TIME-VARYING-GEOMETRY OBJECT SURVEILLANCE USING A MULTI-CAMERA ACTIVE-VISION SYSTEM

Matthew Mackay, Robert G. Fenton, and Beno Benhabib

Department of Mechanical and Industrial Engineering,
University of Toronto, 5 King's College Road
Toronto, ON, Canada, M5S 3G1
[mackay@mie.utoronto.ca]

Abstract

This paper presents a novel, agent-based sensing-system reconfiguration methodology for the recognition of time-varying-geometry targets (objects or subjects). A multi-camera active-vision system is used to improve form-recognition performance by selecting near-optimal viewpoints along a prediction time horizon. The proposed method seeks to maximize the target visibility in a cluttered, dynamic environment. Simulated experiments clearly show a tangible potential performance gain.

Keywords: Surveillance, Sensing-System Reconfiguration, Active Vision, Form Recognition

Static Environment

The survey paper [1] characterizes sensor-planning methods as either *generate-and-test* or *synthesis*. Generate-and-test methods discretize the domain to limit the number of configurations that must be considered, and evaluate possible configurations with respect to task constraints (e.g., [10]). Synthesis methods characterize task requirements analytically, and determine sensor poses by finding a solution to the set of constraints given by the current system state (e.g., [11], [12]). One can note, however, that most examples of reconfiguration in static environments tend to be application-specific (e.g., [13], [14]).

Dynamic Environment

A natural extension to exploring a static environment with mobile cameras is the consideration of moving targets, obstacles, and sensors – a dynamic environment [15]. For example, in [7] an 11-camera system was used to examine the effects of viewpoint on recognition rates for human gait. In [16] and [17], multiple mobile sensors were positioned on-line, for the surveillance of maneuvering targets in the presence of static obstacles. More recently, agent-based planning methods were applied to the on-line sensing-system reconfiguration problem ([3], [18], and [19]). Other examples include ([20]-[23]).

Static-Form Recognition

A logical starting point in any time-varying geometry-recognition algorithm is the identification of a single, static form. However, since this would require an existing database of characteristic data for known poses, past work focused on merely reconstructing the model of an unknown object ([24]-[28]). Earlier human-gait recognition works advocated that it might be possible to uniquely identify an individual based on their gait (e.g., [29]). Research in the area began with algorithms designed to distinguish the current form of a human given a single image [30]. Using key-point markers, it has been shown that the gait of an individual can be uniquely distinguished at a rate above that of random chance [31]. The research results reported in [32] showed that automatic face and gait recognition can be combined, using decision-level data fusion, for human identification.

Dynamic-Form Recognition

Many time-varying-geometry objects exhibit specific, repeatable sequences of form that one might wish to recognize [33]. Common recognition approaches have been classified into three general categories: *template matching*, *semantic approaches*, and *statistical approaches* [15].

The two principal qualitative goals that the surveillance system should achieve are:

- *Real-time operation:* All operations must be limited in computational complexity and depth, such that real-time operation of the system is not compromised.
- *Robustness:* The system must be robust to faults, and the likelihood of false identification or classification must be minimized.

The performance of a surveillance system can, thus, be characterized by the success of the vision task in recognizing the target form and its current action. This task depends primarily on the quantity and quality of the sensor data that is collected, characterized herein by a visibility metric, V. This metric in turn depends on the current form and pose of the OoI, the poses of the obstacles, and the poses of the cameras. However, the only variables that the sensing system has direct control over are the poses of the cameras.

The visibility metric for the i^{th} camera at the j^{th} demand instant, t_j, is expressed herein as a function of $\boldsymbol{p}_{S_i}^j$, the pose of the i^{th} sensor, S_i, at the j^{th} instant:

$$V_i^j = f_i^j\left(\boldsymbol{p}_{S_i}^j\right), \tag{1}$$

where pose is defined as a 6D vector $\begin{bmatrix} x & y & z & \varphi & \psi & \theta \end{bmatrix}$ representing position, (x,y,z), and orientation, (φ,ψ,θ). Thus, this paper proposes a global formulation of the reconfiguration problem for a sensing system with n_{sens} sensors, n_{obs} obstacles, and with prediction over the time horizon ending at the m^{th} demand instant:

For each demand instant, t_j, j=1 to m, perform the following:

For each sensor, S_i, i=1 to n_{sens}, solve the following:

Given: $\boldsymbol{p}_{S_i}^0, \boldsymbol{p}_{OoI}^0, \boldsymbol{u}^0, \boldsymbol{p}_{obs_k}^0$; k = 1, to n_{obs}, (2)

Maximize: $Pr = g\left(V_i^l\right)$; l = 1 to j, (3)

Subject to: $\boldsymbol{p}_{S_i}^l \in P_i$, (4)

$\boldsymbol{p}_{S_i}^l \in A_i^l$, (5)

$V_i^l \geq V_{\boldsymbol{min}}$; l = l to j, (6)

End of loop.

Continue while: $t_{proc} < t_{max}$,

Above $\boldsymbol{p}_{OoI}^j$ is the pose of the OoI at the j^{th} demand instant, $\boldsymbol{p}_{obs_k}^j$ is the pose of the k^{th} obstacle at the j^{th} demand instant, $\boldsymbol{u}^j$ is the feature vector of the OoI at the j^{th} demand instant,

3. PROPOSED METHODOLOGY

The proposed methodology advocates sensing-system reconfiguration via an agent-based approach, Figure 1. The individual modules are described below.

