

Inference of 3D Human Body Posture from Multiple Cameras for Vision-Based User Interface

Isaac Cohen, Gérard Medioni, Haisong Gu
Institute for Robotics and Intelligent Systems
University of Southern California
{icohen|medioni}@iris.usc.edu

ABSTRACT

In this paper we presents a silhouette-based method for 3D human body reconstruction from multiple views. The silhouettes are extracted automatically from the captured frames using a background-learning approach. The multiple views are integrated using the epipolar geometry relating the pair of synchronized video streams. The 3D model of the human body is derived from the integration of the median axis of the silhouettes along with a Generalized Cylinders description of the 3D shape. The proposed approach is validated by a set of experiments capturing various human postures.

1. INTRODUCTION

Human body motion tracking and analysis has received a significant amount of attention in the computer vision research community in the past decade. This has been motivated in part by the desire to replace expensive and cumbersome motion capture systems such as body suits, optical (marker-based) or magnetic tracking systems, and data gloves. These systems, while reasonably accurate and in some cases fast, are restrictive enough that they cannot be used as part of the widely and easily accessible immersive environments we envision.

Early systems to visually track people tended to be very slow and were typically tested on just a few video sequences [18, 19, 23, 7, 14, 6]. Some focused on the problem of recovering structure from motion; for example, Webb and Aggarwal [23] studied motion of articulated objects representing human body kinematics, while Rashid [19] and Goddard [6] used moving light displays as input. In contrast, Krueger [13] built an interactive system called VIDEOPLACE that enabled

users to interact with a synthetic environment via real-time analysis of body silhouettes.

In recent years, computer vision researchers have continued to investigate the recovery and analysis of full-body structure [11, 5] and to develop real-time interactive systems with more sophisticated 2D and 3D tracking and representation [22, 1, 20, 2, 10, 21, 3]. The Pfinder system [24] at the MIT Media Lab was perhaps the first to perform 2D person segmentation, tracking and interpretation while providing a real-time person model suitable to support a variety of interactive applications. More recent systems have tracked body posture and 3D locations of heads and hands in the context of controlling virtual characters [8] and interactive special effects [3].

Human gestures involve the simple movements of body parts, requiring an observation in a close-up range such as movements of finger joints and the rotation of arms or a head. Gestures may be communicated by facial expressions, by hand signs or by body limbs. Based on the repetitiveness of movements, these gestures may be further categorized as being periodic or not.

Understanding the human motion from a single visual stream is challenging since only the $2D$ projection of these motions is captured. Several body motion are therefore difficult to capture and understand and multiple views are required to disambiguate or identify the human motion. We propose a multi-sensor approach to capture the real $3D$ body motion from video streams. Several approaches have been proposed for estimating human postures in $3D$, these approaches rely on three to an array of cameras to capture the human motion [9, 12].

Our approach, is based on an incremental integration of the $2D$ views captured by a set of two or more GlobeAll camera systems [17]. The GlobeAll device

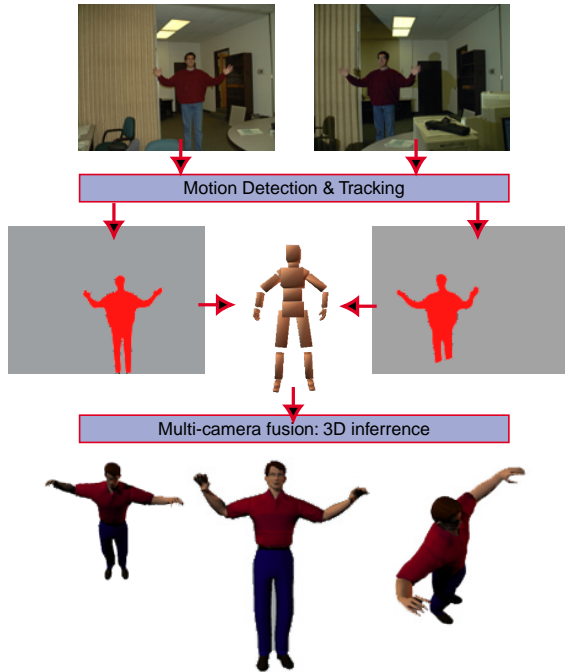


Figure 1: 3D inference of a moving person from multiple cameras.

is made of an array of cameras mounted on a fixed frame allowing to acquire a high resolution wide field of view images. The camera images are registered so that a large region of awareness can be maintained as a two-dimensional mosaic. Based on this mosaic, any arbitrary intermediate view (according to the desired region of the interest) can be created from the collection of images continuously acquired by the array of cameras. The system is functionally equivalent to a mobile camera platform, but it performs its pan-tilt-zoom operations electronically rather than mechanically.

