Image-Based Reconstruction for View-Independent Human Motion Recognition

Robert Bodor **Bennett Jackson** **Osama Masoud** **Nikolaos Papanikolopoulos**
{rbodor, jackson, masoud, npapas}@cs.umn.edu
AIRVL, Dept. of Computer Science and Engineering, University of Minnesota.

Abstract-- *In this paper, we introduce a novel method for employing image-based rendering to extend the range of use of human motion recognition systems. We demonstrate the use of image-based rendering to generate additional training sets for view-dependent human motion recognition systems. Input views orthogonal to the direction of motion are created automatically to construct the proper view from a combination of non-orthogonal views taken from several cameras. To extend motion recognition systems, image-based rendering can be utilized in two ways: (i) to generate additional training sets for these systems containing a large number of non-orthogonal views, and (ii) to generate orthogonal views (the views those systems are trained to recognize) from a combination of non-orthogonal views taken from several cameras. In this case, image-based rendering is used to generate views orthogonal to the mean direction of motion. We tested the method using an existing view-dependent human motion recognition system on two different sequences of motion, and promising initial results were obtained.*

Index Terms—image-based rendering, human motion recognition, computer vision, visual tracking, shape reconstruction.

I. Introduction

The problem of using computer vision to track and understand the behavior of human beings is a very important one. It has applications in the areas of human-computer interaction, user interface design, robot learning, and surveillance, among others.

Much work has been done in the field of human motion recognition. A great deal of this work has been done using a single camera providing input to a training-based learning method ([1], [3], [4], [8], [9], [10], [12], [16], [17], [18], [22], [23]). Methods that rely on a single camera, are highly view-dependent. They are usually designed to recognize motion when viewing all subjects from a single viewpoint (e.g., from the side, with the subject moving in a direction orthogonal to the camera's line of sight, as in Figure 1). The training becomes increasingly ineffective as the angle between the camera and the direction of motion varies away from orthogonal. As a result, these methods have limited real-world applications, since it is often impossible to limit the direction of motion of people.

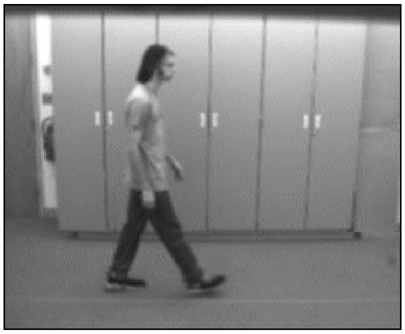

Figure 1. A single frame of human motion taken from a camera positioned orthogonal to the direction of motion.

Many of the methods track a single person indoors for both gesture recognition and whole body motion recognition ([1], [4], [22], [23]). Bregler details a method for motion analysis by modeling the articulated dynamics of the human body in outdoor scenes taken from a single camera [3]. Analysis of human gaits is also very popular in this domain. Statistical learning methods were developed by Fablet and Black to characterize walking gaits from 12 angles [8]. Additionally, Polana and Nelson studied the periodicity of the legs while walking or running to classify pedestrian motion [17]. Gavrila and Davis [10] used multiple synchronized cameras to reconstruct 3D body pose and study human motion based on 3D features, especially 3D joint angles. Experimental data was taken from four near-orthogonal views of front, left, right, and back. Oliver *et. al.* used coupled hidden Markov models along with a Kalman filter in a hierarchical approach to track and categorize motion [16].