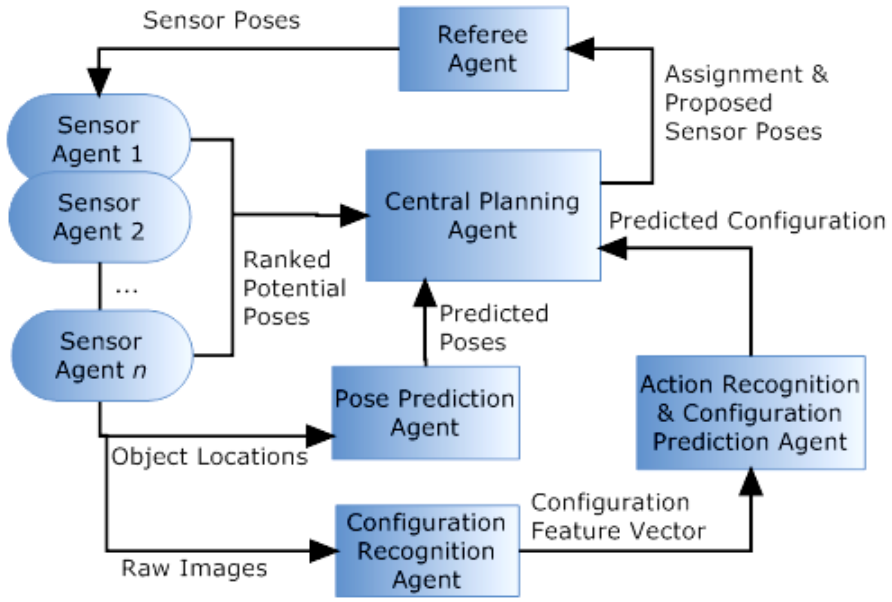

Figure 1. Structure of proposed agent-based methodology.

Sensor Agents

At the lowest level, each sensor agent may be associated with a sensor (i.e., camera) present in the given physical system. The exact configuration (in terms of number and composition of the sensor set) can be determined through a number of established methods ([1], [10], and [11]). It is assumed that each camera is reconfigurable in terms of its pose and that each is limited in capability by positional and rotational velocity and acceleration:

$$t_d = t_1 - t_0 \,, \tag{9}$$

$$\boldsymbol{L}_{min} < \boldsymbol{x}_1 < \boldsymbol{L}_{max} \,, \tag{10}$$

$$\boldsymbol{x}_{L-} < \boldsymbol{x}_1 < \boldsymbol{x}_{L+} \,, \tag{11}$$

$$\boldsymbol{x}_{L-} = f(\boldsymbol{x}_0, t_d),\ \boldsymbol{x}_{L+} = f(\boldsymbol{x}_0, t_d) \,, \tag{12}$$

where $\boldsymbol{x}_0$ is the initial position, $\boldsymbol{x}_1$ is the final position, t_0 is the initial time, t_1 is the final time, t_d is the total time between the demand instants, and $L_{min/max}$ are the outer limits of the motion

