Accepted for publication:
Second International Workshop on
Articulated Motion and Deformable Objects
Palma de Mallorca, Spain, 21-23 November, 2002.

3D Body Reconstruction for Immersive Interaction

Isaac Cohen and Mun Wai Lee
{icohen|munlee}@iris.usc.edu

Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA 90089-0273

Abstract. In this paper we present an approach for capturing 3D body motion and inferring human body posture from detected silhouettes. We show that the integration of two or more silhouettes allows us to perform a 3D body reconstruction while each silhouette can be used for identifying human body postures. The 3D reconstruction is based on the representation of body parts using Generalized Cylinders providing an estimation of the 3D shape of the human body. The 3D shape description is refined by fitting an articulated body model using a particle filter technique. Identifying human body posture from the 2D silhouettes can reduce the complexity of the particle filtering by reducing the search space. We present an appearance-based learning method that uses a shape descriptor of the $2D$ silhouette for classifying and identifying human posture. The proposed method does not require an articulated body model fitted onto the reconstructed 3D geometry of the human body: It complements the articulated body model since we can define a mapping between the observed shape and the learned descriptions for inferring the articulated body model.

Keywords: *3D body reconstruction, articulated body model, particle filter, posture recognition, support vector machines*

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 by the ambitious goal of achieving a vision-based perceptual user interface in which the state and the action of the user(s) are automatically inferred from a set video cameras. The objective is to extend the current mouse-keyboard interaction techniques in order to allow the user to behave naturally in the immersed environment, as the system perceives and responds appropriately to user actions. Understanding human action in an environment is a challenging task as it involves different granularity in its analysis and description according to the targeted application. For example, describing a human activity in term of its trajectory constitutes a first level of representation, which may be satisfactory for surveillance applications but remains quite insufficient for understanding human gesture in an interactive environment. Indeed, in such situations richer descriptions are required in order to understand the human activity. Deriving a compact

description or signature of the perceived 2D silhouette or of the reconstructed body model is challenging task since the object is an articulated object.

Expensive and cumbersome motion capture systems such as body suits, optical (marker-based) or magnetic tracking systems, and data gloves provide a partial solution. 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: very often a recalibration for specific gestures is required.

Various methods have been proposed for the estimation and analysis of full-body structure (see [10] and references therein). The objective is the development of real-time interactive systems with more sophisticated 2D and 3D tracking and representations [14, 20]. Understanding the human motion from a single visual stream is challenging since only the 2D projection of these motions is captured. Recently several researchers focused on the inference of 3D body model from a monocular 2D camera using a human body model[18, 7]. The main drawback of these techniques is that roughly one third of the degrees of freedom of the human model are nearly unobservable due to motion ambiguities and self-occlusion. Multiple views are therefore required to disambiguate or identify the human motion.

In this paper we present our approach for capturing 3D body motion and inferring human body posture from detected silhouettes. We show that body silhouettes allow to identify the human body posture and furthermore the integration of two or more silhouettes allows us to perform a 3D body reconstruction. 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 [8, 15]. Our approach is based on an incremental integration of 2D silhouettes captured by two or more cameras and the representation of the human body shape using generalized cylinders (GC). This GC-based reconstruction is then augmented by fitting an articulated body model. The articulated model selected in this paper consists of 10 joints and 14 segments leading to a 32 degrees of freedom model. The articulated body model is fitted and tracked in time using a particle filter method. We have defined a likelihood function based on similarity between the extracted body medial axes and the articulated body model. This allows us to avoid the need of estimating the widths of body segments that vary among people. The skeleton structure, on the other hand, does not have a significant variance among people (except for a global scaling factor), and can be used for fitting the model to the extracted data without any additional information.

Identifying the human posture from the set of degrees of freedom (*dof*) of the reconstructed articulated body model requires the mapping of the large space defined by the *dof* onto the corresponding canonical description of human postures. We present an appearance-based learning formalism that uses a shape descriptor of the $2D$ silhouette for classifying and identifying human posture. The proposed method does not require an articulated body model fitted onto the reconstructed 3D geometry of the human body. It complements the articulated body model since we can define a mapping between the observed shape and the learned descriptions for inferring the state of the *dof* of the articulated body model. Our approach is based on a shape descriptor of the 2D silhouettes and a learning algorithm based on Support Vector Machine (SVM) in order to account for variability in human posture.