Each camera system provides data to characterize the articulated body model and the associated key feature points. The integration, of different views of the articulated body model provides a 3D representation of the human body motion. This 3D representation, inferred from synchronized video streams and mapped onto a generic articulated body model provides an efficient approach for disambiguating human motion. Indeed, the combination of the 3D body model along with the 3D geometry of the sensors allow to merge the 2D tracks of key feature points, extracted from each video stream, into real 3D tracks. These 3D tracks provide more detailed and reliable description of human

body motion for analyzing and understanding the status of the user during his interaction with a machine driven system. The overview of the approach for inferring the 3D posture of the person is illustrated in figure 1.

2. BODY DETECTION

The objective of the incremental integration of the 2D views captured by the GlobeAll devices is the inference of the 3D body posture. The different video stream acquired by the cameras can be merged in order to reconstruct the 3D representation of the scene. However, since we are only interested in the characterization of the body posture and the human gesture, a detection of the person has to be performed in each video stream.

The video sequences are captured by a static camera which allows us to use a detection of objects in the scene based on a background modeling technique. A pixel-based background model of the scene can be estimated using a Gaussian distribution scheme. From a large set of images we fit a Gaussian distribution model to each pixel [24]. This distribution reflects the pixel changes over time and can be used to characterize the background or static part of the scene. There are different implementations of the algorithm: fixed or adaptive. In the fixed implementation, the background properties are inferred once for all and are not updated. This is commonly used in indoor situations where we can acquire a sufficient large number of frames of the static scene. In outdoor scenes, an adaptive implementation performs better as it allows to update the background model to adjust to change in illuminations.

The background learning technique is based on a Gaussian modeling of the pixel change over time. For each pixel P_i we estimate the Gaussian distribution from a collection of frames acquired by the camera.

Video streams acquired by stationary cameras provide a high temporal sampling of the scene. This large number of samples in time for each pixel can be efficiently used for inferring a segmentation of the frames onto foreground and background regions.

Foreground regions are associated to moving objects in the scene, while background represent the static regions in the viewed scene.

The segmentation of each frame onto foreground/background regions is derived from a statistical modeling of the temporal sequence of each pixel gray levels. For each pixel (i, j) , the temporal sequence:

$$S_{ij}(t_0, t_1) = \{I(i, j, t_0), I(i, j, t_0 + 1), \dots, I(i, j, t_1)\}$$



Figure 2: Body detection from a video stream. The two images correspond to two frames simultaneously acquired by two video cameras.

represents the collection of measured gray level, or the temporal distribution associated to the pixel (i, j)

The temporal sequence $S_{ij}(t_0, t_1)$ can be modeled using a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ defined by the probability:

$$P(g) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(g - \mu)^2 / 2\sigma^2)$$

The parameters μ and σ of the Gaussian distribution are estimated by:

$$\mu = \frac{1}{N} \sum_{i=1}^N g_i$$

and

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (g_i - \mu)^2$$

The foreground/background segmentation of a newly acquired frame is performed as follows: For each pixel,

- Update the mean and variance of the temporal sequence as follows:

$$\mu_{new} = (1 - \alpha)\mu + \alpha + \alpha I_{new}(i, j)$$

$$\sigma_{new}^2 = \max(\sigma_{min}^2, (1 - \alpha)\sigma^2 + \alpha(I_{new}(i, j) - \mu_{new})^2)$$

where α is a learning rate and σ_{min} is a threshold for minimal standard deviation.

- Pixels are flagged as foreground if $\sigma > \tau$ where τ is a given threshold.

In figure 2 we show an example of such detection. The figure shows two frames extracted from the two video streams and the corresponding detections.