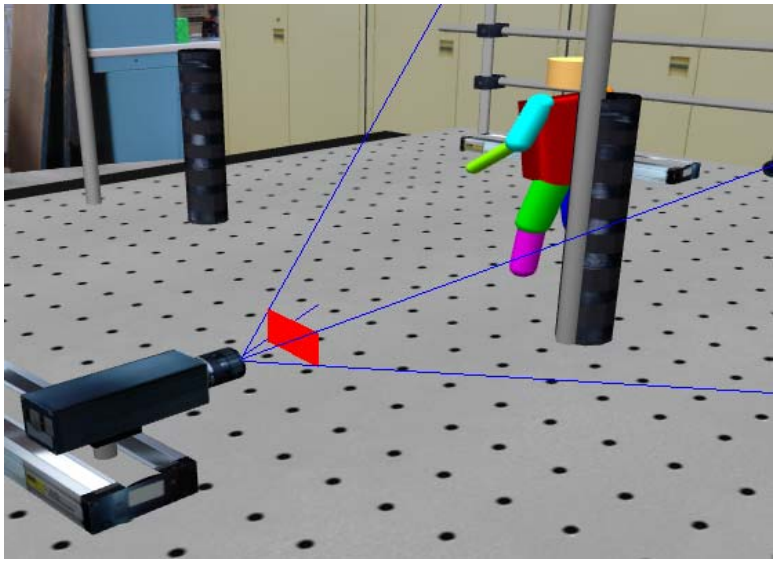

Figure 2 - (Top) Example of 3-D simulation showing projection plane

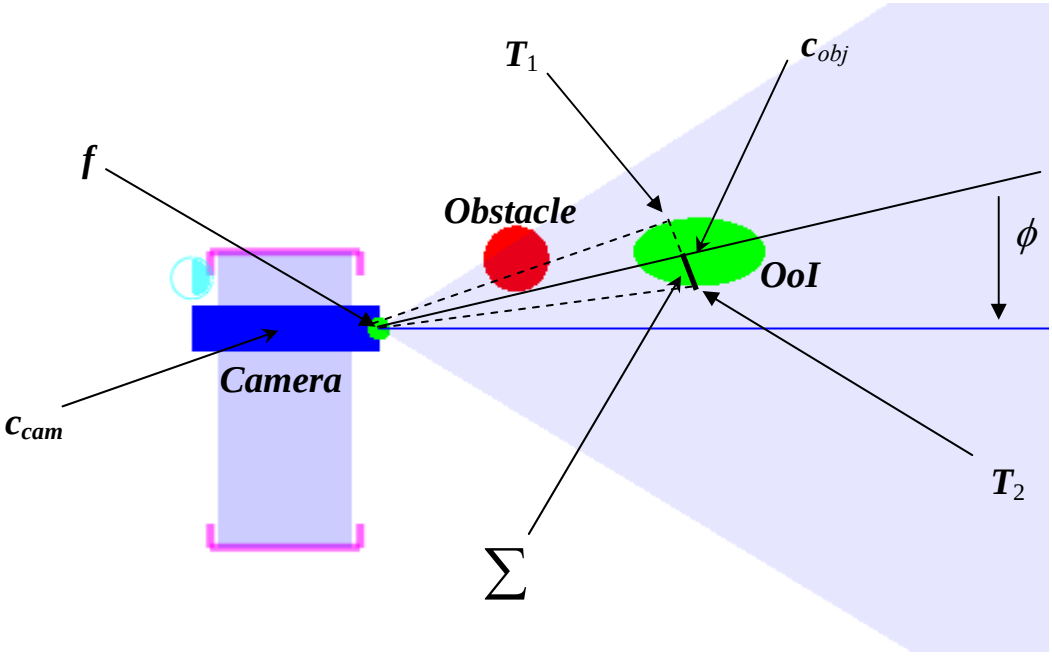

Figure 3 - Top-down view (different scene) of projection of the virtual OoI/obstacle cylinders onto camera plane.

Also in (13), the vector $\boldsymbol{f}$ represents the focal point of the camera, and $\boldsymbol{c}_{obj}$ is the center of the OoI. The center term gives the distance from the focal point to the OoI center; essentially, it is a metric of the size of the OoI in the final image. It is normalized by d_{max}, which is the maximum possible distance from the focal point of the camera to the OoI that is considered to be within the confines of the workspace. d_{max} can be found off-line by solving Equation (16). ϕ is the angle between the focal line of the camera and the line passing through both the camera's rotation center, $\boldsymbol{c}_{cam}$, and the OoI's center, $\boldsymbol{c}_{obj}$. It is a measure of centering of the OoI in the camera image, and normalized by the maximum possible difference, ϕ_{max}. This

which sub-parts of the OoI are not currently well represented in the dataset is also included.

Pose-Prediction Agent

This agent predicts the future poses of the OoI and all obstacles in the workspace from historical data. A number of well established options exist, such as the Kalman Filter (KF) and its variants [40].

Referee Agent

This agent ensures that global rules are not violated – rules imposed on the overall system behavior that are not captured directly by the optimization problem, or by the specifications of the other system agents. Typically, such rules are highly application specific, and have a variety of uses. For example, a rule could be defined to guarantee the assignment of a minimum number of cameras at each instant to the surveillance of the OoI. While a real-world application would have a significantly more demanding set of rules, this single rule does serve a purpose for the simulations that follow.

Form- and Action-Recognition Agents

The proposed static-form recognition method is model based. Data on object forms are stored as a feature vector derived from geometrical data – the feature vector consists of a list of interest point locations on the OoI, relative to an origin point on the object. The system must be able to determine the location of the reference point in some coordinate system (possibly world coordinates) and the locations of as many interest points as possible in this same system. There are many computer vision methods available for this purpose, such as local PCA (Principal Component Analysis), Harris Corner points, Harris DOG (Difference of Gaussian) points, and Image Neighborhood Descriptors [41].

In order to recognize the current OoI action, the sensing system would also have to identify the location in a database sequence of two distinct target forms, referred to as the start and end frames. Using time normalized input data, a metric of distance from the library data could be formulated for each database set that contains both the start and end forms. A simple approach would be to consider the most current data as the end frame, and use some number of previous frames. However, this has the potential to miss action transitions, and introduce other artifacts. Thus, a continuous, depth-limited scan must be implemented.

A fixed assignment of three cameras to the nearest instant, t_1, and one camera to the farthest instant, t_2, is used for these experiments. Before the *deadline* (or t_0), the following rules are used herein to determine camera assignment and determination of their poses:

- The three cameras with the highest maximized expected visibility at time instant t_1 are assigned to service the OoI at that instant. They begin moving at the end of the decision deadline, (i.e., at time instant t_0) and continue until they reach their final pose (which must happen no later than t_1).
- The agent with the lowest maximized expected visibility metric is asked to re-evaluate its metric for an additional demand instant in the future, t_2, and is assigned to this instant.
- Any cameras with an expected visibility metric less than the minimum, V_{min}, at the final pose chosen above are not used in the fusion process for the nearest instant, t_1. These sensors still move to the positions determined by Rules 1 and 2, in anticipation of future demand instants.

The software developed can produce (simulated) images from any of the (virtual) cameras, and includes a segmented model that approximates the human form. A simple walking action was used in all the simulated experiments. Custom surface models of all objects in the environment were created, and camera calibration matrices were generated based on data from physical cameras in our matching physical setup. Form recognition was achieved by a model-based algorithm using color segmentation, Appendix A.

