I²MTC 2008 – IEEE International Instrumentation and
Measurement Technology Conference
Victoria, Vancouver Island, Canada, May 12–15, 2008

Volumetric Modeling with Multiple Cameras for Markerless Motion Capture in Complex Scenes

Silvain Bériault, Martin Côté, Pierre Payeur
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5
[sberi081, mcote, ppayeur]@site.uottawa.ca

Abstract *– Recently, significant advances have been made in many sub-areas regarding the problem of markerless human motion capture. However, current markerless systems tend to introduce major simplifications, especially in early stages of the process, that temper the robustness and the generality of any subsequent modules and, consequently, of the whole application. This paper concentrates on improving the aspects of multi-camera system design, multi-camera calibration and shape-from-silhouette reconstruction. A thoughtful system analysis is first proposed with the objective of achieving an optimal synchronized multi-camera system. This multi-camera system is then accurately calibrated using a flexible method which allows free camera positioning. A novel region-based silhouette extraction procedure is proposed to remove the requirement of static and highly contrasting backgrounds. The outcome of our work is the achievement of robust voxel data reconstruction and coloring in complex and unconstrained scenes.*

Keywords *– motion capture, multi-camera calibration, JSEG segmentation, shape-from-silhouette, voxel coloring.*

I. INTRODUCTION

Computer-based monitoring of human activities is of great interest in a wide variety of applications. Motion capture is heavily used for animated television shows, in cinema industry, and for advanced realistic 3D video games. In biomedical engineering, human motion capture finds applications in computerized gait analysis for rehabilitation and prevention of injuries. Prompt recognition of human activities can also be applied to the field of surveillance and security, for example, to detect and locate hostile behaviors. In this research project, we are particularly interested in monitoring human motion, using passive vision, for piano-playing performance evaluation as well as prevention of injuries. This particular application poses several challenges and problems: complexity of the background and of the human postures, limited extent of movement, non-empty scenes, self-occlusion, etc.

Systems for motion capture can be distinguished in two main categories: marker-based and markerless systems. Marker-based systems, like the Vicon system [1], are characterized by the fact that the performers must wear multiple markers in order to capture their various movements. Marker-based solutions are typically more robust and almost always preferred to markerless solutions because they support complex and rapidly varying human postures. However, marker-based solutions also admit several drawbacks. They typically require very specialized and high cost equipments, and lengthy setup time in installing the markers. Furthermore, wearing such markers can be cumbersome, uncomfortable, and can interfere with natural motion of the performers. Markerless solutions attempt to remove those constraints by using solely passive vision for gesture monitoring. The current trend in markerless motion capture is to merge the content of multiple views of a performer into a consistent 3D model, which is ultimately used to estimate and analyze the human posture at every instant in time. In recent applications, voxel data, generated by the intersection of multiple silhouette images of the human body, is almost always used because this volumetric representation, due to its Cartesian nature, automatically lends itself to post-processing analysis such as the human posture extraction.

Early attempts, by Mikic *et al.* [2] and Cheung *et al.* [3] resulted in systems that can track the human body using binary voxel data with acceptable robustness. However several constraints are imposed to the system. Performers are required either to wear tight clothing of a distinct color [2] or to evolve in a completely static scene [3]. In both cases, only a subset of simple and non-occlusive postures is supported. In a later iteration [4], Cheung *et al.* proposed to incorporate color information to volumetric models at the expense of removing any real-time aspirations. Color information provides a complementary cue to resolve ambiguities occurring in complex human postures. This idea has been further investigated in the work of Kehl *et al.* [5]. In their system, a fast voxel coloring scheme is proposed with very satisfying results. Very recently, Caillette [6] proposed a real-time motion capture implementation also based on colored voxels. To achieve a real-time system, many tradeoffs and simplifications had to be made, regarding the complexity of the multi-camera setup and of the applied algorithms, making this work unsuitable for real-world applications.

