IEEE International Conference on Robotics and Automation 2003

Foveated Observation of Shape and Motion

James Davis
Honda Research Institute USA, Inc.
800 California St #300, Mountain View CA 94041
jedavis@ieee.org

Xing Chen
Stanford University
376 Gates Building, Stanford CA 94305
xcchen@graphics.stanford.edu

***Abstract* – Robotic navigation and interaction frequently require that the shape and motion of external objects and events be observed. Many interesting events occur at mixed scales. Subtle localized shape and motion often occurs together with long range movements. One of the chief challenges in recovering these events is to obtain high resolution imagery suitable for resolving small details, while simultaneously increasing the working volume in which recovery is possible.**

This paper proposes an architecture for mixed scale motion recovery. Robust coverage of a large working volume is provided by a wide area tracking system. This system localizes interesting motions, and guides a separate foveated system of pan-tilt cameras to observe the detailed event at high resolution. We demonstrate two applications, foveated structured light scanning and the capture of muscle deformation while walking. Both applications allow subtle detailed recovery that would not be possible using existing single scale systems.

I. INTRODUCTION

Robotic systems that interact with real environments need some method of recovering the shape and motion of external events that occur. For example a humanoid robot may need to locate actual humans, determine their location, track their motion, and finally watch for hand gestures in order to accomplish some collaborative task.

Existing systems that capture motion using identifiable markers are used widely in the entertainment and biomedical industries. In addition, current research efforts are bringing markerless capture of body motion, facial deformation, and hand gesture within reach. Since a robot that interacts with humans will need to observe "body language" and interpret facial expressions, it is likely that these advances will enable a large number of applications in robotic interaction with humans.

Despite recent advances, many applications are currently beyond reach. Among these are a class of applications that require the recovery of extremely detailed motion, relative to the size of the working volume. For example, in a soccer stadium, player motion is at a much smaller scale than the field itself; a person's facial expression is at a smaller scale than the person's motion in a room; and the gestures of a hand are at a smaller scale than the space within an arm's reach.

Due to the fundamental tradeoff in vision based sensing systems between imager resolution and field of view, existing motion recovery systems have primarily focused on single scale motions. In these existing systems the magnitude of a motion is at approximately the same scale as the volume in which the motion occurs. For instance, research on gesture recognition often requires that a user's hand be confined to a relatively small experimental working volume. Under real world conditions, the user's hand might of course be located anywhere in the room. One of the chief challenges to increasing the range of recoverable motion is to develop systems which can robustly capture motion at multiple scales.

In this paper we introduce a foveated sensing system which can capture detailed shape and motion that occurs imbedded in a much larger working area. We expect that this system will ultimately be integrated into a set of autonomous robots that can interact with the environment, however the immediate goal of this project has been to investigate the sensing technology itself. Thus a "3D Room" which can record the shape and motion of events that occur within the space has been designed. This Foveated sensing system can recover details that would be beyond the reach of more traditional single scale systems.

The system described in this paper uses a two level hierarchy to recover detailed shape and motion. A wide area tracking system is used to robustly localize motion in a large working volume. This information steers automated pan-tilt cameras which are zoomed onto the much smaller volume in which the relevant action occurs. By partitioning the goals of