4.2 Experimental Results

Numerous (simulated) experiments were performed to verify achievable, tangible improvement on form recognition through sensing-system re-configurability. Two primary experiments are presented herein. In the first experiment, we seek to quantify how system re-configurability affects form recognition of time-varying-geometry subjects (from noisy images) in the presence of static obstacles but with ideal (perfect) OoI motion prediction. In the second experiment, we seek to quantify the effect of real-world tracking through the addition of noise on the OoI motion prediction.

Results from additional two experiments are presented in Appendix C on non-uniform importance and self-occlusions.

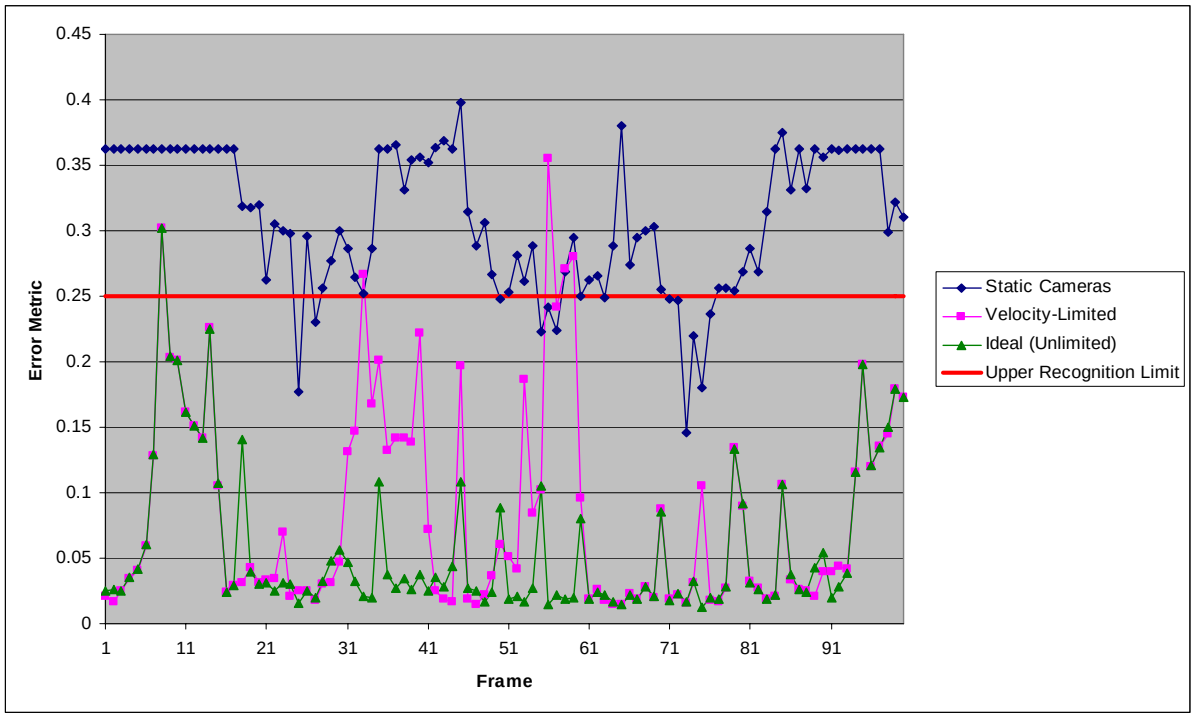

Figure 5. Comparison of error metric over three trials of 100 frames each, with walking action performed by the target.

Experiment 2 – Effect of OoI-Pose-Prediction Noise

From the results of Experiment 1, one can conclude that sensing-system re-configurability tangibly reduces the average error-metric values and improves form-recognition performance. Additional simulated experiments are presented herein to validate the performance of the system under non-ideal OoI pose estimation, which would be inherent in any real-world application.

For these simulated experiments, the prediction agent implementation is that of a Kalman Filter (KF), with second-order (position, velocity, and acceleration) state variables. The input observations to the KF are taken from the form-recognition agent, which tracks the position of the head of the subject as a reference center-point. Only 2D tracking is considered, as it is assumed for these trials that the subject does not change elevation significantly. Input images to the system still come from the simulation environment developed for the previous trials.

A total of four trials were performed. As before, an error metric upper limit of 0.25 was selected, and real-world velocity- and acceleration-constrained reconfigurations were used. The initial positions of all obstacles and the subject are similar to the previous experiments. The four trials consisted of (1) Ideal prediction, Static Obstacles (similar to Trial 2 from