Instead of modeling each temporal sequence with a Gaussian model, we can rely on the fact that a background pixel is a pixel for which the gray level distribution does not change. A robust method to extract these pixels is by extraction the *mod* of the temporal sequence distribution. The *mod* of the distribution is computed by identifying the gray value that maximize the histogram of the temporal sequence $S_{ij}(t_0, t_1)$.

3. MULTI-CAMERAS FUSION: INFERENCE OF 3D

Acquiring 3D models from images is a challenging task and many techniques have been developed over the last few years. In most of the cases these 3D models are inferred from a sequence of images acquired moving cameras. In our application, we are interested in reconstructing the 3D body model of a moving person from multiple synchronized video streams.

Three dimensional object reconstruction from a set of cameras is a classical problem in computer vision. In the recent years the topics have gained much attention due to some nice applications such as the *Virtualized Reality* [16]. Volumetric scene modeling was initially proposed by Martin [15] and have gained recently a lot of attention due to its generic usage. This method also known as *shape from silhouettes* allows to combine silhouette image with calibration information for each camera to characterize the 3D space where the object must lie. Many algorithms have been developed for constructing volumetric models from a set of silhouettes (see [4] and references therein). The main steps of these algorithms are the intersection of the silhouettes and the efficient encoding of the 3D space. These techniques rely on calibrated cameras to infer a 3D reconstruction of the silhouettes and are usually time consuming.

In our application we are interested in reconstructing the body model of a human. Therefore, additional knowledge can be used to simplify the shape-from-silhouette problem. Human body model can be represented using an articulated model based on generalized cylinders (GC). These GC correspond to the elementary shape description of the human body and conse-



Figure 3: Median axis of the silhouettes displayed in figure 2

quently locating these elementary GC allow to derive a human body model.

Relying on a GC-based model to describe the human body model simplify tremendously the problem of the integration of multiple views into a 3D model. Indeed, it was shown that the 3D reconstruction of a GC can be obtained from the 3D reconstruction of its axis. We propose to use this property to infer 3D model of a human body from its silhouettes. In the following we assume that the epipolar geometry is known. The epipolar geometry is inferred by using a modified version of the method proposed by Zhang *et al* [26].

3.1. SILHOUETTES' MEDIAN POINTS

The description of a GC is provided by its median axis and the set of radius function representing the radius of the GC cross sections. Therefore, identifying the median axis in each silhouette and reconstructing its 3D location using the relative position of the cameras provides a 3D representation of the GC. The extraction of the median axis in each silhouette can be performed using morphological tools. We chose to compute the median axis using the rectified images. Given a pair of images and an estimated fundamental matrix, we can map the images into a map in order to minimize the epipolar distortion. Equivalently, we can compute directly the median axis of the silhouettes along the epipolar lines. In figure 3 we display the median points of the left and right silhouettes. These median points are used to fit a GC to the reconstructed 3D points.

3.2. FITTING GENERALIZED CYLINDERS TO SILHOUETTES

The use of rectified silhouettes image allows us to

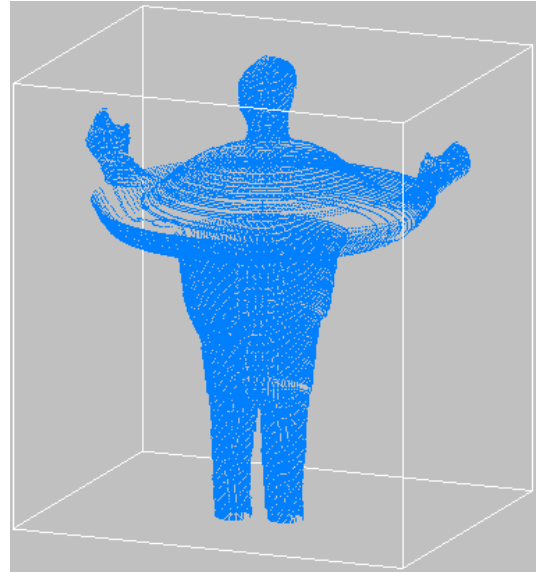


Figure 4: 3D representation of the body posture corresponding to the silhouettes given in 2. The 3D shape is described by a collection of circles corresponding to the cross section of the GC model with the epipolar plane.

easily characterize for each median point a circle centered at the median point and which intersect the silhouette in two points. Indeed, the estimation of the circle radius is done by a search along the epipolar lines (or horizontal line in the case of a rectified image). These circles represent the cross sections of the GC model approximating the 3D data and allows us to characterize the approximating GC through a set of median points and associated radius. This approach relies on the symmetries in the human body shape. Similar approaches were used for object description. However these method enforce a global symmetry constraint to infer the 3D shape from a single 2D image [25]. In our approach, we are enforcing these symmetries only locally along the median axis of the recovered silhouettes.