2 3D Articulated Body Model

2.1 3D Shape from Silhouettes

The integration of the silhouettes of the object captured from the different view points allows us to reconstruct the 3D object shape. The 2D body silhouettes detected using a background learning method, define a set of rays from the camera center to the object in the scene. For each camera, this set of rays defines a generalized cone within which the object lies. Integrating the multiple views by intersecting the corresponding generalized cones defines a volume in the scene where the object lies. This volume provides an approximation of the true 3D shape that depends on the complexity of the shape of the object and on the number of views. Several algorithms have been proposed for the construction of volumetric models from a set of silhouettes. These techniques are based on the discretization of the visible volume by voxels which back-projection corresponds to the observed silhouettes cones. These techniques rely on various implementations of the intersection of the generalized cones. These volume intersection techniques are time consuming and require a large number of views (i.e. cameras). The construction of the 3D human body shape is based on the use of two or more apparent contours and *a priori* geometric model of the body shape. Shape from silhouettes techniques methods focus on inferring a surface description of the object from its occluding contours. They also require a large number of views in order to infer a good shape representation. The *a priori* shape model based on the description of body parts using Generalized Cylinder (GC) allows us to derive a 3D reconstruction from two or more silhouettes.

The human body silhouettes are extracted from multiples calibrated cameras. The estimation of the epipolar geometry between any two views is performed beforehand using a set of feature points. Image pair from these two views are rectified using a planar homography transformation so that the epipolar lines are aligned to image scan lines using the method described in [9]. This rectification reduces the numerical complexity of the registration of apparent contours along the epipolar lines. Cross sections of the extracted silhouettes in the two images are matched along the rectified scan lines. This is repeated for every image pairs to obtain multiple sets of matched cross sections.

Each pair of matched cross sections defines a quadrilateral in the 3D space. A GC is formed by fitting a circle within this quadrilateral. After fitting, the center of the circle defines a median axis of the GC. If more than two cameras are used, a reconstructed circle can be validated by projecting the circle to other views and check if it falls within the silhouette. This validation process eliminates false matches of the silhouettes cross sections.

2.2 Initialization of the Articulated Body Model

A human articulated body model is used for tracking. This model consists of 10 joint and 14 segments, representing the head, torso and limbs. Body constraints such as the limits of the joint angles are expressed in the model. The model consists of 32 degrees of freedom that include the global translation, rotation and scale, and local joint rotations. Fitting and tracking are performed using particle filter, which is described in the next

section. The tracking however requires good initialization. This section describes the initialization process during which the model is fitted to the set of median points using a 3-level hierarchical approach.

1. *Extract the main axis of the body:* The main axis can be extracted using applying principle component analysis (PCA) on the median points and choosing the first principle axis. However, outstretched arms can affect the principal axes. Therefore, we perform the PCA in two stages. The PCA is performed on all the median points, and the first principle axis is extracted. We then remove half of the median points that are furthest away from the first principle axis. Using the remaining half set of median points, we performed PCA again, and extract the first principle axis as the main axis of the body. In this way, the extracted main axis is less sensitive to an outstretched arm or other outliers.
2. *Extract the position of neck and torso:* Using the body's main axis extracted in Level 1, the height of the person can be determined. We can also coarsely estimate the positions of the neck and the torso. We use these estimates to define search regions in which we locate the neck and torso more accurately using heuristics rules. For the neck, we find the position along the main axis where the width of the silhouette is the narrowest. For the torso, we find a maximal-sized rectangle box that is enclosed by the middle section of the silhouette.
3. *Extract the limbs:* The limbs are extracted from the median points using a PCA approach on the points that are not close to the main body axis. In addition, we use geometric constraints that characterize the human body such as the length ratio of the limbs. This is especially useful in extracting the positions of elbows and knees.

In Figure 1.a we show three views of a gesturing person captured by the system and the corresponding silhouettes extracted using background subtraction. After matching of the silhouettes cross sections, the 3D body reconstruction using GCs is performed as shown in Figure 1.b. Due to the effect of shadows, some parts of the ground are reconstructed around the legs of the person, but these can be easily removed using the knowledge of the ground plane. The obtained GC-based 3D body reconstruction displayed in Figure 1.c is used for initializing the articulated body model. The errors of the initialization can be seen at the intermediate nodes such as shoulders and elbows. This is expected because the locations of these joints are not precisely extracted during this initialization. Nonetheless, this coarse fitting provides a sufficiently good initialization of the articulated model. In the subsequent tracking process, the particle filter will provide a better fitting.