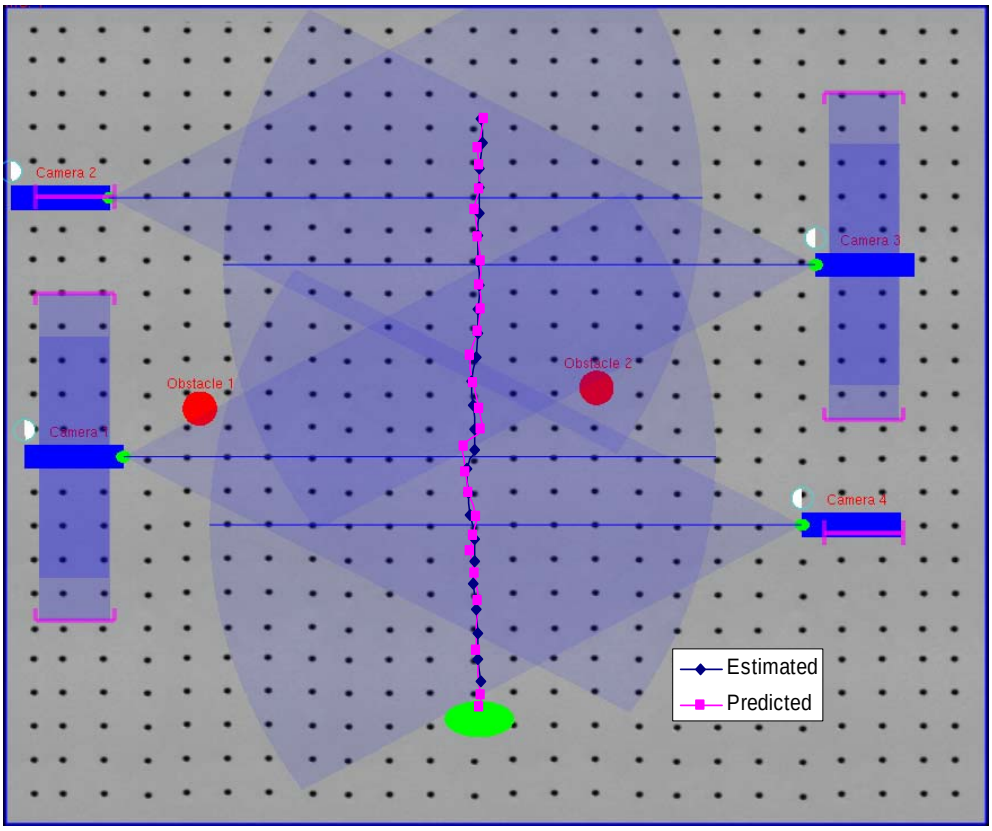

Figure 7. Subject pose estimation for Trial 3.

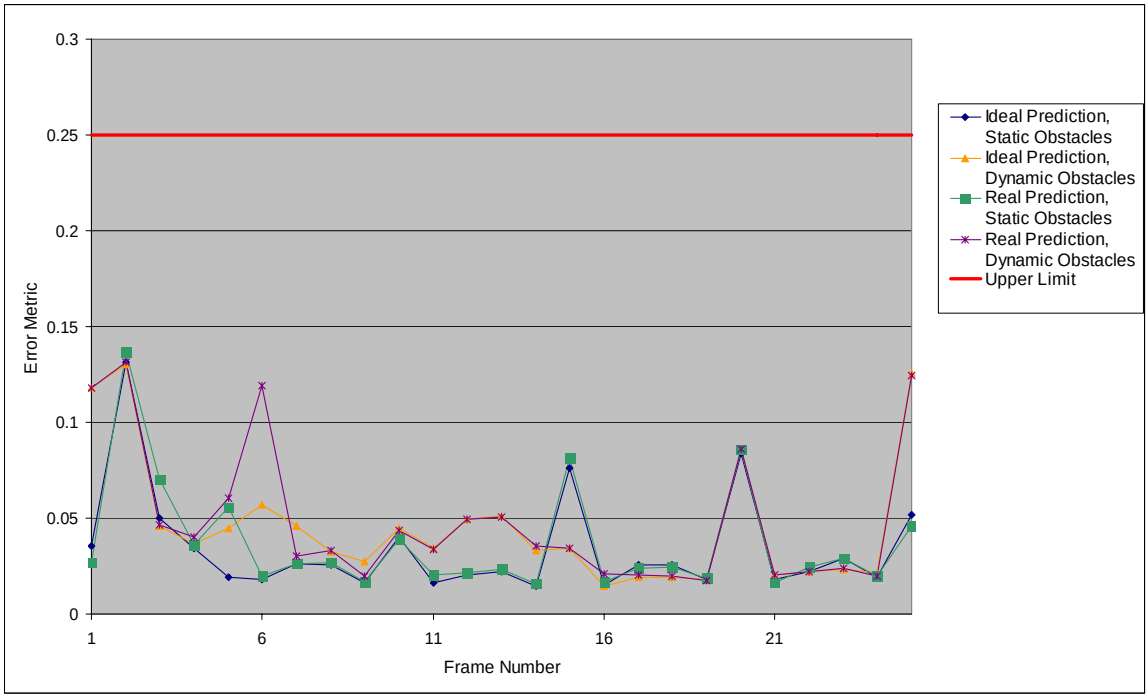

Figure 8. Error metric comparisons.

each database set containing these start and end forms is calculated. This distance is compared to a second limit, and a match is determined if the distance is sufficiently small.

It is important to note that all vision algorithms used in the simulated experiments are tolerant to partial occlusions, image noise, etc. In addition, the methodology itself is designed to be robust to uncertainty introduced at various points, such as during pose and form recovery. As such, the experiments have been carefully designed to highlight only a single factor at a time for comparison.