Rosales and Sclaroff [18] developed a trajectory-based recognition system that is trained on multiple view angles of pedestrians in outdoor settings. This system can avoid the common problem of being limited in its ability to recognize motion from only a small set of view angles. However, the method requires a tremendous amount of training data. Difranco *et. al.* [6] tackle the problem of reconstructing poses from challenging sequences like athletics and dance from a single view point based on 2D correspondences specified either manually or by separate 2D registration algorithms. They also describe an interactive system that makes it easy for users to obtain 3D reconstructions very quickly based on a small amount of manual effort. In [11], Luck *et. al.* generate a 3D voxelized model of a moving human, extracting the moving form using adaptive background subtraction and thresholding in real-time. Joint angle constraints are used to reconstruct an 11-degree of freedom model of the body. Additionally, silhouettes have been used widely as tools for recognition. Weik and Liedtke [21] use 16 cameras and a shape-from-silhouette method to build a model template for the body, which is then matched to volume data generated in each successive frame. Pose analysis is performed based on an automatic reconstruction of the subject's skeleton. Silhouettes are also used by [2] and [5] to classify poses.

II. Description of the Work

We are studying the use of image-based rendering to generate additional training sets for these motion recognition systems. Orthogonal input views can be created automatically using image-based rendering to construct the proper view from a combination of non-orthogonal views taken from several cameras (Figure 2).

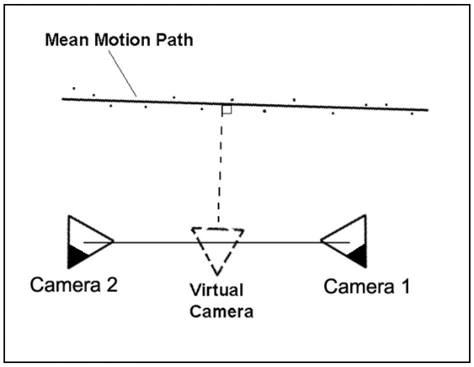

Figure 2. Camera positions relative to motion path.

Image-based rendering can be applied to motion recognition in two ways: (i) to generate additional training sets containing a large number of non-orthogonal views and (ii) to generate orthogonal views (the views that 2D motion recognition systems are trained for) from a combination of non-orthogonal views taken from several cameras. Figure 3 shows two examples of non-orthogonal novel views generated from three different real camera views using image-based rendering.

The advantage of the first approach is that action recognition can still be performed using a single camera. However, view-dependent recognition may not be suitable for certain views (e.g., an action performed along the optical axis). This may be due to shortcomings in the specific recognition algorithm or merely because the information content from that view is not sufficient for action recognition. The second approach resolves this issue by always generating a content-rich orthogonal view, or other appropriate view optimized for recognition of specific actions. This requires the use of multiple synchronized cameras during the recognition phase (rather than the training phase). The bulk of this paper focuses on this second approach.

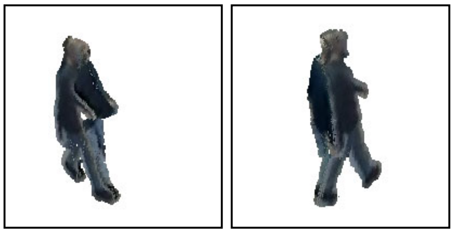

Figure 3. Two non-orthogonal novel views generated by the image-based renderer.

To test the system, images created by the image-based renderer are processed by an existing view-dependent human motion recognition system (subsection F).

The system consists of the following steps:

1) Capture synchronized video sequences from several cameras spread around the perimeter of an area of interest.
2) Calibrate the cameras and compensate for wide-angle lens distortions.
3) Segment the subject in the images (separate foreground / background).
4) Reconstruct the motion path of the subject in the 3D world frame.
5) Calculate the optimal "virtual" camera intrinsic and extrinsic parameters for use in the image-based renderer.
6) Use an image-based renderer to create orthogonal views as training and test data.
7) Test the views using a view-dependent motion recognition system.

A. *Image Capture*

The use of image-based rendering requires capturing multiple views of the same scene simultaneously. We implemented this approach for our human motion recognition application with three video cameras positioned around one half of a room (see Figure 4).