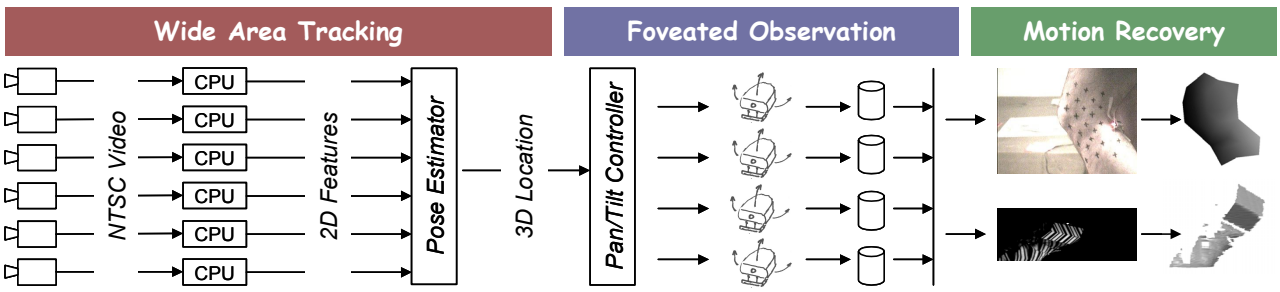

Figure 1. Overview of our mixed scale motion recovery system. A wide area tracking system robustly locates the target in a large working volume. The foveated observation system uses this information to direct high resolution imager towards the target. The high resolution video is streamed to disk, for later off-line analysis and motion recovery.

enlarging coverage area and obtaining high resolution, into disjoint subsystems, each can be optimized without making unnecessary tradeoffs. In addition to proposing an architecture for multiple scale motion recovery, we demonstrate two applications enabled by our system that would not be possible with existing single scale systems.

II. RELATED WORK

In existing single scale motion recovery systems, a tradeoff must be made between increasing coverage area and providing the maximum image resolution. Systems have been built at both ends of the design spectrum.

A number of researchers have attempted to capture actions which cover a wide area. Kanade's Virtualized Reality dome, and 3D Room both use many cameras to cover a space the size of a typical office laboratory [1] [2]. Similarly, commercial motion capture systems, such as those from Vicon and Motion Analysis, can track retro-reflective markers in a room size environment [3] [4]. At the far extreme, outdoor surveillance systems often cover extremely large areas [5]. However, all of these systems focus on capturing large scale motions that occur in large scale environments.

There has also been a great deal of research on capturing extremely detailed motion. For instance, Guenter et. al. capture the subtle deformations of a face that occur while speaking [6]. The speaker's head is required to remain centered in the field of view of several narrowly focused cameras. Other researchers have captured the motion of a hand in order to recognize gestures [7]. These systems are focused on the recovery of detailed motion given high resolution images, but do not address the capture of such images. So typically motion is restricted to a small area in order to obtain maximum resolution.

Simple foveated vision systems have been built. These systems typically pair a singlewide-angle view camera with a pan-tilt-zoom camera to provide high resolution images of a target. For instance, Greiffenhagen et. al. use a panoramic imager to locate a person, and a pan-tilt camera to capture the face [8]. Goodridge and Kay built a system which fuses audio measurements and video from a single wide angle camera to localize a speaker. A pan-tilt camera can then be oriented in the correct direction [9]. These systems image the environment from a single direction so can not robustly handle complex motion and occlusions.

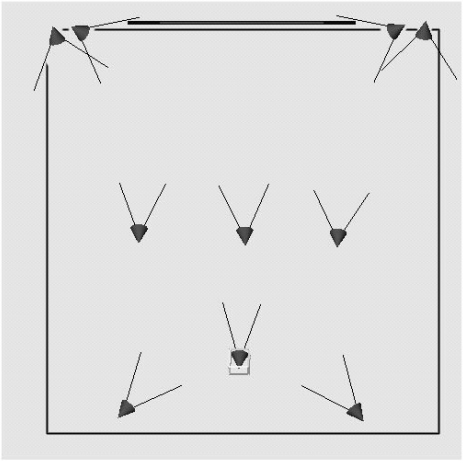

Figure 2: Placement of cameras in our wide area tracking system. These cameras can be placed to optimize coverage and robustness, without making tradeoffs to obtain good resolution.

We demonstrate a mixed scale motion capture system, which provides robust target localization over a wide scalable coverage area. Foveated high resolution cameras provide detailed imagery from several viewpoints, allowing recovery of subtle motion not previously possible.