From this analysis many issues remain unresolved and prevent markerless solutions to make marker-based systems obsolete, from a commercial standpoint. This paper concentrates on resolving issues related to multi-camera system design and its calibration, silhouette extraction and volumetric reconstruction in minimally constrained environments. A synchronized multi-camera system architecture for human motion capture is developed in section II. Section III overviews a flexible multi-camera calibration

approach, which allows free camera positioning. Section IV presents a novel technique for human silhouette extraction which removes the requirement of using a static and highly contrasting background. In section V, the camera calibration data and human silhouette data from all views are combined to obtain a colored voxel model of the performer. Results of the calibration, silhouette extraction and volumetric reconstruction procedures are presented in the final section.

II. MULTI-CAMERA SYSTEM ARCHITECTURE

The system design aspect of markerless motion capture applications is often overlooked. However, inadequate selection of camera equipment or improper system design can temper the spatial and temporal quality of the input video data and thus impact every subsequent module which deals with the actual gesture analysis. For this reason, multi-camera system design is the first topic addressed in this paper.

A. Hardware Camera Setup

Our acquisition system is shown in Fig. 1. It is composed of 3 Pentium IV 3.40 GHz computers and 8 Point Grey Research® Flea2 IEEE1394b Firewire cameras. All cameras are mounted to a reconfigurable structure. This structure allows free positioning of cameras all around the workspace. The structure itself can be enlarged to accommodate various sizes of working volume. The camera setup used to monitor the gesture of pianist musicians occupies a volume of approximately 2.5 m x 2.5 m x 2.5 m.

The decision of using Flea2 cameras is motivated by multiple factors. To monitor human activities with high precision, the use of global shutter exposure is clearly a predominant requirement because it allows all pixels to be measured simultaneously in contrast with a rolling shutter where the pixels are measured sequentially line by line. The Flea2 cameras also allow multiple mechanisms for multi-camera frame synchronization, which is essential especially when cameras are distributed across multiple computers. These cameras can operate at a high frame rate (60 fps) and therefore provides high flexibility in adjusting the temporal resolution to match the speed of motion to be captured. The frame resolution is limited to 640x480 pixels but it is sufficient for the purpose of volumetric reconstruction. Indeed, in current applications, a resolution of 320x240 is often privileged because it allows faster silhouette extraction and also because of inherent memory limitations in the resolution of voxel models, making irrelevant the use of higher image resolutions. Furthermore, cameras possess internal color calibration functionalities and several pre-processing functionalities to enhance the quality of the acquired video. Finally, the IEEE1394b bus speed allows for multiple cameras to be connected to a single acquisition node at high frame rate thus helping reduce the global system cost.

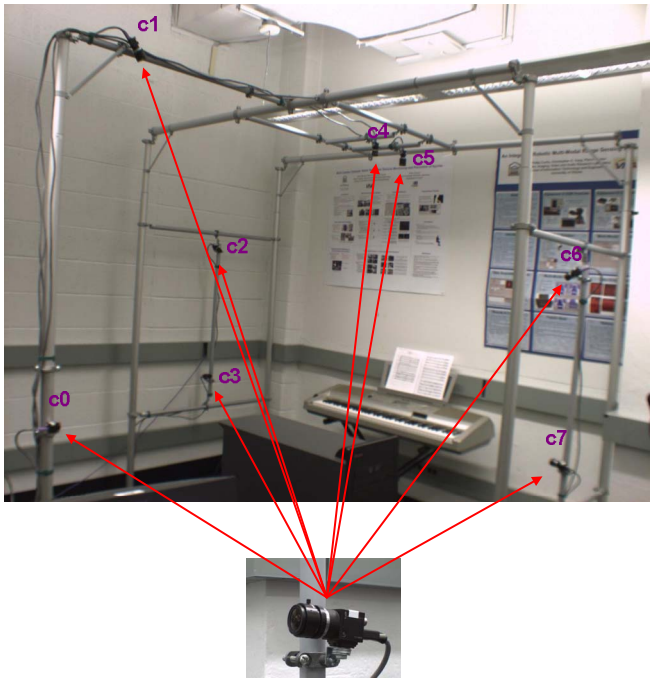

Fig. 1. Setup used for multi-camera video acquisition.