We used three Panasonic GP-KR222 digital video cameras, with Rainbow auto-iris lenses. Two of the lenses were wide-angle, with focal lengths of 3.5mm, while the third had a focal length of 6mm. Each camera was connected to a VCR, and the video was recorded onto three VHS tapes. The video from the tapes was then digitized and separated into sequences of single frames, and all three sequences were synchronized. The cameras were calibrated to determine all intrinsic and extrinsic parameters using the method of Masoud *et. al.* described in [13]. In addition, the images were corrected for lens distortion due to the wide-angle. For part of this procedure we used the process of Ojanen [15].

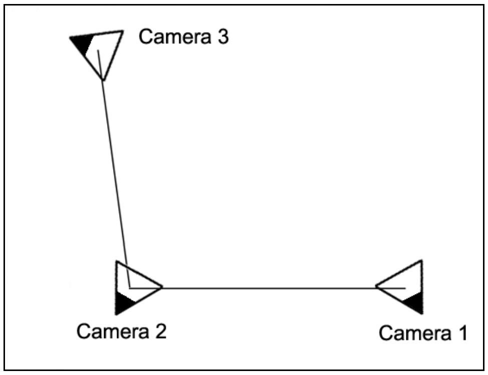

Figure 4. Layout of cameras for synchronized capture.

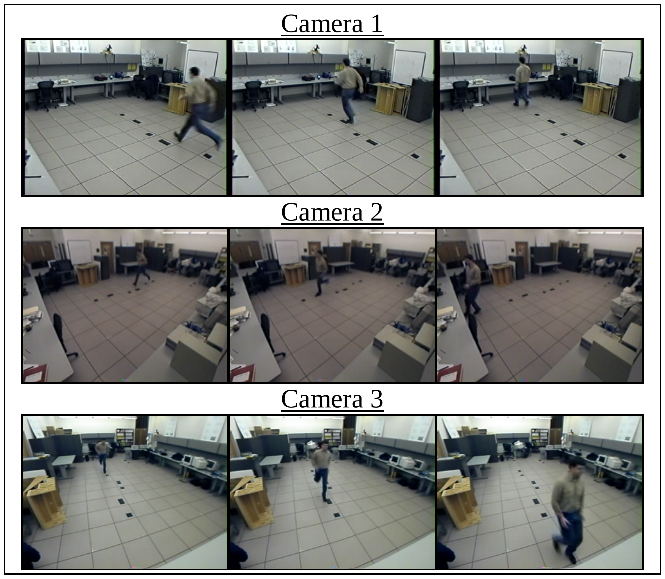

Figure 5. Images of a single running sequence captured from 3 cameras positioned around a room (undistorted).

B. *Foreground Segmentation*

It was necessary to isolate the human for reconstruction in each frame. This required an automated process to separate the human (foreground) from the background of the room. We attempted a variety of approaches to this problem, and found that the best results could be achieved by using a combination of simple background subtraction, chromaticity analysis, and morphological operations.

Each frame in the motion sequence was subtracted from the background image in grayscale. This is a simple, standard approach, and is well known to fail in many cases where adaptive backgrounding has been found to be more robust, particularly in changing environments such as outdoor scenes [20]. Since our sequences were taken indoors, background subtraction performed quite well, with the exception of leaving significant shadows below the subject in the foreground image (see Figure 6). This was particularly problematic in our case, since articulation of the legs is critical for motion recognition.

To address this problem, we converted the RGB images into chromaticity space in an approach similar to that described by Elgammal *et. al.* [7]. We then applied a subtraction and threshold operation to the each frame (see Figure 6). This approach worked very well at removing shadows, but had the negative effect of also removing much of the valid foreground.