4. EXPERIMENTAL RESULTS

We have implemented the proposed approach in order to capture human body gesture in the context of a vision-based user interface. The objective here is to infer a reliable and efficient body model in order to understand the user gestures. We have experimented our approach using two views of the user. The system's components are two synchronized video cam-

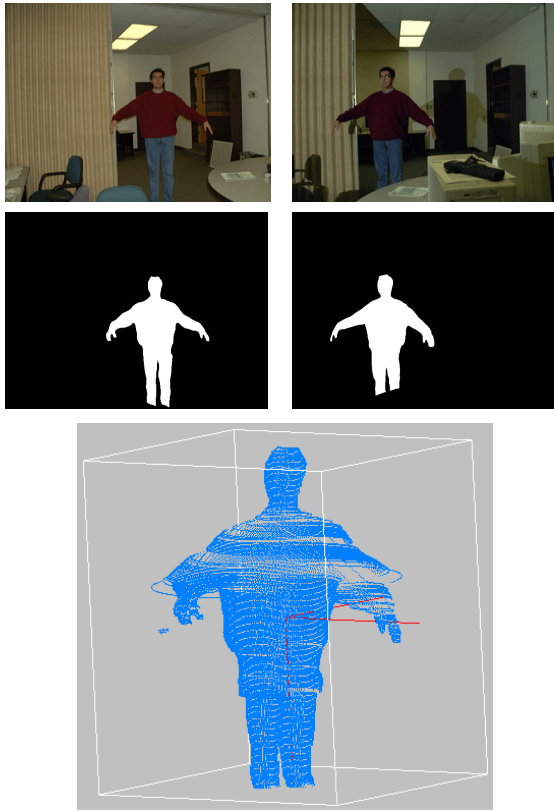


Figure 5: A second example of 3D reconstruction of a human body. We show the acquired frames, the corresponding silhouettes and the integration of the silhouettes into a 3D model.

eras with a large baseline (approximately 3 meters). The cameras are static and therefore we can easily estimate the epipolar geometry associated to the two video streams captured by the cameras. The frames captured by the cameras allows us to infer the silhouette from each view point using a background learning method as described in section . These silhouettes are then processed in order to extract the median axis points. These two views are then integrated using the essential matrix associated to the camera rig generating a set of points in 3D. These 3D points are merged into a collection of circles that describe the cross sections of the generalized cylinder modeling the human body.

In figures 4, 5 and 6 we show various 3D reconstructions of body postures obtained with the proposed approach.

5. CONCLUSION

In this paper we have presented a first toward the use of multiple video streams for a vision-based user

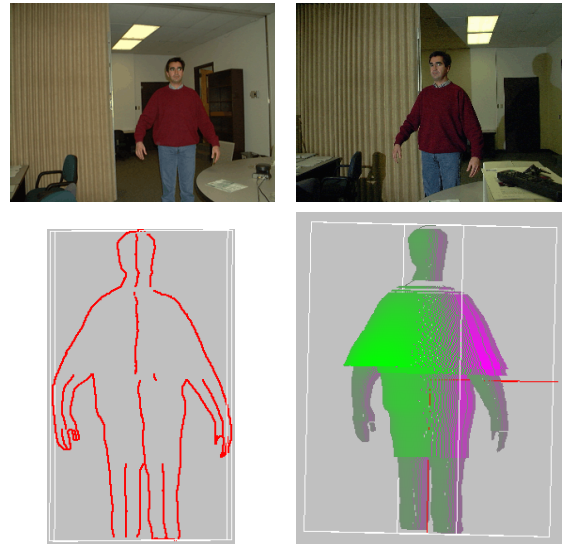


Figure 6: A third reconstruction of a human body. We show the acquired frames and the corresponding 3D model.

interface. The inference of the 3D body model of a human from multiple synchronized video streams provides us more reliable description of the human body motion. This 3D model and the tracking of particular body parts will allow us to analyze and understand the status of the user during his/her interaction with a system. The remaining issues to address are the 3D tracking of the articulated body model (derived from the GC representation) and the interpretation of these track based on the task context.