B. System Architecture

The cameras are distributed over multiple acquisition nodes according to the diagram of Fig. 2. The main purpose of the acquisition nodes is to receive and store frames of video from multiple cameras. All acquisition nodes are daisy-chained by an IEEE1394a link for inter-camera synchronization. A main computer serves the purpose of merging the content of all synchronized video streams featuring the performer. In off-line applications, where the motion capture is performed only once the video acquisition is completed, it is not necessary to use a separate computer for this task.

C. Inter-Camera Synchronization

Inter-camera synchronization serves the same purpose as the global shutter exposure featured independently in every camera. Global shutter exposure ensures that all pixels in a frame are measured at the exact same time. This allows all pixels in a frame to be spatially synchronized. Analogically, inter-camera synchronization ensures that concurrent video frames are exposed simultaneously across all cameras. All cameras in the network are synchronized with high precision using the Point Grey Research® MultiSync software. For all acquired frames of video in each view, timestamps are extracted and saved along with the image data. The main computer has the special task of comparing timestamps from all received images in order to regroup and process together frames that occurred simultaneously.

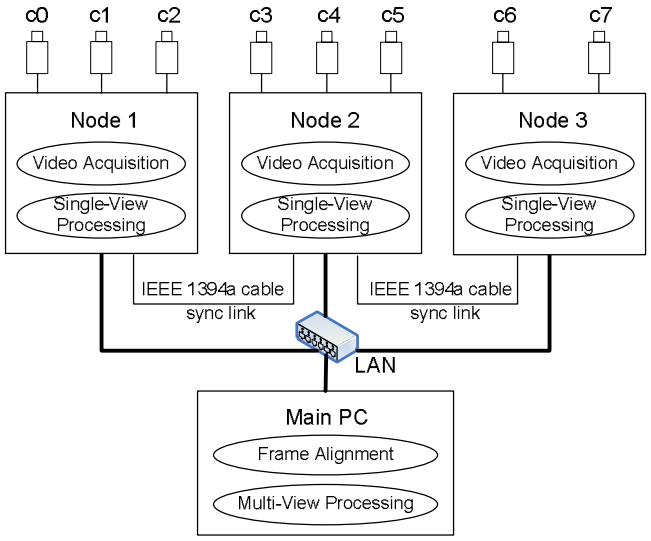

Fig. 2. System architecture for multi-computer synchronized video acquisition.

III. MULTI-CAMERA CALIBRATION

As a premise to multi-view reconstruction, the multi-camera system, previously presented, needs to be calibrated in order to align all cameras to a common global coordinate system. It is well established that multi-camera calibration is a complex problem since many requirements need to be satisfied. In particular, classical approaches that use complex calibration targets are not scalable, require an empty working volume and do not allow free camera positioning.

A. Camera Calibration Scheme

The proposed calibration procedure [7] intends to answer the major flexibility issues associated with classical approaches and, in particular, allows for very large baselines and major orientation changes between camera poses. The procedure is executed in two stages. In the first stage, the intrinsic parameters and lens distortion coefficients are computed, only once, for each camera independently using the classical multi-frame approach of Zhang [8]. Then cameras are positioned in their final configuration. The second stage is the core of the procedure and consists of extrinsically registering all cameras to a global coordinate system using a single visible marker waved over the entire working volume.

B. Framework for Extrinsic Camera Calibration

The proposed framework counts on seven major steps to achieve complete and accurate estimation of the extrinsic parameters. 1) The first step consists of creating a virtual 3D calibration object by waving a small visible marker, obtained with a light-emitting diode, over the full working volume to collect image matches across the entire camera network. 2) Those matches are regrouped by pair of cameras and the fundamental matrix is computed for each pair that contains enough matches. A robust RANSAC implementation [9] is used to eliminate outliers. 3) Each pair-wise fundamental matrix is then decomposed to extract stereo relations up to a scale factor using the method described by Hartley and Zisserman [10]. 4) Pair-related scale factors are solved incrementally by intersecting translation vectors in the 3D space such that a consistent camera network is defined up to a global scale factor. Links are scaled in a preferential order, using a weighted graph analysis, to minimize the accumulation and propagation of errors in links located far from the reference camera. 5) Upon completion of the weighted camera graph, all cameras are unified to a common global coordinate system, therefore resulting in an initial estimate of the extrinsic parameters for all cameras. This is done by linking all cameras to the reference camera using the rule of the shortest path. 6) This estimate is then optimized using a bundle adjustment [11] which is implemented in conjunction with the framework proposed by Lourakis *et al.* [12]. 7) The camera network can optionally be rescaled to absolute dimensions because it can be advantageous, from a human operator perspective, to represent the final calibration with meaningful (i.e. metric) units. This scale factor can be estimated coarsely by measuring the baseline between any camera and the reference camera or more precisely by using a dual-marker calibration target instead of a single marker one.