2.3 Articulated Model Fitting and Tracking

The tracking method is based on particle filtering, also known as the Condensation algorithm [12]. This method, which uses the Bayesian framework, is a robust online filtering technique for tracking in the presence of clutters.

In particle filtering, the posterior density is represented by a set of N particles denoted by $\{\theta^i_{t-1}\}_{i=1..N}$, with weights $\{q^i_{t-1}\}_{i=1..N}$. There are 3 basic steps: selection, prediction and updating.

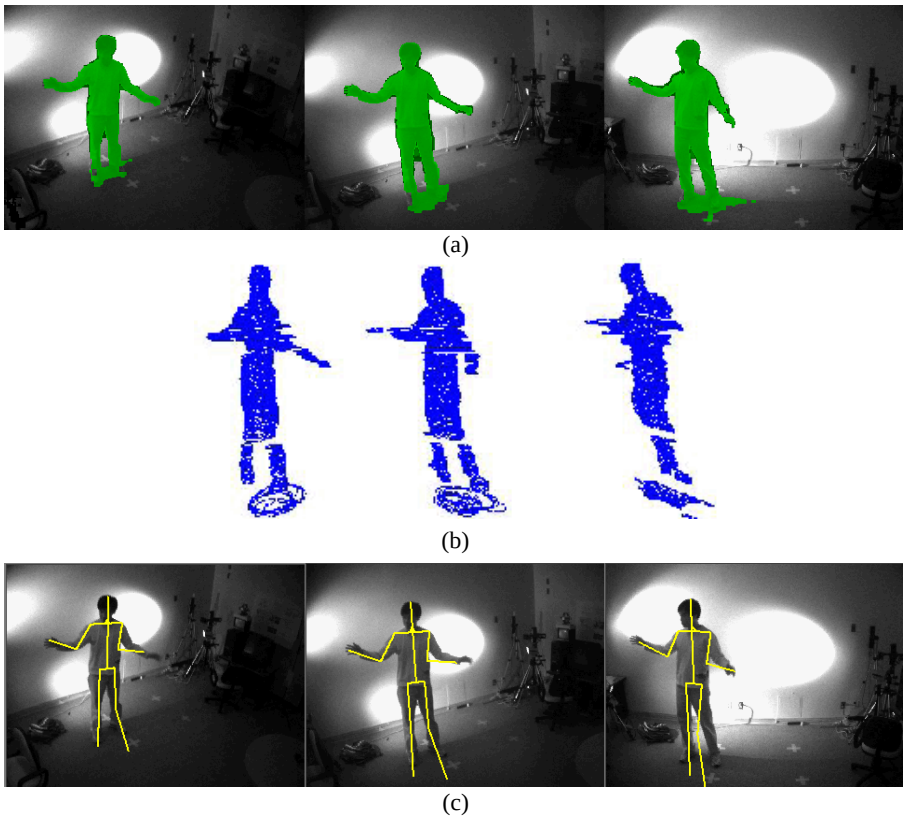


Fig. 1. 3D body reconstruction from three views using a GC-based shape description. (a) displays the views of a gesturing person with the corresponding silhouettes. (b) shows three view points of the reconstructed 3D model. (c) shows the derived initialization of the 3D articulated body model back-projected onto 2D views.

1. *Selection.* Resample with replacement the N particles $\{\tilde{\theta}^i_{t-1}\}_{i=1..N}$, from the set $\{\theta^i_{t-1}\}_{i=1..N}$. The probability of selecting a particle θ^i_{t-1} is proportional to its normalized weight q^i_{t-1}.
2. *Prediction.* The samples are updated according to a stochastic diffusion model,

$$\theta^i_t = \tilde{\theta}^i_{t-1} + w$$

   where w is a vector of standard normal random variables.
3. *Updating.* Given an observation z_t, the weights are updated by the likelihood estimates,

$$q^i_t \propto p(z_t|\theta^i_t)$$

   The weights are normalized so that $\sum_{i=1}^{N} q^i_t = 1$.

