 2003]
3. Barron, C., Kakadiaris, A.: Estimating Anthropometry and Pose from a single uncalibrated image. *Computer Vision and Image Understanding*, Vol. 81, 269-284, 2001.
4. Biovision: <http://www.biovision.com> [June 2003]
5. BodySkanner: <http://www.scansuccess.com> [June 2003]
6. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. 23rd Conf. on Computer graphics and interactive techniques, pp. 302-312, 1996.
7. Cyberware: <http://www.cyberware.com> [June 2003]
8. D'Apuzzo, N., Plankers, R., Fua, P., Gruen, A., Thalmann, D.: Modeling human bodies from video sequences. *Videometrics Conference, SPIE Proc.*, Vol. 3461 (1999), 36-47
9. Dobashi, Y., Kaneda, K., et. al.: A simple, efficient method for realistic animation of clouds. *Proc. of 27th Conf. Computer graphics and interactive techniques*, pp. 19-28, 2000.
10. Gavrilu, D.M, Davis, L.S.: 3-D Model-based Tracking of Humans in Action: a Multi-View Approach. *CVPR Proceedings*, 1996.
11. Hilton, A., Beresfors, D., Gentils, T., Smith, R., Sun, W., Illingworth, J.: Whole-body modeling of people from multiview images to populate virtual worlds. *The Visual Computer*, Vol. 16, 411-436, Springer-Verlag, 2001
12. Howe, N., Leventon, M., Freeman, W.: Bayesian reconstruction of 3D human motion from single-camera video. *Advances in Neural Information Processing System*, Vol. 12, 820-826, MIT Press, 2000.
13. Human Figure Drawing Proportion: www.mauigateway.com/~donjusko/human.htm
14. Ilic, S., Fua, P.: From explicit to implicit surfaces for visualization, animation and modeling. *Proc. of Inter. Workshop on visualization and animation of reality based 3D models*, Vulpera, Switzerland, 2003.
15. Krishnamurthy, V., Levoy, M.: Fitting smooth surfaces to dense polygon meshes. *Proc. of 23rd Conf. on Computer graphics and interactive techniques*, pp. 313-324, 1996.
16. Lee, H.J., Chen, Z.: Determination of human body posture from a single view. *Computer Vision, Graphics, Image Process*, Vol. 30 (1985), 148-168.
17. Learning Maya 2.5: Alias Wavefront, 1998.
18. Lightwave: <http://www.lightwave3d.com> [June 2003]
19. Motion Analysis: <http://www.motionanalysis.com/> [June 2003]
20. Remondino, F.: 3D reconstruction of static human body with a digital camera. *Videometrics Conference, SPIE Proc.*, Vol. 5013, pp. 38-45, 2003.
21. SculpLand: <http://www.sanynet.ne.jp/~nakajima/SculpLand.html> [June 2003]
22. Sidenbladh, H., Black, M., Fleet, D.: Stochastic Tracking of 3D Human Figures Using 2D Image Motion. *ECCV*, D. Vernon (Ed.), Springer Verlag, LNCS 1843, pp. 702-718, 2000.
23. Taylor, C.T.: Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. *Computer Vision and Image Understanding*. Vol. 80, 349-363
24. Van den Heuvel, F.A.: 3D reconstruction from a single image using geometric constraints. *ISPRS Journal for Photogrammetry and Remote Sensing*, 53, No. 6, pp. 354-368, 1998.
25. Vedula, S., Baker, S.: Three Dimensional Scene Flow. *ICCV '99*, Vol. 2, pp. 722-729.
26. Visual Body Proportion: <http://www2.evansville.edu/drawinglab/body.html> [June 2003]
27. Vitus: <http://www.vitus.de> [June 2003]