IV. SILHOUETTE EXTRACTION

It has been established previously that current markerless solutions to the problem of human motion capture impose unreasonable constraints and assumptions in order to operate on individuals. These limitations typically consist of static, highly contrasting backgrounds or assumptions on the degree of motion exhibited by the performer. Most current systems utilize a background subtraction technique to extract human silhouettes. However, image segmentation in unconstrained environments is complex and remains an open problem [13]. This section overviews a novel approach, developed within this project, to resolve the constraint of static backgrounds using a region-based segmentation methodology. The reader is referred to the original paper [14] for implementation details. The proposed technique uses color-texture information to produce homogenous regions within a set of frames that are then tracked throughout the sequence. The technique is based on Deng and Manjunath's JSEG implementation [15] with key improvements making it more appropriate to the context of human beings performance evaluation. The algorithm is structured as a set of five key processes: clustering, soft-classification, *J*-Value segmentation, merging and tracking.

1) Clustering: As a precursor to the actual segmentation, the video first undergoes a clustering process. Originally proposed by Deng *et al.* [15] was a *k*-means based approach which assumes that the colors present within a scene follow Gaussian-like statistics. This hypothesis cannot always be

guaranteed for complex scenes. Wang *et al.* [16] also reached this conclusion and modified the approach to use a nonparametric clustering technique called the Fast Adaptive Mean-Shift (FAMS) introduced by Georgescu *et al.* [17]. It is used within our approach to cluster color distributions within a video without applying assumptions to their distributions.

2) Soft-classification: In a list of improvements to the original JSEG algorithm, Wang *et al.* [16] introduced the concept of soft-classification maps. These maps represent a measured membership value that a pixel has to its assigned cluster. These values allow the JSEG algorithm to soften the color-texture edges between two similar cluster distributions. The classification maps are created using normalized 3D histograms of pixel intensities [14]. This nonparametric representation of the clusters allows for better results in the segmentation process.

3) J-Value segmentation: The JSEG process, introduced by Deng *et al.* [15], allows images to be segmented into regions based on a color-texture homogeneity criterion. This criterion, called the *J*-Value, is computed for every pixel in an image and is based on a neighborhood cluster distribution. Once all *J*-Values are computed the result is a gradient image representing the edges of color-texture areas. Using a seed growing algorithm these areas are labeled. These regions can be refined by iteratively re-computing *J*-Values using a smaller set of neighborhood pixels. As a start-point, key regions representing the subject of interest are identified from the first frame by a human operator.

4) Region merging: The JSEG algorithm suffers from a problem of over-segmentation. This issue, originally addressed using a color merging process [15], is solved here by incorporating Hernandez *et al.*'s joint-criteria merging algorithm [18]. The latter allows both edge and color information to be taken into consideration in order to produce better defined regions.

5) Region tracking over time: The tracking algorithm is based on a hybrid approach [14] that cuts the video into blocks. Regions within a block are tracked based on Deng *et al.*'s J_t-Value computation for temporal seed correspondence [15]. Regions between blocks are tracked using Withers *et al.*'s research on cell tracking [19] using overlap-ratios.

V. VOLUMETRIC RECONSTRUCTION

A premise to most recent vision-based motion capture applications is the computation of a volumetric model of the targeted performer. Nowadays, voxel data often serves as an intermediate representation in estimating the actual posture of the human body. A voxel-based model can be obtained by intersecting silhouette images from multiple views as shown in the diagram of Fig. 3. Synchronized frames of color video are acquired using the system developed in Section II. Silhouettes are then extracted and back-projected in the 3D space using the camera calibration data computed in Section III and the segmentation technique described in Section IV.