Particle filtering has previously been used for human body tracking in [6], which uses a likelihood function based on edge and region measurements. This requires prior knowledge about the widths of articulated body segments.

In our method, the likelihood function is based on similarity between the extracted body median axes, with the articulated body model. In this way, we avoid the need to estimate the widths of body segments that vary among people. The skeleton structure, on the other hand, does not have significant variance among people (except for a global scaling factor), and can be used to fit to the extracted data without any addition information.

For each particle θ_t^i, we compute the positions of a set of L line-segments, $M_t^i = \{l_1, l_2, ..., l_L\}$ that represent the skeleton of the human body. From the GC reconstruction, we have extracted a set of median axes, denoted by $\{x_t^j\}_{j=1...K_t}$, where K_t is the number of median axes. We want to match these median axes to the line segments of the body model. For each median axis, x_t^j, the error of fit of the model to the 3D reconstruction is given by:

$$\mathcal{E}(x_t^j, M_t^i) = d(x_t^j, l_m)$$

where $m = \arg\min\ d(x_t^j, l_i)$ and $d(x_t^j, l_m)$ is the Euclidean distance of the point x_t^j to the line segment l_m. Given a set of K_t median axes, the error of fit of the model to the data is defined by:

$$\delta_t^i = \frac{1}{K_t} \sum_{j=1}^{K_t} \mathcal{E}(x_t^j, M_t^i).$$

The likelihood estimate is computed using a zero-mean Gaussian function,

$$p(z_t|\theta_t^i) = p\left(\{x_t^j\}_{j=1,...,K_t}|M_t^i\right) = \mathcal{N}(\delta_t^i, 0, \sigma).$$

where σ^2 is the noise variance. This likelihood estimate is used to update the weight q_t^i of the particle θ_t^i.

We have used this particle-based tracking method for fitting and tracking an articulated body model from multi-view video sequences. The video sequences were captured in our laboratory using three synchronized cameras. The particle filter is initialized using the hierarchical fitting method described in Section 2.2. A total of 400 particles were used for tracking. The 2D projection of the fitted articulated body model is shown in Figure 2, along with the corresponding 3D body model rendered at arbitrary views.

3 Human Posture Recognition

Identifying the human posture from the set of degrees of freedom (*dof*) of the reconstructed articulated body model requires the mapping of the large space defined by the *dof* onto the corresponding canonical description of human postures. We present in this paper a global description or signature of human shape allowing to identify the posture of the observed human from 2D silhouettes. This signature has to account for variability in characterizing a posture or a gesture. Indeed, several people will perform similar gestures differently and therefore identifying a gesture from the 2D/3D shape descriptions

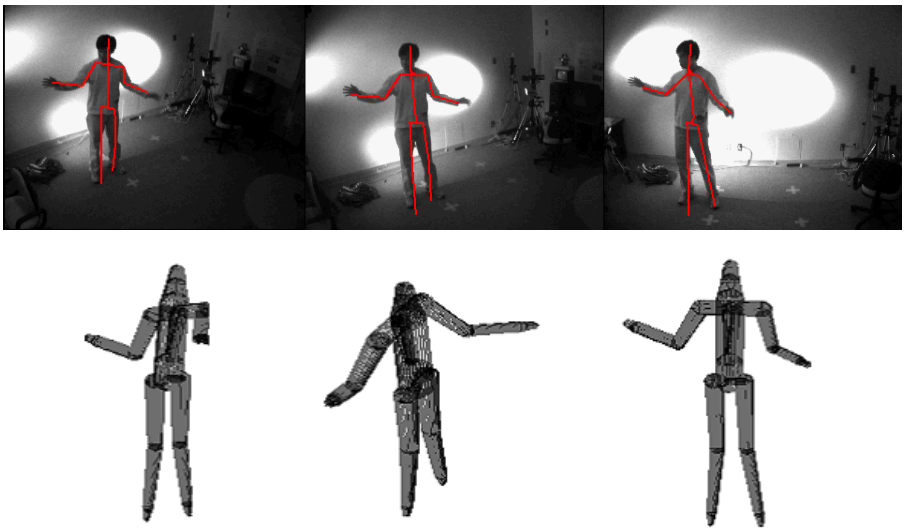


Fig. 2. Articulated body model fitting and tracking derived from the GC-based 3D reconstruction. The first row illustrates the back-projection of the articulated model onto the acquired 2D images. The reconstructed 3D model obtained using 400 particles is rendered from different view points.

will require a learning step. We present an appearance-based learning formalism that uses a shape descriptor of the $2D$ silhouette for classifying and identifying human posture. The proposed method does not require an articulated body model fitted onto the reconstructed 3D geometry of the human body. In fact, it complements the articulated body model since we can define a mapping between the observed shape and the learned descriptions for inferring the articulated body model. In the following section we will present the shape descriptor considered and the learning algorithm based on Support Vector Machine (SVM).