III. SYSTEM ARCHITECTURE

Recovering motion in a large working volume requires that imagers have a relatively wide viewing angle, in order cover the entire volume. Unfortunately, this is a direct conflict with the need for a relatively narrow field of view in order to maximize the effective resolution of the imager. We address this challenge by noting that motion often occurs at mixed scales, the actual area of interest at any time is often a small subset of the total working volume.

Our system employs both a wide area tracking system to localize motion in a large volume, and a detailed capture system to recover high resolution motion. An overview of this architecture can be seen in Figure 1. The wide area tracking system can be optimized for coverage and robustness without making tradeoffs for high resolution. In addition, the detailed motion recovery system can obtain optimal imaging resolution without making tradeoffs for a larger working volume. Combining these components results in a system capable of capturing subtle motion over a wide area. In the following subsections we describe each subsystem in detail.

A. Wide area tracking

A desirable wide area tracking system should robustly handle complex motion and scalably cover a large area. We use a network of independent video cameras to observe the environment. These cameras are currently mounted on the walls of a room. The system described here has ten cameras which provide robust tracking of an area of 4 x 5 x 2 meters from many surrounding viewpoints. Additional cameras can easily be added to increase the coverage area. We calibrate the intrinsic parameters of each camera independently [10]. However, to ensure globally consistent tracking, the external pose of all cameras are optimized with respect to each other [11]. Figure 2 shows an overview of camera placement in our system.

In addition to physical coverage of space, a scalable tracking system needs to manage CPU load. By coupling a digitizer and local CPU with each camera, we ensure that computational bandwidth is sufficient to process each additional video stream. We have used both SGI Indy workstations and Wintel PCs with Matrox Meteor digitizing boards. The local CPU computes image feature coordinates from the high bandwidth video streams and communicates only the low bandwidth feature data via ethernet to a central estimator which integrates data from all observing cameras.

Each camera and associated CPU tracks the desired target in a subsection of the entire working volume. The object or feature parameters detected in each video stream are communicated to a

central estimator which integrates information from all cameras into a single estimate of object pose. A large body of research exists on tracking objects in a single video stream, many of which are applicable [12] [13]. While we have experimented with several more complex tracking methods, the emphasis of our wide area tracker is coverage and robustness. In order to ensure robustness, we use LED light sources to mark target objects, and a simple image intensity thresholding to locate targets in each video stream.

Features from each observing camera are integrated into a single estimate of object pose. Efficiency is important since these estimates will be used to guide the detailed tracking system to cover the appropriate subspace. Therefore, we choose to use an extended Kalman filter (EKF) to integrate observations because its efficient computation enables a high update rate [14]. The EKF estimates position and velocity of the target from the 2D observations of the mounted LEDs.

B. Foveated motion recovery

The goal of our detailed motion recovery sub-system is to provide high resolution coverage of the desired target. Traditional tracking systems must trade off resolution for coverage area, thus optimal motion recovery can not be obtained. In our system, the wide area tracking system guides the detailed tracking system to cover only the sub-volume in which the target currently resides. Since only the necessary subspace receives attention from the detailed tracker, optimal resolution and accuracy can be maintained.

The detailed tracker must selectively cover a sub-volume of the total working area. We use four imagers mounted on pan-tilt mechanisms to achieve this goal. The pan-tilt settings of these cameras can be operated via RS-232 serial control, so that any part of the working volume can be contained within the current viewing area. The Sony EVI-D30 cameras we use also allow automated control of zoom and focus. Figure 3 shows the location of these detailed cameras in our system.

Pan-tilt camera model: As the camera pans and tilts around its internal axes, the extrinsic pose of the camera changes. This motion must be calibrated in order to correctly steer the camera observation area to the desired target. As shown in Figure 4, we model the pan-tilt mechanism as a pair of arbitrary axes in space around which the camera rotates. Neither the axes nor the cameras nodal point are constrained to be coincident. Given a camera calibrated extrinsically in some known pan-tilt parameter configuration, any other configuration can be generated by rotating the known camera position and orientation around each of the axes.