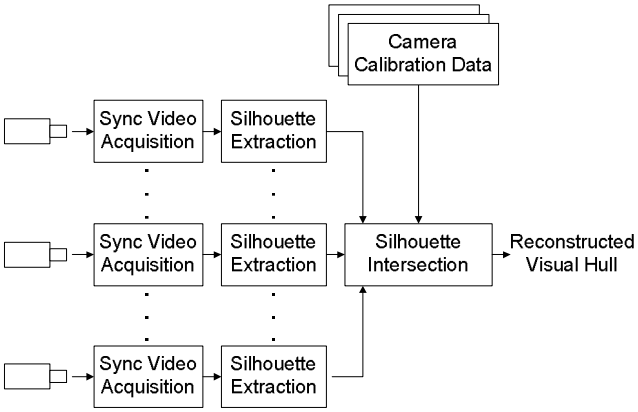

Fig. 3. High-level view of shape-from-silhouette reconstruction process.

A. Voxel Occupancy Evaluation

Constructing a voxel model fundamentally consists of evaluating the occupancy (foreground or background) of each voxel independently. This is done by computing the volumetric intersection of multiple silhouettes back-projected in the 3D space. In practice, a voxel is labeled as foreground only if the voxel projects in a region that pertains to the targeted performer in every image view. Consequently, a voxel is labeled as background if it falls in a background region in at least one image view. To account for possible imperfections in the silhouette extraction process, this criterion can be relaxed to two views to the expense of possibly resulting in a coarser volumetric representation as some background voxels could be misclassified as foreground.

B. Voxel Coloring

In addition to the voxel occupancy information, a color is assigned to each surface voxel based on the color information provided by the original video streams and can serve as a supplementary cue to disambiguate complex postures. Instead of using highly photorealistic, but very computationally demanding, texturing approaches that rely on multiple plane-sweeps [20], our approach is inspired from Kehl *et al.*'s method [5] which presents a good compromise between efficiency and texturing quality. Our algorithm proposes the use of depth buffers (one per camera) to detect cases of voxel occlusions. Depth buffers maintain the Euclidian distance to the closest foreground voxel, for every pixel in every view. In the original approach of Kehl *et al.* [5], the coloring was based only on the point projection of voxel centers, rather than the surface projection of full voxels, for the sake of speed. This approximation leads to an insufficient condition for occlusion detection, especially when the voxel resolution is unbalanced with respect to the working image resolution.

Once every surface voxel has been visited, the average color among all views that can see a voxel without occlusion is used as the final color for that particular voxel. For special

cases where all views are disqualified, the minimal depth distance separating this voxel from the occluding voxel is used to pick the best view. Finally, as Caillette [6] observed, it can be relevant, for the purpose of human posture estimation, to propagate voxel color to interior voxels. Our implementation allows color to be propagated to interior voxels simply by re-applying this scheme, layer-by-layer, to every interior voxel.

VI. RESULTS

A. Multi-Camera Calibration

The calibration procedure of section III results in the accurate estimation of the global camera structure with an average reprojection error of less than ½ pixel even when using highly distorting wide angle lenses. Fig. 4 shows a virtual model of the camera positioning after the calibration of the physical setup displayed in Fig. 1. The bottom-left camera was chosen to be the reference camera (in red). The black dots represent the position of all calibration points forming the virtual calibration object generated by waving the visual marker. They are displayed to demonstrate that uniform coverage of the working volume is obtained.

B. Silhouette Extraction

Results of our region-based silhouette extraction procedure are shown in Fig. 5 for four camera views. The first frame of video is segmented in many regions based on similarities between neighboring pixels, without any prior knowledge of the background. Particular regions of interests are then manually chosen, but the tracking is automatic afterwards. In comparison with work from other motion capture implementations reviewed in this paper, this new scheme does not compromise the overall perceptual quality of extracted silhouettes. To the contrary, segmentation results are improved since region-based segmentation does not suffer from the problem of shadows inherent to most background subtraction methods. In particular, our results show that shadowed pixels over the keyboard are not mistakenly incorporated to the foreground silhouettes, therefore yielding better extraction and reconstruction of the pianist hands.