3.1 Human Body Shape Description

Shape descriptors have been well studied in various fields as they are used for determining the similarity between two shapes. The derived descriptors can be classified in terms of the shapes they characterize *i.e.* 2D contours, 3D surfaces, 3D volumes... to name a few; bending energy functions [23], spin images [13], harmonic shape images [24], shape context descriptors [2]. These descriptors were mainly used for shape matching and therefore focused on characterizing the local properties of the shape. Global models, assume a description of the objects into a set of features or parts segment. Common description rely on parametric models [3, 19, 22], deformable regions [1, 4, 5], shock graphs [17]... Shape similarity is then measured by comparing location of features and their spatial distributions. The performances of these approaches depend on the difficult task of segmenting the shape into its corresponding parts. These techniques perform well in the case of shapes of fixed configurations and are not suitable

for modeling variability in the observed shapes such as a gesturing person. Finally, a third description approach is based on modeling the geometric distribution of the shape properties such as histograms of angles [11], algebraic moments [16]... These descriptions are view dependent and do not perform well as the localization of the features is lost in the statistical representation used (commonly a histogram).

We present a statistical shape description model that preserves the localization of the geometric features considered. This global representation allows a robust description of shape that accommodates variation of the shape. Indeed, as one would expect a small shape variation induces a small change in the object description. Moreover, this variation is localized and does not interfere with the global representation of the object. These properties of the proposed shape description are crucial for efficiently representing human shape and its variations.

The shape descriptor used is a generalization of the shape context descriptors (SCD) [2], it extends the local representation into a global shape description. Given a 2D silhouette of a human shape we compute a reference circle $\mathcal{C}_{\mathcal{R}}$ defined by the centroid of the silhouette and its main axis. This circle is uniformly sampled into a set of points P_i. We then consider a polar encoding of the projection of the silhouette onto the set of points P_i. For every point Q_j of the silhouette we accumulate (r, q) where $Q_j - P_i = r(\cos q, \sin q)$.

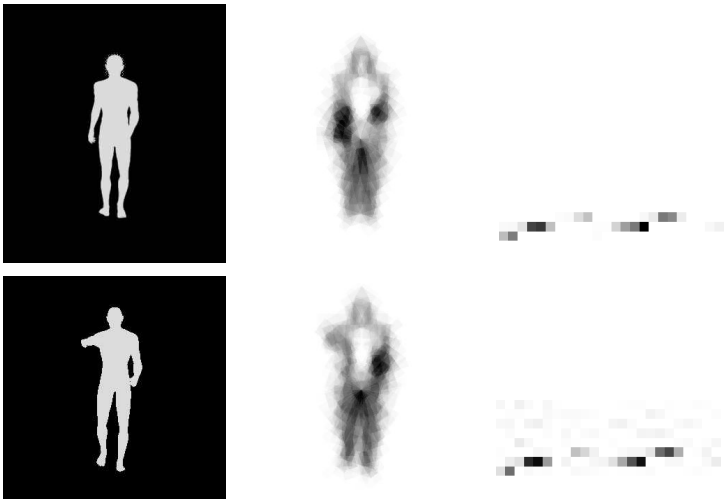


Fig. 3. Examples of global shape descriptors of a walking person (first row), and walking and pointing (second row). The middle column depicts the encoding process while the right column displays the shape signature in polar coordinates.

The polar maps derived for each point P_i are then summed and constitute the signature of the shape. In Figure 3 we show an example of such representation where the signature function, in polar coordinates was centered on the centroid and rendered in

Cartesian coordinates. Note that this only a visualization of the representation map. The real signature is in the polar coordinate system.