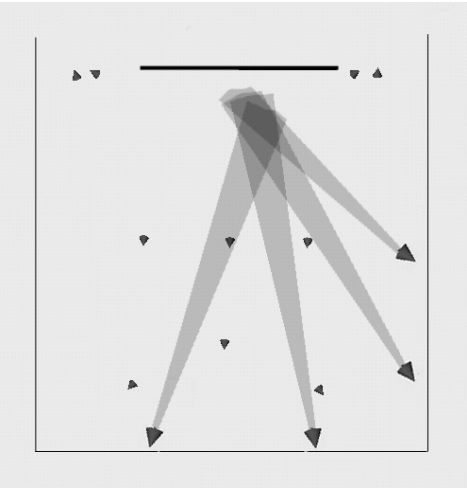

Figure 3. High resolution imagers in the detailed tracking sub-system are guided to constantly observe the target as it moves around the entire working volume.

A traditional static camera can be calibrated by placing a target with known features in the field of view. The 3D location of the features are registered in correspondence with their observed image coordinates. Given an ideal pin-hole camera model, each feature point satisfies

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = \mathbf{P} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

where (x, y, z) is a known feature location in world coordinates, and $(U/W, V/W)$ is the 2D image coordinates observed by the camera. By measuring many points, the parameters of the camera projection matrix, $\mathbf{P}$, can be determined. The camera projection matrix is typically parameterized by its position, orientation, and focal length. A pan-tilt camera is additionally parameterized by two axes. Thus, the transform from world coordinates to image plane can be written as

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = \mathbf{R}(\mathbf{A}_P, \theta_P) \cdot \mathbf{R}(\mathbf{A}_T, \theta_T) \cdot \mathbf{P} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

where $\mathbf{R}$ is a rotation matrix defined by an axis, $\mathbf{A}$, and rotation angle θ. An axis is parameterized by a 3D point and direction. Given a set of known 3D features, their corresponding image coordinates and the pan-tilt configuration (θ_P, θ_T) used to make each observation, the parameters of $\mathbf{P}$, $\mathbf{A_P}$, $\mathbf{A_T}$ can be solved jointly by setting up a set of non-linear equations.

Calibrating the camera model that we have defined is an additional challenge. Since the camera can pan and tilt to cover a very large working volume with very high precision 3D features that span the volume cannot be easily acquired using a single small target as is traditionally done when solving for camera extrinsic parameters. Instead, the wide area tracking system itself can be used to construct a large virtual calibration target [11]. The pan-tilt camera is rotated to observe the virtual calibration target from many pan-tilt configurations. The best fit axes and camera pose can now be found by solving the systems of non-linear equations shown above.

Camera control module: The detailed tracking system must be guided so that only the active area of the entire working volume is observed. This would be difficult to achieve if the detailed tracking system was operating independently and tracking the target by itself. The foveated cameras would have to exhaustively search the entire working volume to identify the target of interest, and then repeat this process whenever tracking ambiguities, occlusions, or unexpected acceleration caused the target object to be lost. Instead we use the wide area tracking system with multiple overlapping

cameras to robustly localize the area of interest and then simply direct the foveated system to make observations in the right direction.

Given a point of interest by the wide area localization, we can find the pan-tilt setting that causes the foveated cameras to observe the correct region. As described previously, we can determine the optical viewing axis given by a base camera calibration and pan-tilt setting. To find the correct setting we can merely search for the optical viewing axis which intersects the desired point in space. The distance between the line given by an optical axis and the desired point can be minimized to find the correct setting. We use a simple gradient descent method and are able to find correct settings in only a few iterations.