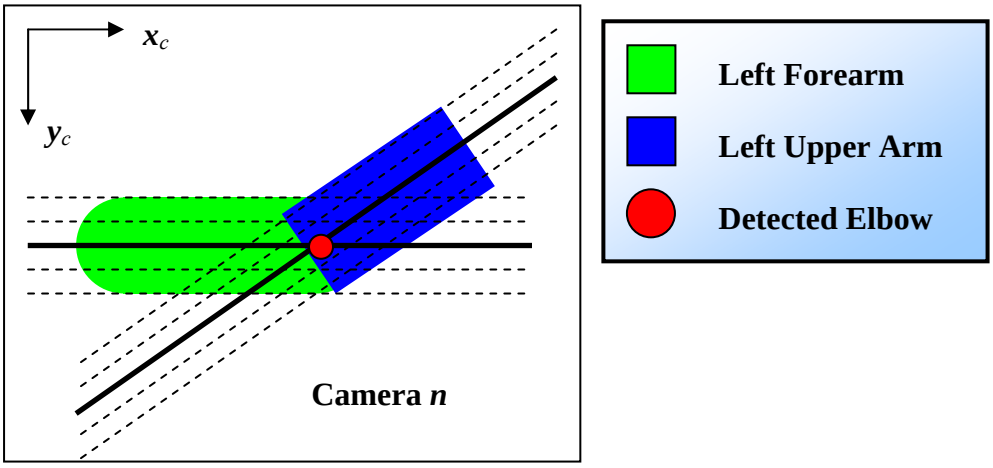

Figure A1. Simplified implementation of key-point detection using color cues.

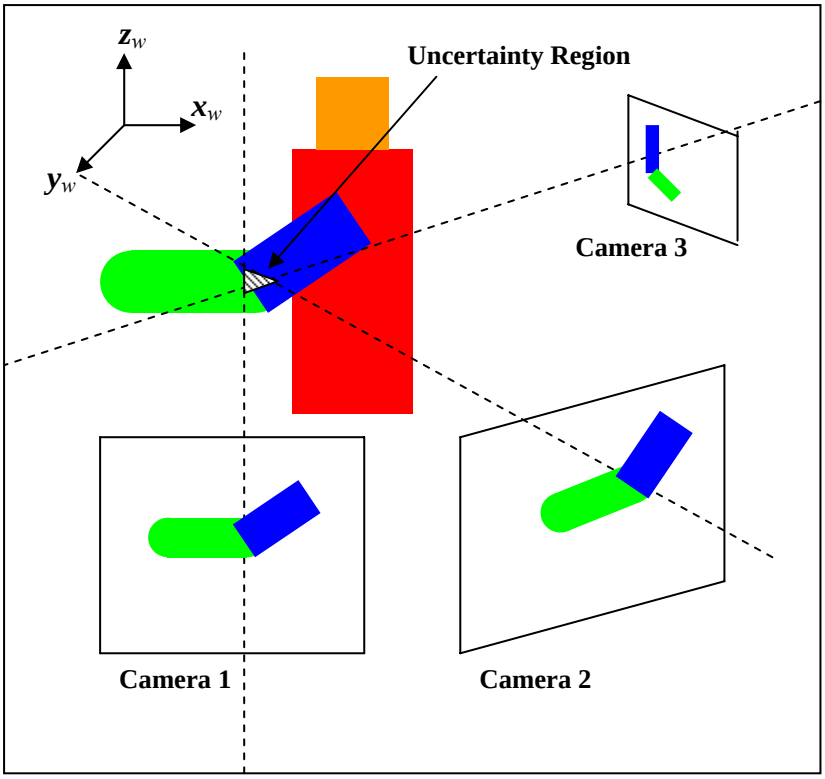

Figure A2. Implementation of algorithm to recover world coordinates of key-points.

Appendix C: Non-Uniform Importance and Self-Occlusions

Experiment – Verification of Non-Uniform Importance

For this set of experiments, the goal is to verify that viewpoints are non-uniform in their importance. Namely, the goal is to show that some views may contain more unique, useful information about the subject than do others. For this experiment, the same initial conditions as in Experiment 1 were used, with the exception of the removal of both obstacles. For each of the four runs, a subset of two cameras is chosen to reconstruct the model of the subject at each demand instant. If the subject were completely uniform in appearance, and all viewpoints offered exactly the same information about the subject, then, one would expect to note very close correlation in each of the error metric graphs. However, as one can note in Figure C1, this is clearly not the case.