References

- [1] A.F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, B 352:1257–1265, 1997.
- [2] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Proc. Third IEEE Conference on Face and Gesture Recognition*, Nara, Japan, April 1998.
- [3] Darrell, T., G. Gordon, J. Woodfill, and M. Harville. A virtual mirror interface using real-time robust face tracking. In *AFG98*, page Applications, 1998.
- [4] C.R. Dyer. Volumetric scene reconstruction from multiple views. In L.S. Davis, editor, *Foundations*

- of *Image Analysis*. Kluwer, Boston, 2001.
- [5] D.M. Gavrila and L.S. Davis. 3-d model-based tracking of human upper body movement: A multi-view approach. In *SCV95*, pages 253–258, 1995.
- [6] N.H. Goddard. The interpretation of visual motion: Recognizing moving light displays. In *Proc. Workshop on Visual Motion*, pages 212–220, 1989.
- [7] D. Hogg. Model-based vision: A program to see a walking person. *IVC*, 1(1):5–20, 1983.
- [8] T. Horprasert, I. Haritaoglu, D. Harwood C. Wren, L. Davis, , and A. Pentland. Real-time 3d motion capture. In *Proc. 1998 Workshop on Perceptual User Interfaces (PUI'98)*, pages 87–90, November 1998.
- [9] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima. Human body postures from trinocular camera images. In *Fourth International Conference on Automatic Face and Gesture Recognition, 2000*, pages 326–331, 190.
- [10] N. Jovic, M. Turk, , and T. Huang. Tracking self-occluding articulated objects in dense disparity maps. In *Proc. ICCV'99*, Corfu, Greece, 1999.
- [11] I.A. Kakadiaris and D. Metaxas. Three-dimensional human body model acquisition from multiple views. *IJCV*, 30(3):191–218, December 1998.
- [12] T. Kanade, H. Saito, and S. Vedula. The 3d room: Digitizing time-varying 3d events by synchronized multiple video streams. In *CMU-RI*, 1998.
- [13] M. W. Krueger. Videoplace - an artificial reality. In *Proc. SIGCHI'85*, pages 35–40, New York, April 1985. ACM.
- [14] M.K. Leung and Y.H. Yang. Human body motion segmentation in a complex scene. *PR*, 20(1):55–64, 1987.
- [15] W.N. Martin and J.K. Aggarwal. Volumetric descriptions of objects from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):150–158, March 1983.
- [16] P.J. Narayanan, P.W. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *Proceedings of the International Conference on Computer Vision*, pages 3–10, 1998.
- [17] M. Nicolescu and G. Medioni. Globeall: Panoramic video for an intelligent room. In *Proc. International Conference on Pattern Recognition (ICPR), 2000*, Barcelona, September 2000.
- [18] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraints propagation. *IEEE Transactions on Machine Intelligence*, 2:522–54, 1980.
- [19] R. F. Rashid. Towards a system for the interpretation of moving light displays. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):574–581, 1980.
- [20] M. Turk. Visual interaction with lifelike characters. In *Proc. Second IEEE Conference on Face and Gesture Recognition*, Killington, October 1996.
- [21] M. Turk and G. Robertson. Perceptual user interfaces. *Communications of the ACM*, march 2000.
- [22] M. Turk and Y. Takebayashi, editors. *Proceedings of the Workshop on Perceptual User Interfaces*, Banff, Canada, 1997.
- [23] J.A. Webb and J.K. Aggarwal. Structure from motion of rigid and jointed objects. *AI*, 19(1):107–130, September 1982.
- [24] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland. Pfunder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, July 1997.
- [25] M. Zerroug and R. Nevatia. From an intensity image to 3-D segmented descriptions. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 108–113, Israel, October 1994.
- [26] Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Technical Report 2273, INRIA, 1994.