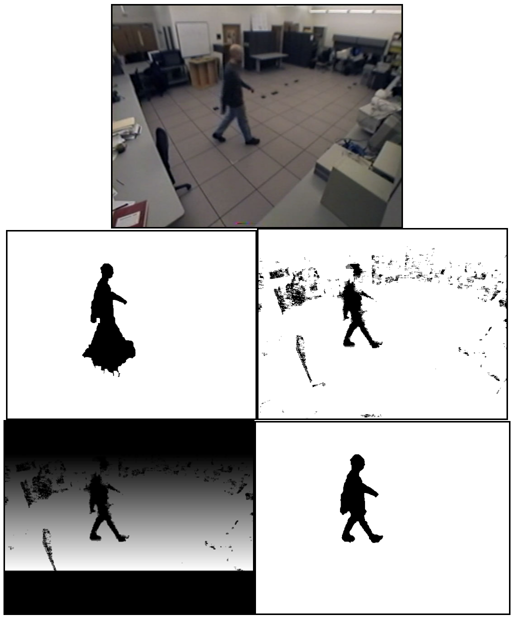

Figure 6. Foreground process (left to right): Source image, background subtraction image containing shadow, chromaticity subtraction image, weighted chromaticity subtraction image, and composite mask with shadow removed.

To retain the best qualities of each method, we generated a composite mask for the final foreground image. This mask was a linear combination of the mask generated by background subtraction and the chromaticity mask. The chromaticity mask was first multiplied by a linear confidence weighting, with full confidence at the bottom of the bounding box of the subject (to remove shadows) and zero confidence at the top (to avoid the chromaticity mask removing valid foreground). Lastly, a series of standard morphological operations such as hole filling, dilation, and erosion were applied to the composite mask.

C. *Calculation of Motion*

From each foreground mask, a perimeter outline image was calculated. This perimeter was used to calculate input silhouette contours for the image-based rendering (see subsection E below). It was also used to extract the centroid (based on the perimeter) and the bottom center of the subject in each image (see Figure 7).

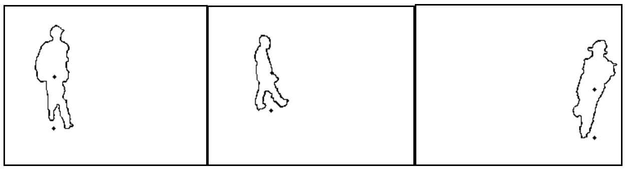

Figure 7. Silhouette perimeter images with centroid and bottom center point displayed.

The bottom center point is simply the point where a vertical line drawn through the centroid of the figure intersects the horizontal line at the bottom of the mask in the perspective of each camera. These points are used to calculate the location of the subject on the plane of the floor in the world frame. We chose the world frame relative to camera 1 and constructed the geometry using the transformations from the calibration step. The subject's position is calculated by projecting each of the three points into the world frame and calculating the Euclidean mean. When this is done at each time step, the subject's entire motion path along the floor can be tracked. Figure 8 shows the three bottom center points projected into the world frame for the first and last positions in the running sequence, along with the averaged position estimates.

D. Determination of Virtual Camera Parameters

To create the optimal input for the motion recognition system, it is critical to position the virtual camera such that it is orthogonal to the mean motion path of the subject. These calculations were done in the world frame using the results of the calibration procedure.

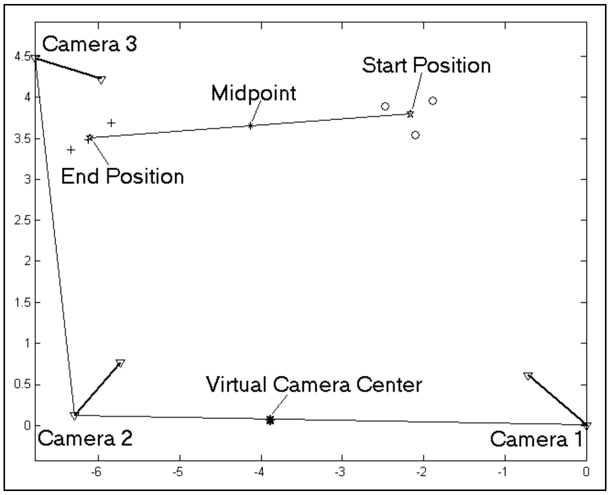

Figure 8. Running path and virtual camera's position relative to fixed cameras for the subject's motion sequence.