Another challenge to real-time foveated tracking is the inevitable latency in the pan-tilt mechanism and in the wide area localization. A target moving with a high velocity is likely to have left the expected observation region by the time that the camera actually observes that region. Because the EKF in the robust wide-area localization sub-system can estimate the velocity as well as the position of a target in space, we can predict the location of the target in the near future and direct the camera toward the predicted location to compensate for latency. We have found that without prediction the target often moves outside the detailed observation region. By using prediction the target stays within the observation region except under extreme acceleration. Empirically, the pan-tilt mechanism of our cameras seems to require about 300ms to reach a target configuration, thus we set the latency compensation such that position is predicted 300ms into the future.

IV. RESULTS

A number of interesting applications can be enabled by a mixed scale motion recovery system. We describe two applications that we have investigated in our lab which are made possible by the system described. The first recovers the shape of a target as it moves about in the space, the second recovers motion that would be too detailed for a single scale system to capture.

C. *Foveated structured light*

Structured light systems recover depth images by projecting known patterns into a scene and analyzing the patterns as observed by a camera [15, 16]. Depth images are useful in a number of domains, including image segmentation, gesture recognition, and model recovery. Spatial resolution of these capture systems is limited by both the projector and imager. While modern projectors can easily display over 1000 stripes, typical NTSC video cameras can only resolve about 250 of these lines uniquely. Spatial resolution is limited by the camera, rather than the projector resolution.

To overcome this limitation, mixed scale recovery can be used. Foveated cameras allow selective attention to different regions in the working volume. In this way spatial resolution is limited only by the projector.

Figure 5 shows an overview of an actual capture session using our foveated architecture. A projector displays a sequence of structured light patterns that cover an area about 1 x 2 x 2 meters. The field of view of the pan-tilt camera is set very narrowly so that it covers only the teapot. The camera stays oriented towards the teapot as it moves through the large working volume. The foveated camera captures a continuous stream of images, which can be used to reconstruct the depth of the target. Figure 6 shows several images taken from a video stream recorded as described. The projected stripe pattern is clearly visible on the moving block shaped target. Below each video frame is a rendering of the polygonized arm and block depth values that were recovered. Each recovered mesh is shown both from the front and side. The resolution with which this shape has been recovered is much higher than would be possible if the entire working volume had been covered equally.

Traditional structured light systems require multi-frame stripe patterns that are unsuitable for moving objects. Our system uses a method that correctly accounts for moving objects [17]. In the example shown here both the camera and projector are functioning at lower than optimal resolution. The mixed scale system

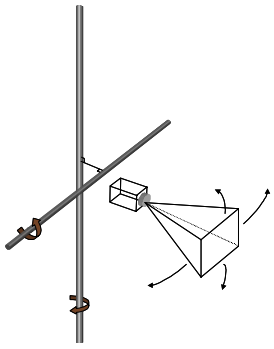

Figure 4. The camera's pan-tilt mechanism is modeled as a pair of arbitrary axes in space. Changes in the camera's pan-tilt parameters cause the camera's position and orientation to rotate around these axes.

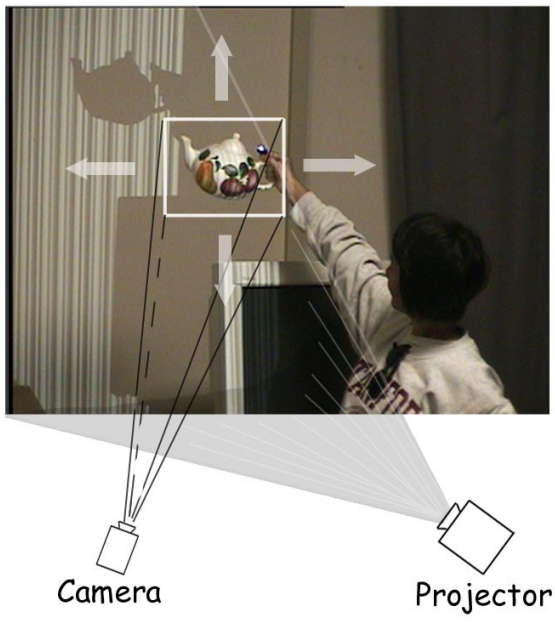

Figure 5 In this foveated structured light depth imager, the projected light covers the entire working volume. The detailed observation system observes only a small region of the entire volume, providing higher resolution depth estimates.

we have demonstrated would work equally well with devices functioning at their maximum resolution. In the future, we would also like to mount the projector on a pan-tilt head, to eliminate this constraint on resolution as well.