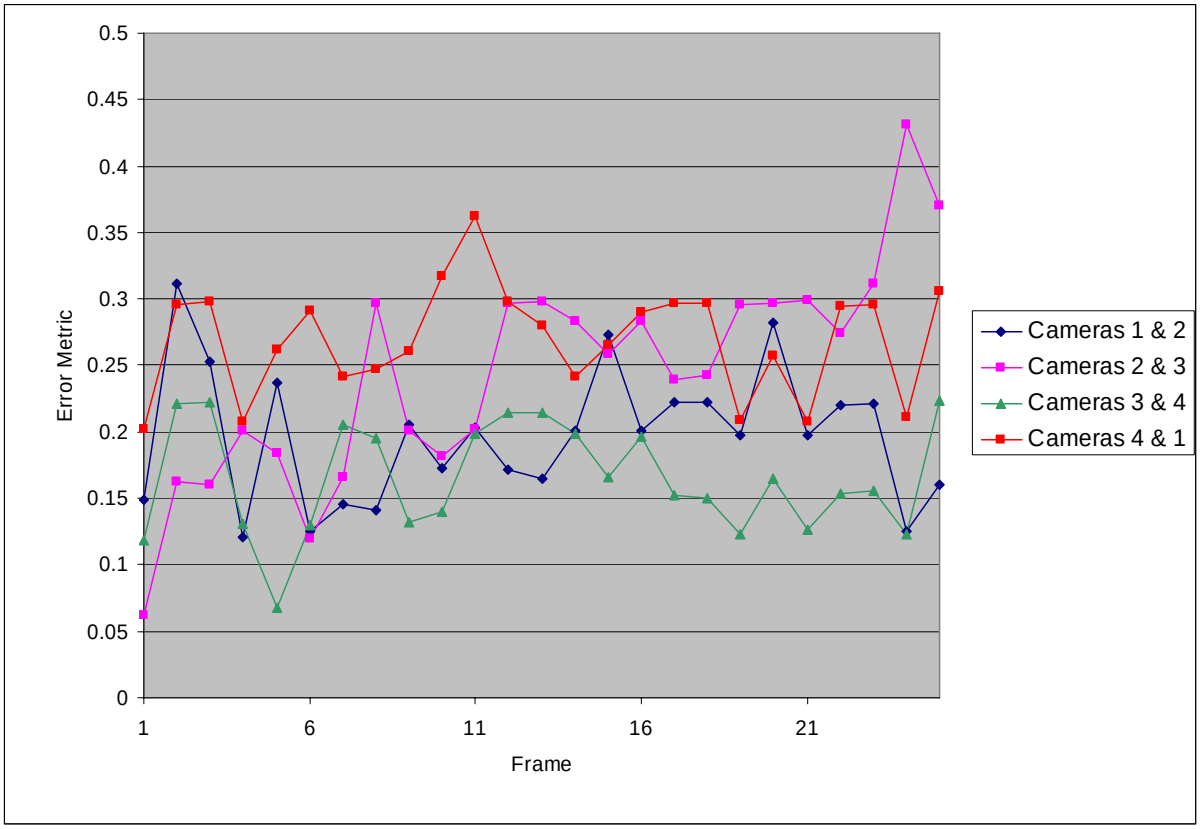

Figure C1. Comparison of error metric for four different sets of two cameras each.

From Figure C1, one can note that, for most frames, there exist major differences in the error metric value since entire segments of the subject are not recovered due to non-uniform viewpoints – as a view from two cameras is the minimum amount of data needed to recover the world position of a key-point under our implementation. In cases such as Frames 2, 5, 24, etc., one can clearly note that different parts of the object are recognized under different

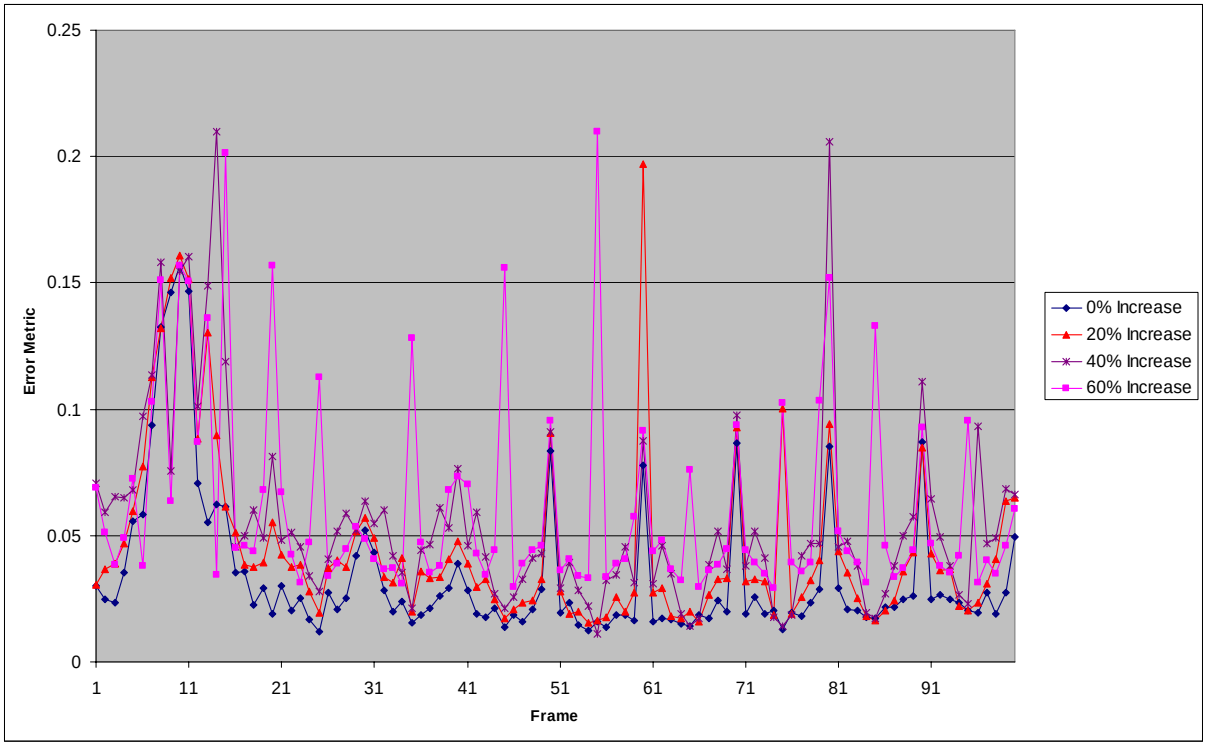

Figure C2. Comparison of error metric for varying degrees of limb size.

