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

1 INTRODUCTION

In recent years, the surveillance of human targets has become an area of intense research. The need to capture detailed data from a complicated, articulated target (such as a human) presents a difficult problem. Many algorithms developed thus far have been designed to work under ideal conditions, with clear, un-occluded images [1]. The effectiveness of such algorithms could be significantly improved if reconfigurable sensing systems were employed in order to reduce uncertainty inherent in the sensing process. However, current sensing-system planning (also known as sensor planning or sensing-system reconfiguration) methods proposed in the literature mostly deal with non-time-varying (fixed) target geometries (e.g., [2]). Thus, the objective of this paper is to present a novel, agent-based sensing-system reconfiguration methodology for the form recognition of time-varying geometry targets (objects or subjects).

Sensing-system reconfiguration is commonly defined as using a formal method to select the number, types, locations, and internal parameters of sensors employed in the surveillance of an object or subject. An effective, real-time surveillance system must be able to cope with the presence of multiple static or dynamic (maneuvering) targets and obstacles as part of its sensing solution [3]. These factors complicate the process of sensing-system reconfiguration, as it applies to time-varying geometry objects, by introducing such complications as: non-uniform importance, self-occlusions, and continuous surveillance ([4], [5]).

Non-uniform importance refers to the fact that viewpoints are differentiated in the useful information they can provide to a vision algorithm – by both their viewpoint and by their time instant [6]. For instance, different views of a human subject would contain different sub-parts of the overall model (form). If the algorithm already has input data containing some parts of the overall model, views containing images of the remaining parts could become relatively more important. Similarly, since the subject would change in form over time, so the same relative viewpoint would not necessarily have the same importance at two different instances.

Below, a detailed literature review discusses past research work on form recognition of time-varying geometry objects and sensing-system reconfiguration. As mentioned above, the majority of past work on sensing-system reconfiguration has dealt with fixed-geometry objects (e.g., [7]). Some past works, however, have applied reconfiguration algorithms developed for fixed-geometry objects to those with time-varying geometries (e.g., [8]). Similarly, an off-line planning method was proposed in [9] to track the motion of an articulated human form using eight static cameras. While such approaches would have merit, the problems mentioned above cannot be directly addressed.

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].

In template matching, input images are compared directly to stored templates and multiple pose matches over time form a sequence template (e.g., [34]). Semantic approaches are model-based approaches, in that a high-level representation of the target may be constructed (e.g., [35]). Statistical approaches can be seen as extension of both previous approaches, in that they attempt to reduce the dimensionality of database matching through statistical operations on the template database (e.g., [36]-[39]).

The key similarity to most of the above form-recognition methods is that sensors are not considered as part of the methodology – input data is taken to be fixed, with no opportunity for quality improvement. However, research has shown that many factors, such as viewing angle, are important factors in recognition performance [7]. Thus, in this paper, a multi-camera active-vision system is used to improve form-recognition performance by selecting near-optimal viewpoints. The proposed method seeks to maximize the visibility of the time-varying-geometry targets, hereafter referred to as Objects of Interest (OoIs), moving in a cluttered, dynamic environment.

2. PROBLEM FORMULATION

A primary assumption in this paper is that Objects of Interest (OoIs) may move on *a priori* unknown paths, in a workspace cluttered with multiple, dynamic obstacles, also moving on *a priori* unknown motion paths. As such, a suitable starting point would be first to define the expected tasks for the surveillance system, followed by the principal qualitative goals. The surveillance system must perform the following tasks:

- *Detection:* All objects in the scene must be detected and categorized as either the OoI or an *Obstacle* upon entering the workspace.
- *Tracking:* Each object must be tracked, and an estimate of its future pose (position and orientation) maintained.
- *Reconfiguration:* Given historical, current, and predicted data about the OoI and obstacles, an achievable set of poses for all sensors that minimizes uncertainty for the surveillance task at hand must be determined.
- *Recognition:* Data from all sensors must be fused into a single estimate of the OoI’s current geometry. A further estimate must reconcile this geometry data with historical data to determine the current action of the OoI.

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 $[x \quad y \quad z \quad \varphi \quad \psi \quad \theta]$ 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 = $1$ 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,

P_i is the discretized set of feasible camera poses for the i^{th} camera, A_i^j is the discretized set of achievable camera poses for the i^{th} camera at the j^{th} demand instant, V_{min} refers to a user-defined threshold of minimum visibility, t_{proc} is the time spent processing data, and t_{max} is the maximum time limit for the selection of a final pose set. The two sets of feasible and achievable poses, P_i and A_i^j, are governed by:

$$A_i^l \subseteq P_i \,, \tag{7}$$

$$\boldsymbol{p} \in P \text{ iff. } \boldsymbol{p}_{low} \leq \boldsymbol{p} \leq \boldsymbol{p}_{upp} \,, \tag{8}$$

where $\boldsymbol{p}_{low}$ and $\boldsymbol{p}_{upp}$ are the lower and upper movement limits of the sensor motion. These limits are defined by the physical constraints of the sensors themselves. The set of achievable poses represents the set of poses (out of all feasible poses) that can be reached given the limited motion capabilities of the sensor and the remaining time. The determination of the subset $A_i^j \subseteq P_i$ depends on the model of motion used, and is not specified.

The proposed performance objective function, *Pr*, depends on the visibility metric of each camera at all demand instants on the horizon. It is a measure of success in achieving the sensing objective [16]. Overall, the proposed formulation seeks first to maximize the visibility of the OoI at the immediate future demand instant, [t^l], for all cameras. If sufficient time remains, the system seeks to maximize expected visibility at [t^l and t^2], then, [t^l, t^2, and t^3], and so on. As such, a higher overall metric value may be achieved at later demand instants, possibly at the expense of near-future visibility.