It is also important to consider the limitations of the image-based renderer when choosing the viewpoint for the virtual camera. The quality of reconstructed views provided by image-based rendering degrades as the viewpoint of the virtual camera moves away from the views of the real cameras in the scene. Therefore, we chose to position the virtual camera center in a plane between the real cameras (see Figure 8). The virtual camera's orientation was calculated to intersect the midpoint of the motion path at 90 degrees. In addition, the field of view of the virtual camera was automatically calculated to encompass the entire motion sequence without introducing wide-angle distortions.

E. Image-Based Rendering of Orthogonal Views

To generate novel views, we use an image-based approach. Image-based methods have several advantages over other volume- or 3D-based methods. Efficiency and realism are two such advantages. Given the parameters of a novel virtual camera, we use an approach similar to view morphing [19]. The views we use to render the new image are those of the nearest two cameras. Our approach differs in that we do not restrict the novel camera center to lie on the line connecting the centers of the two cameras. In order for view morphing to work, depth information in the form of pixel correspondences needs to be provided. To compute this, we use an efficient epipolar line clipping algorithm [14] which is also image-based. This method uses the silhouettes of an object to compute the depth map of the object's visual hull. Once the depth map is computed, it is trivial to compute pixel correspondences.

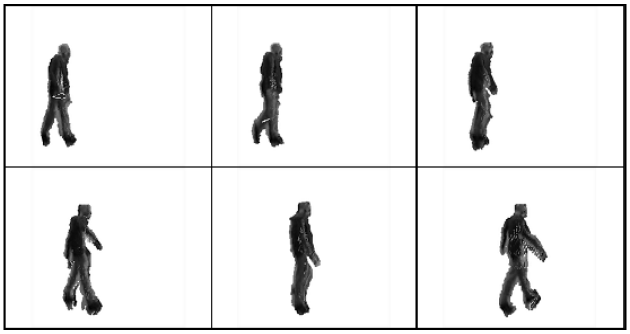

Figure 9. Rendered orthogonal walking sequence.

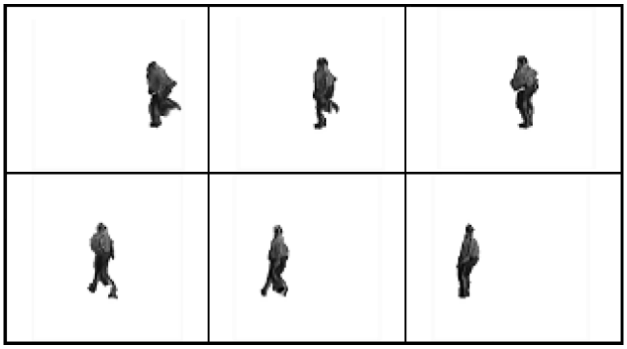

Figure 10. Rendered orthogonal running sequence.

F. Motion Recognition System

As mentioned earlier, we use a view-dependent human motion recognition method [12] to study the feasibility of our approach. This motion recognition system uses principal component analysis to represent a periodic action as a set of points (a manifold) in a low dimensional eigenspace. Training involves computing, for each action, an "average" manifold, which is used as a representation for that action. The manifold of a new action (whose images are computed orthogonally using image-based rendering) is then compared to all representative manifolds by computing a distance measure. The minimum distance is used to determine the classification result.

The images of an action are first processed to produce *feature images*. These images capture instantaneous motion since they are computed by recursively filtering the sequence images. Figure 11 shows some examples of feature images computed for the sequence in Figure 9.