REFERENCES

[1] K.A. Tarabanis, P.K. Allen, and R.Y. Tsai, "A Survey of Sensor Planning in Computer Vision," *IEEE Transactions on Robotics and Automation*, vol. 11, no. 1, pp 86–104, Feb. 1995.

[2] J. Miura and K. Ikeuchi, "Task-Oriented Generation of Visual Sensing Strategies in Assembly Tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 2, pp. 126-138, Feb. 1998.

[3] M.D. Naish, E.A. Croft, and B. Benhabib, "Coordinated Dispatching of Proximity Sensors for the Surveillance of Maneuvering Targets," *Journal of Robotics and Computer Integrated Manufacturing*, Vol. 19, No. 3, pp. 283-299, 2003.

[4] S. Sakane, T. Sato, and M. Kakikura, "Model-Based Planning of Visual Sensors Using a Hand-Eye Action Simulator: HEAVEN," *Proc. of Conf. on Advanced Robotics*, pp. 163–174, Versailles, France, Oct. 1987.

[5] C.K. Cowan and P.D. Kovesik, "Automated Sensor Placement for Vision Task Requirements," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 3, pp. 407-416, May 1988.

[6] R. Bodor, P. Schrater, and N. Papanikolopoulos, "Multi-Camera Positioning to Optimize Task Observability," *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, pp. 552-557, 2005.

[7] S. Yu, D. Tan, and T. Tan, "A Framework for Evaluating the Effect of View Angle, Clothing, and Carrying Condition on Gait Recognition," *Proc. of Int. Conf. on Pattern Recognition,* pp. 441-444, *Hong Kong,* 2006.

[23] A. Bakhtari, and B. Benhabib, "An Active Vision System for Multi-Target Surveillance in Dynamic Environments," *IEEE Transactions on Systems, Man, and Cybernetic*s, vol. 37, no. 1, pp. 190-198, 2007.

[24] E. Marchand and F. Chaumette, "Active Vision for Complete Scene Reconstruction and Exploration," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 21, no. 1, pp. 65-72, Jan. 1999.

[25] R. Pito, "A Solution to the Next Best View Problem for Automated Surface Acquisition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 1016-1030, Oct. 1999.

[26] S.D. Roy, S. Chaudhury, and S. Banerjee, "Isolated 3-D Object Recognition through Next View Planning," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 30, no. 1, pp. 67-76, Jan. 2000.

[27] J.E. Banta, L.M. Wong, C. Dumont, and M.A. Abidi, "A Next-Best-View System for Autonomous 3-D Object Reconstruction," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 30, no. 5, pp. 589-598, Sept. 2000.

[28] S.Y. Chen and Y.F. Li, "Vision Sensor Planning for 3-D Model Acquisition," *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 35, no. 5, pp. 894-904, Oct. 2005.

[29] G. Johansson, "Visual Perception of Biological Motion and a Model for its Analysis," *Percept. Psychophys.*, vol. 14, no. 2, pp. 201-211*,* 1973.

[30] R. Chellappa, A.K. Roy-Chowdhury, and S K. Zhou, "*Recognition of Humans and Their Activities Using Video,"* San Rafael, CA: Morgan & Claypool Pub., pp. 53-92, 2005.

[31] J.E. Cutting and L.T. Kozlowski, "Recognizing Friends by Their Walk: Gait Perception without Familiarity Cues," *Bull. Psychonom. Soc.,* vol. 9, no. 5, pp. 353-356, 1977.

[32] A. Kale, A.K. Roy-Chowdhury, and R. Chellappa, "Fusion of Gait and Face for Human Identification," *Proc. of ICASSP04,* pp. 901-904, Montreal, Canada, 2004.

[33] H. Kobayashi and F. Hara, "Facial Interaction between Animated 3D Face Robot and Human Beings," *Proc. of IEEE Int. Conf. on Computational Cybernetics and Simulation,* pp. 3732-3737, *Orlando, FL*, 1997.

[34] M. Dimitrijevic, V. Lepetit and P. Fua, "Human Body Pose Recognition Using Spatio-Temporal Templates," *ICCV workshop on Modeling People and Human Interaction*, pp. 127-139, Beijing, China, October 2005.

[35] D. Cunado, M. S. Nixon, and J. Carter, "Automatic Extraction and Description of Human Gait Models for Recognition Purposes," *Computer Vision and Image Understanding*, vol. 90, pp. 1-41, 2003.

[36] G. V. Veres, L. Gordon, J. N. Carter, and M.S. Nixon, "What Image Information is Important in Silhouette-Based Gait Recognition?" *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition,* pp. 776-782, Washington, D.C., 2004.

[37] X. Weimin, L. Ying, H, Hongzhe, X. Lun, W. Zhiliang, and C. Fengjun, "New Approach of Gait Recognition for Human ID," *Proc. of ICSP04,* pp. 199-202, Beijing, China, 2004.

[38] N. Rajpoot and K. Masood, "Human Gait Recognition with 3D Wavelets and Kernel based Subspace Projections," *Proc. of Workshop on Human Activity Recognition and Modeling*, HAREM 2005, Oxford, UK, 2005.