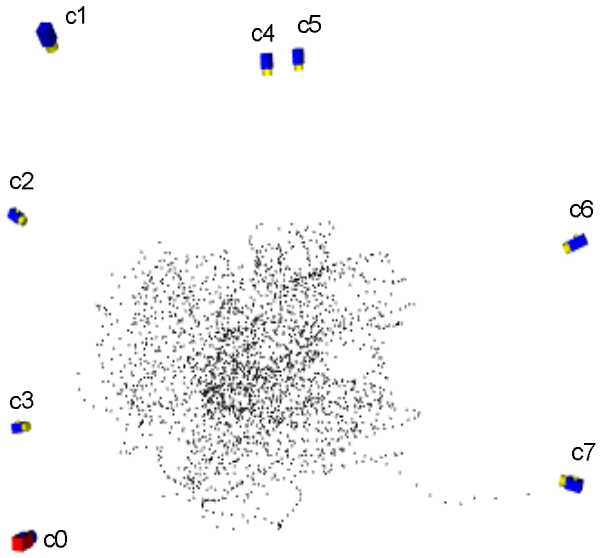

Fig. 4. A virtual model representing the physical camera structure of Fig. 1.

C. Volumetric Reconstruction

Volumetric reconstruction results are shown in Fig. 6. The voxel occupancy is computed using the binary silhouette data with satisfying precision for human posture estimation. The color content of the original videos is then incorporated to the voxel data with competitive accuracy to Kehl *et al.*'s work [5] and especially in presence of complex self-occluding postures such as the pianist posture.

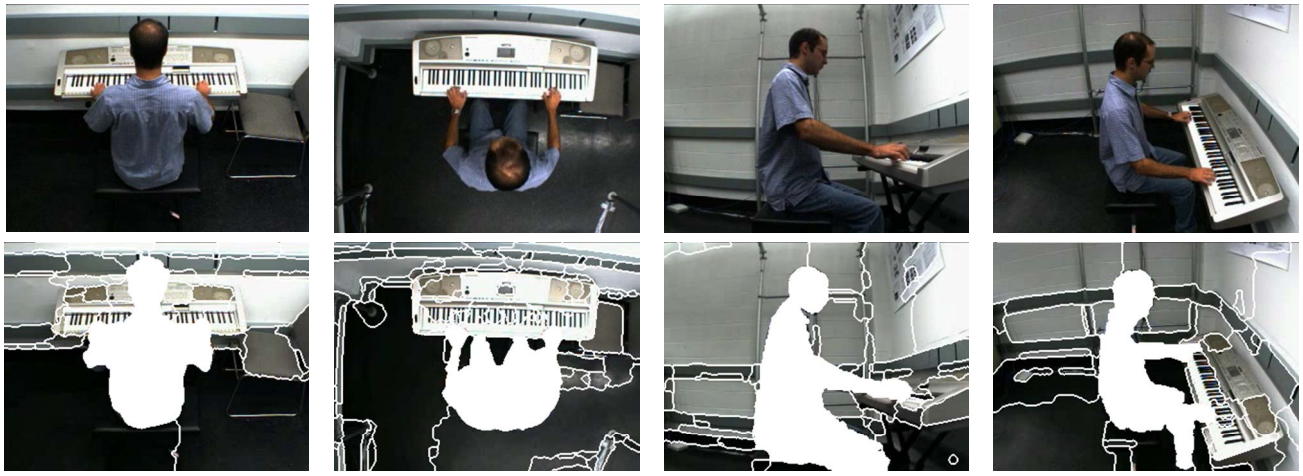

Fig. 5. Synchronized color and segmented frames for four camera views.

Fig. 6. Four views of a reconstructed colored voxel model of a pianist.