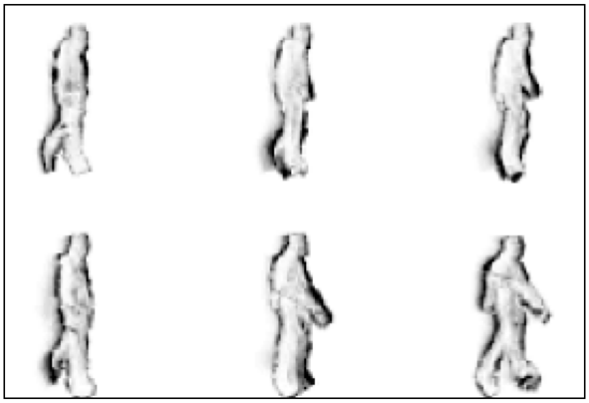

Figure 11. Examples of feature images corresponding to the snapshots shown in Figure 9. These images are computed by recursive filtering and are used by the recognition algorithm.

In the training phase, a basis consisting of the principle components is first computed using all feature images of the training data. Then, each feature image is projected using the basis. Thus, an action can be represented as a sequence of points in an eigenspace (a manifold). For every action, the average manifold is computed and stored in a database. Testing is performed by computing the manifold of the test action and finding the action in the database whose manifold is most similar. The comparison is done using a Hausdorff-like metric that is invariant to the phase and speed of the action. In our experiments, the system was trained to classify an action into one of eight categories: walk, run, skip, march, hop, line-walk, side-walk, and side-skip. Training sequences were provided by eight subjects performing all eight actions while moving orthogonal to the camera (a total of 64 sequences). The subjects used in testing were not among those eight subjects.

III. RESULTS

We tested the virtual views produced by the image-based renderer on the walking and running sequences shown above, and used the original training data captured from sequences taken orthogonal to the subject's motion, as in Figure 1. Each image sequence was processed in three ways and each was input to the motion recognition system.

The first set of three input sequences consisted of the source video sequences taken from all three cameras**.** In all cases, the system failed to identify the motion. This was true even when part of the sequence was orthogonal to the camera and the mean motion path was near orthogonal.

Test Sequence	**Angle from**	**Result**
	Orthogonal	
Subject A Walking		
Camera 1 source images	74.1°	Failed
Camera 2 source images	15.9°	Failed
Camera 3 source images	81°	Failed
Foreground images	15.9°	Failed
Image-based rendered orthogonal images	0°	Motion recognized
Subject B Running		
Camera 1 source images	44.5°	Failed
Camera 2 source images	45.5°	Failed
Camera 3 source images	69.9°	Failed
Foreground images	44.5°	Failed
Image-based rendered orthogonal images	0°	Motion recognized

The fourth input sequence in each case was the set of foreground images of the camera that had the smallest angle away from orthogonal (see Figure 8). We tested the foreground sequence to ensure that any classification result was not due to simply removing the background and thus somehow better isolating the subject's motion for the recognition system. In this case, the system also failed to classify the motion correctly.

The fifth input sequence was the set of virtual images generated by the image-based renderer from a view orthogonal to the motion path (see Figures 9 and 10). This sequence was correctly identified for both motions.

IV. CONCLUSIONS/FUTURE WORK

The method described here may be used in several ways to reduce the constraints necessary to employ 2D recognition in many environments. This may be done by creating a comprehensive set of training data for 2D motion recognition methods with views of motion from all angles, or by converting non-orthogonal views taken from a single camera into orthogonal views for recognition. In addition, the method can be used to provide input to a three-dimensional motion recognition system because the system creates a 3D volume model of the subject over time using only foreground segmentation.

In the future, we intend to test the method comprehensively with a large base of subjects and motion types. In addition, we intend to test the system in outdoor as well as indoor environments, and attempt to determine image capture parameters such as the minimum angle where recognition begins to degrade and the number of cameras necessary to achieve various levels of performance.

V. Acknowledgements

This work has been supported by the National Science Foundation through the NSF grant #IIS-0219863.