3.2 Human Posture Inference

Deriving the human posture from the shape descriptor of its silhouettes in 2D or from the reconstructed shape in 3D is a challenging task since it requires taking into account posture variability across people. A method commonly used relies on the articulated body model in order to infer the human posture. The recovery of an articulated body model still requires the interpretation of the 30plus-degree of freedom in order to infer the human posture. This interpretation has to take into account posture variability and errors in the estimation of the articulated model in order to perform an efficient analysis of the 30plus-D parameter space.

In this paper we combine model-based and appearance-based description of human activity for the inference of the person posture. We reduce the complexity of the search in the large parameter space of articulated body model by integrating the appearance-based shape descriptor introduced in the previous section.

We use a Support Vector Machine formalism [21] to learn and classify the set of heterogeneous information provided by the appearance-based descriptors and the degrees of freedom of the articulated body model. The main advantage of using a SVM is its ability to compress the information contained in the training set, since only support vectors are required for the classification.

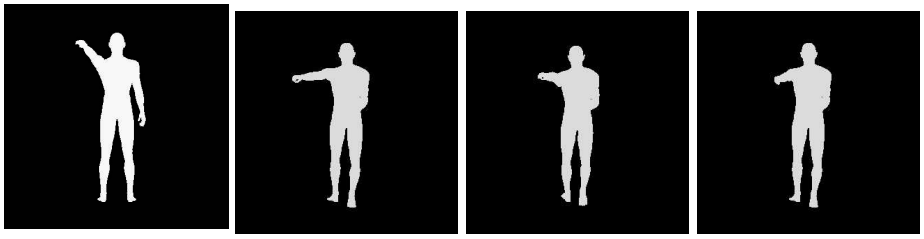
The main issues in using a machine learning approach are the selection of the features used as training data set and the choice of the data set for training the model. While an articulated body model provides a natural set of features to consider for training purposes, it is time consuming and it is difficult to acquire the 30plus degrees of freedom of the selected model. Conversely, appearance-based shapes are very easy to collect but a correct representation of the shape has to be selected in order to be significant for learning.

The problem we are addressing here is the definition of a decision function that from a set of observations $x \in \mathcal{X} = \{x_i, i = 0..N\}$ and the corresponding labels $y \in \mathcal{Y} = \{y_i, i = 0..N\}$ will make accurate classification of unseen values of x. A very successful approach for solving this supervised learning problem is the support vector machine (SVM) [21]. In this work we are interested in a classification of the observed human postures, therefore the set of available labels is limited to $\mathcal{Y} = \{-1, 1\}$. The decision function is defined by the SVM is:

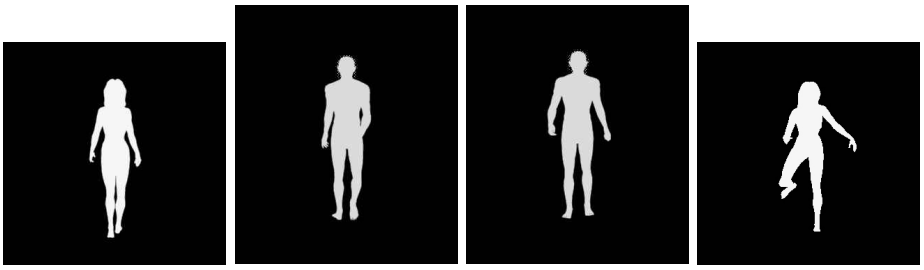
$$f(x) = sgn\left(\sum_{i=0}^{l-1} \alpha_i^0 y_i K(x_i, x) + b\right),$$

where the coefficients α_i^0 are obtained by maximizing the functional:

$$W(\alpha) = \sum_{i=0}^{l-1} \alpha_i - \frac{1}{2} \sum_{i,j=0}^{l-1} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$



a)Right-hand pointing gestures



b)Non-pointing gestures

Fig. 4. a) and b) Examples of silhouettes used for training the SVM for recognizing right-hand pointing gesture.

under the constraints:

$$\sum_{i=0}^{l-1} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0.$$

The coefficients α_i^0 define a maximal margin hyper-plan in the high dimensional feature space where the data are mapped through the non linear function Φ such that $\Phi(x_i) \cdot \Phi(x_j) = K(x_i, x_j)$. Various kernels K are commonly used (linear, exponential, polynomial...) we will use a linear kernel K using therefore a linear mapping between the feature space and the representation space.