VII. CONCLUSIONS

Markerless human motion capture in unconstrained environment remains a challenging computer vision problem. Our implementation relies on the use of a synchronized and fully calibrated multi-camera system. Our solution innovates, with respect to existing systems, by the use of an enhanced region-based silhouette extraction scheme which removes the constraint of static and highly contrasting background. Extracted silhouettes are also combined into a consistent voxel model with coherent texturing. Future work will concentrate on using 3D depth information, as a supplementary cue to color and edge information, to disambiguate special cases in the silhouette extraction procedure, and on high-level analysis of voxel data to extract human kinematics information for gesture quantification and analysis.

ACKNOWLEDGMENTS

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Vicon Motion Capture System, 2007, www.vicon.com.

[2] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Human Body Model Acquisition and Tracking Using Voxel Data", *International Journal of Computer Vision*, vol. 53, no. 3, pp. 199-223, 2003.

[3] G. Cheung, T. Kanade, J. Bouguet, and M. Holler, "A Real Time System for Robust 3D Voxel Reconstruction of Human Motions", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 714-720, June 2000.

[4] G. Cheung, S. Baker, and T. Kanade, "Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 77-84, Madison, June 2003.

[5] R. Kehl, M. Bray, and L. Van Gool, "Full Body Tracking from Multiple Views Using Stochastic Sampling", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 129-136, 2005.

[6] F. Caillette, *Real-Time Markerless 3-D Human Body Tracking*, Ph.D Thesis, University of Manchester, 2006.

[7] S. Bériault, P. Payeur, and G. Comeau, "Flexible Multi-Camera Network Calibration for Human Gesture Monitoring", *Proc. of the IEEE International Workshop on Robotic and Sensors Environments*, pp. 125-130, Ottawa, ON, Oct. 2007.

[8] Z. Zhang, "A Flexible New Technique for Camera Calibration", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, November 2000.

[9] M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with applications to Image Analysis and Automated Cartography", *Communications of the ACM*, vol. 24, no.6, June 1981.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2000.

[11] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment: A Modern Synthesis", Vision Algorithms: Theory and Practice, *LNCS*, vol. 1883, pp. 298-372, Springer-Verlag, 2000.

[12] M. Lourakis and A. Argyros, "The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm", Institute of Computer Science, Forth, Technical Report 340, August 2004.

[13] M. Côté, P. Payeur, and G. Comeau, "Comparative Study of Adaptive Image Segmentation Techniques for Gesture Analysis in Unconstrained Environments", *Proc. of the IEEE International Workshop on Imaging Systems and Techniques*, pp. 28-33, Minori, Italy, 2006.

[14] M. Côté, P. Payeur, and G. Comeau, "Video Segmentation for Markerless Motion Capture in Unconstrained Environments", International Symposium on Visual Computing, *LNCS*, vol. 4842, pp. 791-800, Springer-Verlag, Nov. 2007.

[15] Y. Deng and B. S. Manjunath, "Unsupervised Segmentation of Color-Texture Regions in Images and Video", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800-810, Aug. 2001.

[16] Y. Wang, J. Yang, and P. Ningsong, "Synergism in Color Image Segmentation", *Proc. of the 8th Pacific Rim International Conference on Artificial Intelligence*, vol. 3157, pp. 751-759, Auckland, New Zealand, Aug. 2004.

[17] B. Georgescu, I. Shimshoni, and P. Meer, "Mean Shift Based Clustering in High Dimensions: A Texture Classification Example", *Proc. of the IEEE International Conference on Computer Vision*, pp. 456-463, Nice, France, 2003.

[18] S. E. Hernandez and K. E Barner, "Joint Region Merging Criteria for Watershed-Based Image Segmentation", *Proc. of the 2000 International Conference on Image Processing*, vol. 2, pp. 108-111, Vancouver, BC, Sept. 2000.

[19] J.A. Withers, K.A. Robbins, "Tracking Cell Splits and Merges", *Proc. of the IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 117-122, San Antonio, TX, 1996.

[20] C. R. Dyer, "Volumetric Scene Reconstruction From Multiple Views", in L.S. Davis, *Foundation of Image Understanding*, Boston, Kluwer, pp. 469-489, 2001.