D. Muscle deformation recovery

Recovering non-rigid, time-varying shape deformation is a particularly challenging task. High resolution imagery is essential in order to resolve and track features over multiple frames. Many applications require a wide area of coverage as well.

One application is the study of tension and abnormal pressure on damaged human joints, on which we are collaborating with a biomechanics research group. By measuring exact skeletal configuration and deformation of tissue surrounding the knee, joint torques can be computed. These computations will ultimately help understand the mechanics of motion sufficiently that robotic medical devices can be built. These measurements must be taken while the subject is actually walking, and commercial motion capture systems do not have sufficient resolution to measure deformation while simultaneously covering a wide area.

We have been able to make preliminary measurements using our mixed scale motion recovery system. A subject's knee is marked with a number of features. The subject walks normally in a large working volume and the system goal is the recovery of the time varying shape of these features.. However, at the scale of the entire working volume, individual features are not resolvable with enough resolution to recover geometry.

Figure 7 shows several frames captured during a walk sequence from our foveated system. The view from each of two zoomed pan-tilt cameras is shown. In these high-resolution views, features are clearly visible. As the person walks, the wide area tracking system guides all pan-tilt cameras so that the knee stays in each video frame. These video streams can be post processed to recover shape and motion.

Our preliminary experiments uses a hierarchical Lucas-Kanade tracker to track the 2D features [12]. After obtaining 2D feature traces, the 3D geometry is recovered by triangulating features between two views. The resulting polygonized model is shown in the bottom row of Figure 7. As a visual aid, the rendered model has also been shown overlain on the video frames captured from camera 2.

V. CONCLUSION

Many real world robotic navigation and interaction tasks occur in situations where the working environment is at a much larger scale than the motions that need to be observed. In these cases neither wide area systems nor narrowly focused high resolution sensors are adequate individually. By combining a wide area localization system with a high resolution imaging configuration we can achieve both goals. The wide area system can be optimized for robustness and coverage, while the detailed imaging system can be optimized for resolution. Since these systems can be optimized separately, the tradeoff between resolution and size of working volume which must be made in previous systems is avoided.

We have built a mixed scale motion recovery system and demonstrated its application in two different scenarios, neither of which would have been practical using exiting single scale methods.

Using a projected structured light method, we recover the shape of a target object. By using a mixed scale system, we are able to allow target motion in a wide area and simultaneously image the target at a high resolution.

We were also able to capture zoomed-in video of a walking subject's knee using our mixed scale capture system from multiple viewpoints. Because the subject must be walking, a large capture area is required. Despite the large working volume, high resolution video was obtained. This video allowed us to construct a time varying deformation model of the tissue surrounding the knee joint.

In conclusion, by partitioning motion recovery tasks into multiple scales and optimizing these subsystems separately, we can enable recovery applications in situations not previously possible.

VI. ACKNOWLEDGEMENTS

This work was partially supported by grants from Intel Corporation, Interval Research, and Sony Corporation. Szymon Rusinkiewicz kindly provided the structured light system, for which we are grateful. In addition, we would like to thank Pat Hanrahan and other members of the Stanford Graphics Lab for helpful comments and discussion.