VI. References

[1] Ben-Arie, J., Wang, Z., Pandit, P., and Rajaram, S. "Human Activity Recognition Using Multidimensional Indexing," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, No. 8, August, 2002.

[2] Bradski, G. and Davis, J. "Motion Segmentation and Pose Recognition with Motion History Gradients," *Int'l Journal of Machine Vision and Applications*, vol. 13, No. 3, 2002, pp. 174-184.

[3] Bregler, C. "Learning and Recognizing Human Dynamics in Video Sequences," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 568-574, 1997.

[4] Cutler, R. and Turk, M. "View-Based Interpretation of Real-Time Optical Flow for Gesture Recognition," *Proc. of the Third IEEE Conf. on Face and Gesture Recognition*, Nara, Japan, April 1998.

[5] Davis J. and Bobick, A. "The Representation and Recognition of Human Movement Using Temporal Templates," *Proc. of the Computer Vision and Pattern Recognition*, pp. 928-934, June 1997.

[6] DiFranco, D.E., Cham, T-J., and Rehg, J.M., "Reconstruction of 3-D Figure Motion from 2-D Correspondences," *Proc. of the 2001 IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 307 –314, 2001.

[7] Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L.S. "Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance," *Proc. of the IEEE*, vol. 90, 2002.

[8] Fablet, R. and Black, M.J. "Automatic Detection and Tracking of Human Motion with a View-Based Representation," *European Conf. on Computer Vision, ECCV'02*, May 2002.

[9] Gavrila, D.M. "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, No. 1, 1999.

[10] Gavrila, D.M. and Davis, L.S., "3-D Model-Based Tracking of Humans in Action: A Multi-View Approach," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 73 –80, 1996.

[11] Luck, J., Small, D., and Little, C.Q. "Real-Time Tracking of Articulated Human Models Using a 3D Shape-from-Silhouette Method," *Intl. Workshop on Robot Vision 2001*, LNCS 1998, pp. 19-26.

[12] Masoud, O. "Tracking and Analysis of Articulated Motion with an Application to Human Motion," Doctoral Thesis, Univ. of Minnesota, March 2000.

[13] Masoud, O., Rogers, S., and Papanikolopoulos, N.P. "Monitoring Weaving Sections," ITS Institute Technical Report CTS 01-06, October 2001.

[14] Matusik, W., Buehler, C., Raskar, R., Gortler, S., and McMillan, L. "Image-Based Visual Hulls," *Proc. of ACM SIGGRAPH 2000*, pp. 369-374, July 2000.

[15] Ojanen, H., "Automatic Correction of Lens Distortion by Using Digital Image Processing," Rutgers University, Dept. of Mathematics technical report, July 1999.

[16] Oliver, N., Rosario, B., and Pentland, A. "A Bayesian Computer Vision System for Modeling Human Interactions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.22, No. 8, August 2000.

[17] Polana, R. and Nelson, R. "Nonparametric Recognition of Nonrigid Motion," Technical Report, University of Rochester, New York, 1994.

[18] Rosales, R. and Sclaroff, S. "3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, June 1999.

[19] Seitz, S.M. and Dyer, C.R., "View Morphing," *Proc. of ACM SIGGRAPH 1996*, pp. 21-30, 1996.

[20] Stauffer, C. and Grimson, W.E. "Learning Patterns of Activity Using Real-Time Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, No. 8, August 2000.

[21] Weik, S. and Liedtke, C.E. "Hierarchical 3D Pose Estimation for Articulated Human Body Models from a Sequence of Volume Data," *Intl. Workshop on Robot Vision 2001*, LNCS 1998, pp. 27-34.

[22] Wren, C.R. and Pentland, A.P. "Dynamic Models of Human Motion," *Proc. Third IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, April, 1998.

[23] Wren, C.R., Azarbayejani, A., Darrell, T., and Pentland, A.P. "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, No. 7, July 1997.