3.3 Training the Support Vector Machine and Classification

The SVM is trained using vectors containing both the 2D shape descriptors defined in the previous section complemented by the degrees of freedom of the articulated body model. The set of instances used for learning are obtained from natural observation as well as computer graphics simulation of human gestures. In the Figure 4 we show some 2D examples used for training the model to identify a right hand pointing from a front view camera. This training dataset was successfully used to classify the right hand pointing gestures captured by a front-view camera and illustrated in Figure 5. This simple example illustrates the use of a global shape description and a SVM for human posture recognition.

(a) Synthetic example (b) (c)

Fig. 5. Examples of right-hand pointing silhouettes identified by the SVM. The SVM classified synthetic examples (generated by Poser) as well as real pointing silhouettes acquired by our system as shown in (b) and (c).

4 Conclusion

We have presented in this paper a method for 3D body reconstruction from multiple silhouettes. The integration of silhouettes using a GC-based description of human parts allows us to derive a good description of the human body in 3D. An articulated body model is fitted to the reconstructed 3D data using a particle filter. The likelihood function used for the particle filtering relies on the similarity between the extracted body median axes and the articulated body model.

The large numerical complexity of the particle filtering prevents us from considering a real-time implementation. However the appearance-based body posture recognition from the 2D silhouettes provides a solution to be considered in the future for reducing the search space of the particle filter technique. Indeed the shape description used along with the SVM learning algorithm allows us to account for variability in body posture while providing a good classification rate.

Acknowledgment

This research was supported by a grant from the Institute for Creative Technologies (ict.usc.edu).

References

1. R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. *Vision Research*, (38):2365–2385, 1998.
2. S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *IEEE Proceedings of the International Conference on Computer Vision*, Vancouver, Canada, July 2001.
3. T. Binford. Visual perception by computer. In *IEEE Conference on Systems Science and Cybernetics*, 1971.

4. I. Cohen, N. Ayache, and P. Sulger. Tracking points on deformable objects using curvature information. In *Proceedings of the Second European Conference on Computer Vision*, Santa Margherita Ligure, Italy, May 1992.
5. I. Cohen and I. Herlin. Curves matching using geodesic paths. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, Santa Barbara, June 1998.
6. J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, Hilton-Head, 2000.
7. D. DiFranco, T. Cham, and J. Rehg. Reconstruction of 3d figure motion from 2d correspondences. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, December 2001.
8. S. Iwasawa et al. Human body postures from trinocular camera images. In *International Conference on Automatic Face and Gesture Recognition*, pages 326–331, 2000.
9. J. Gluckman and S. K. Nayar. Rectifying transformations that minimize resampling effects. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, Kauai, December 2001.
10. A. Hilton and P. Fua. Modeling people toward vision-based understanding of a person's shape, appearance, and movement. *Computer Vision and Image Understanding*, 81(3):227–230, 2001.
11. K. Ikeuchi, T. Shakunaga, M. Wheeler, and T. Yamazaki. Invariant histograms and deformable template matching for sar target recognition. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, 1996.
12. M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proceedings of the European Conference on Computer Vision*, pages 343–356, 1996.
13. A. E. Johnson and M. Hebert. Using spin-images for efficient multiple model recognition in cluttered 3-D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.
14. I.A. Kakadiaris and D.Metaxas. Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, 30(3):227–230, 1998.
15. T. Kanade, H. Saito, and S. Vedula. The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams. Technical report, CMU-RI, 1998.
16. R. J. Prokop and A. P. Reeves. A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP: Graphics Models and Image Processsing*, 54(5):438–460, 1992.
17. K. Siddiqi, A. Shokoufandeh, S. J. Dickinson, and S. W. Zucker. Shock graphs and shape matching. *Computer Vision*, pages 222–229, 1998.
18. C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, December 2001.
19. F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990.
20. M. Turk and G. Robertson. Perceptual user interfaces. *Communications of the ACM*, March 2000.
21. V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
22. K. Wu and M. Levine. Recovering parametrics geons from multiview range data. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, pages 159–166, June 1994.
23. I. Young, J.Walker, and J. Bowie. An analysis technique for biological shape. *Computer Graphics and Image Processing*, (25):357–370, 1974.
24. D. Zhang and M. Hebert. Harmonic maps and their applications in surface matching. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, 1999.