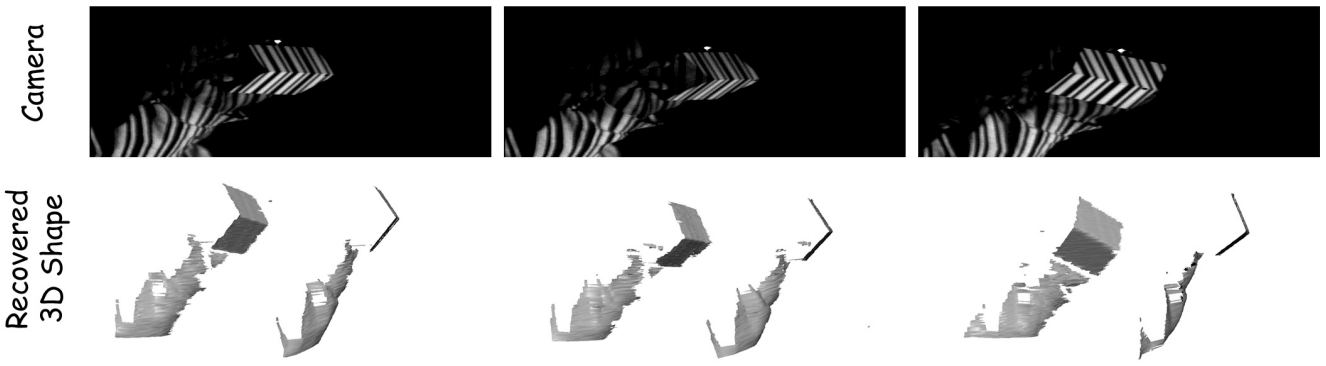

Figure 6: Top: Several frames from the captured video of an object being moved in the volume of structured light. Bottom: The recovered depth map rendered from two viewpoints.

VII. REFERENCES

[1] P. Rander, "A Multi-Camera Method for 3D Digitization of Dynamic, Real-World Events," PhD Dissertation, Robotics Institute, Carnegie Mellon University, 1998.

[2] H. Saito, S. Baba, M. Kimura, S. Vedula, and T. Kanade, "Appearance-based virtual view generation of temporally-varying events from multi-camera images in the 3D room," in *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*. Los Alamitos, CA, USA: IEEE Comput. Soc, 1999, pp. xi+546.

[3] Vicon, http://www.vicon.com.

[4] MotionAnalysis, http://www.motionanalysis.com.

[5] G. P. Stein, "Tracking from Multiple View Points: Self-calibration of Space and Time," Proceedings of *IEEE Computer Vision and Pattern Recognition*, pp. I:521-527, 1999.

[6] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin, "Making faces [facial animation]," in *Computer Graphics. SIGGRAPH 98 Conference Proceedings*. New York, NY, USA: Acm, 1998, pp. 472.

[7] R. Grzeszczuk, G. Bradski, M. Chu, and J.-Y. Bouguet, "Stereo Based Gesture Recognition Invariant to 3D Pose and Lighting," presented at Proceedings of IEEE Computer Vision and Pattern Recognition, 2000.

[8] M. Greiffenhagen, V. Ramesh, D. Comaniciu, and N. Heinrich, "Statistical Modeling and Performance Characterization of a Real-Time Dual Camera Surveillance System," Proceedings of *IEEE Computer Vision and Pattern Recognition*, 2000.

[9] S. G. Goodridge and M. G. Kay, "Multimedia sensor fusion for intelligent camera control," in *1996 IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems (Cat. No.96TH8242)*. New York, NY, USA: Ieee, 1996, pp. xv+848.

[10] J. Heikkila and O. Silven, "A Four-Step Camera Calibration Procedure with Implicit Image Correction," *IEEE Computer Vision and Pattern Recognition*, pp. 1106-1112, 1997.

[11] X. Chen and J. Davis, "Wide Area Camera Calibration Using Virtual Calibration Objects," *IEEE Computer Vision and Pattern Recognition*, 2000.

[12] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, 1981.

[13] O. Faugeras, *Three-Dimensional Computer Vision, A Geometric Viewpoint*: MIT Press, 1993.

[14] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering. Second Edition*, 1992.

[15] J. L. Posdamer and M. D. Altschuler, "Surface measurement by space-encoded projected beam systems," *Computer Graphics and Image Processing*, vol. 18, pp. 1-17, 1982.

[16] K. Sato and S. Inokuchi, "Range-imaging system utilizing nematic liquid crystal mask," in *Proceedings of the First International Conference on Computer Vision (Cat. No.87CH2465-3)*. Washington, DC, USA: IEEE Comput. Soc. Press, 1987, pp. xii+734.

[17] O. Hall-Holt and S. Rusinkiewicz, "Stripe Boundary Codes for Real-Time Structured-Light Range Scanning of Moving Objects," *IEEE ICCV*, 2001.

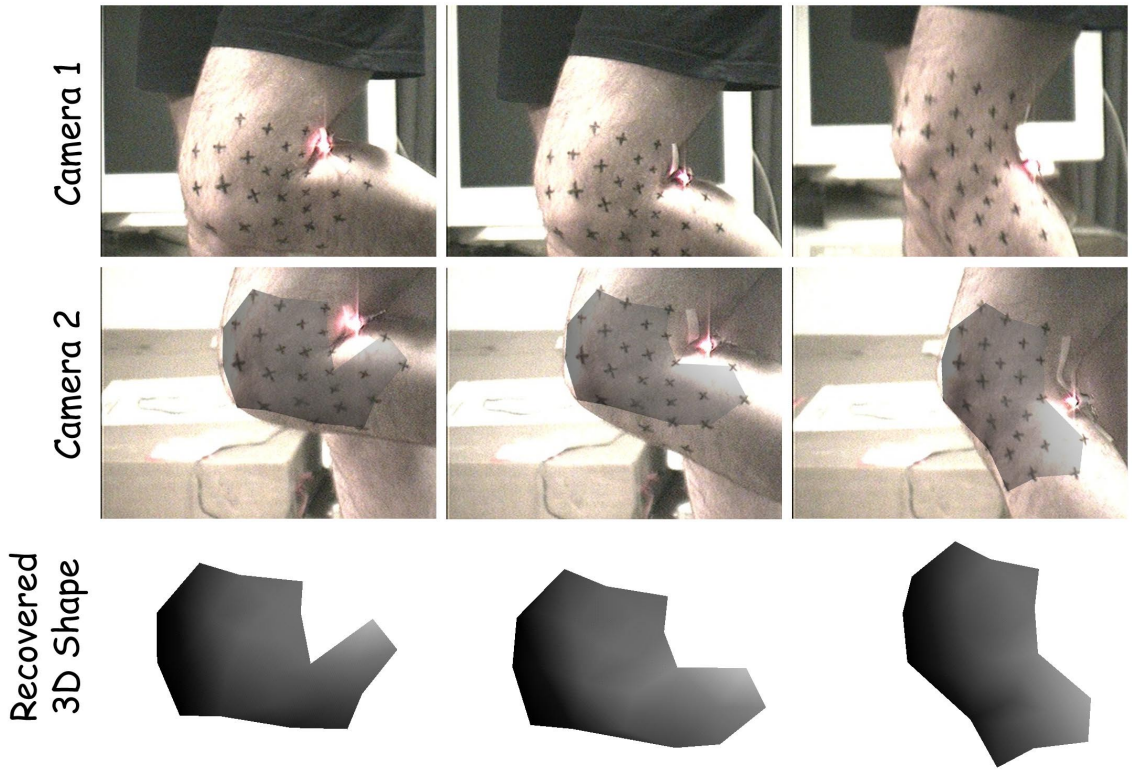

Figure 7. High resolution video of the subjects knee joint is captured from several views by our foveated motion capture system. The top two rows show synchronized video frames from two different viewpoints. The bottom row shows geometry recovered by triangulation. As a visual aid, the rendered geometry is shown superimposed on the video from camera 2.