The above trade-off can be controlled by adjusting the minimum desired visibility, which in turn controls the camera assignment at each demand instant. If the condition $V_i^l \geq V_{\boldsymbol{min}}$ cannot be met for a given camera, it is considered unassigned, and should be moved in anticipation of future demand (a second maximization of expected visibility, at a demand instant further into the future, would take place). However, empirically-determined weights could be assigned to nearer demand instants in order to minimize exposure to future uncertainties in estimated poses.

One may note that the computational complexity is bounded, though the determination of poses at each future instant does depend on poses determined for previous instants, back to the current time, t^0.

3. PROPOSED METHODOLOGY

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

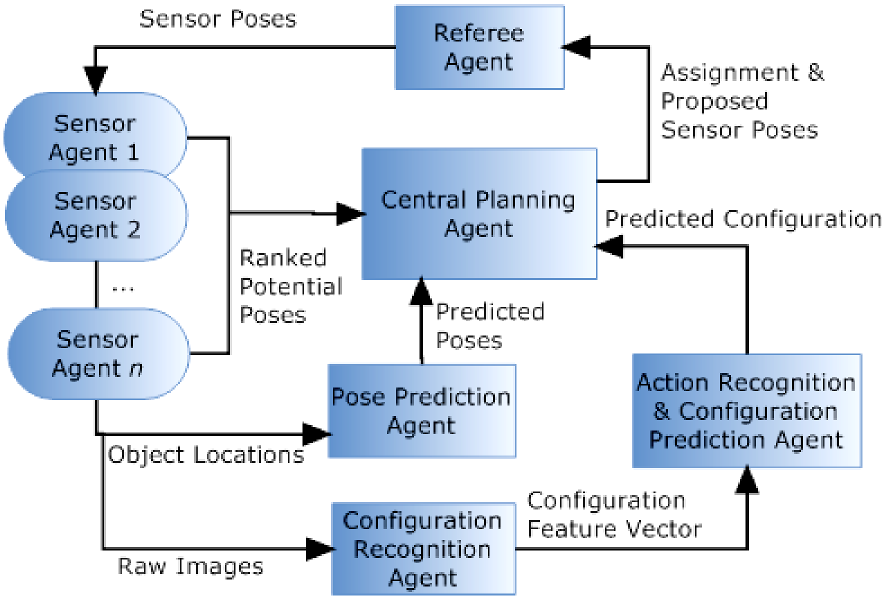


Figure 1. Structure of proposed agent-based methodology.

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:

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

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

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

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

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 axis. Similarly, $\boldsymbol{x}_{L-}$ and $\boldsymbol{x}_{L+}$ are the minimum and maximum poses achievable, respectively, given the capabilities of the sensor, the current pose, and the time remaining. This is captured by equation (9), which defines an arbitrary function for the mapping – the function will depend on the model of motion being used. A similar set of equations can be used to determine the rotational limits in terms of angular velocity and acceleration. This position space can be discretized into n_{pos} possible final positions, where $n_{pos} \propto (t_1 - t_0)$, to bound computational complexity. A continuous algorithm that is limited in iterative depth could also be used. The visibility metric can be evaluated at each discretized sensor pose.

Herein, all known obstacles and the OoI are modeled as elliptical cylinders. A clipped projection plan can be established and all objects projected onto this plane, Figure 2, 3. A sorting algorithm can, then, be used to produce an ordered list from highest to lowest visibility, which is passed from the sensor agent to the central planner. The visibility metric is, thus, defined as:

$$V = \frac{\left[W_{area}\left(\frac{1}{\left\|\boldsymbol{T}_1 - \boldsymbol{T}_2\right\|}\right)\left(\sum_{i=1}^{n_{area}} L_i\right) + W_{dist}\left(\frac{1}{d_{\boldsymbol{max}}}\right)\left(\boldsymbol{f} - \boldsymbol{c}_{obj}\right) + W_{angle}\left(\frac{1}{\phi_{\boldsymbol{max}}}\right)(\phi) \right]}{\left(W_{area} + W_{dist} + W_{angle}\right)}, \tag{13}$$

where

$$\phi = \boldsymbol{cos}^{-1}\left(\left(\frac{\boldsymbol{c}_{obj} - \boldsymbol{c}_{cam}}{\left\|\boldsymbol{c}_{obj} - \boldsymbol{c}_{cam}\right\|}\right) \bullet \left(\frac{\boldsymbol{f} - \boldsymbol{c}_{cam}}{\left\|\boldsymbol{f} - \boldsymbol{c}_{cam}\right\|}\right)\right), \tag{14}$$

$$\phi_{\boldsymbol{max}} = \boldsymbol{max}\left(\left|L_{\phi+} - L_{\phi-}\right|, \left|L_{\theta+} - L_{\theta+}\right|\right). \tag{15}$$

In (13), W_{area}, W_{dist}, and W_{angle} are the weighting constants for the metrics of visible OoI area, zoom factor (size of OoI in the camera view), and angular distance from a view with the OoI centered in the image, respectively. The vectors $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ define the two tangent points to the bounding ellipse (2D), which are also on lines that pass through the focal point of the camera. The sum is over all distinct, visible portions of these lines, L_i, – namely, the sum gives the total length of all the line segments, from $\boldsymbol{T}_1$ to $\boldsymbol{T}_2$, that are visible (not outside the limits of the camera field of view, and not occluded by any obstacles). This sum is normalized by the maximum possible value, which is simply the total length of the line segment between the two tangent points (complete visibility).

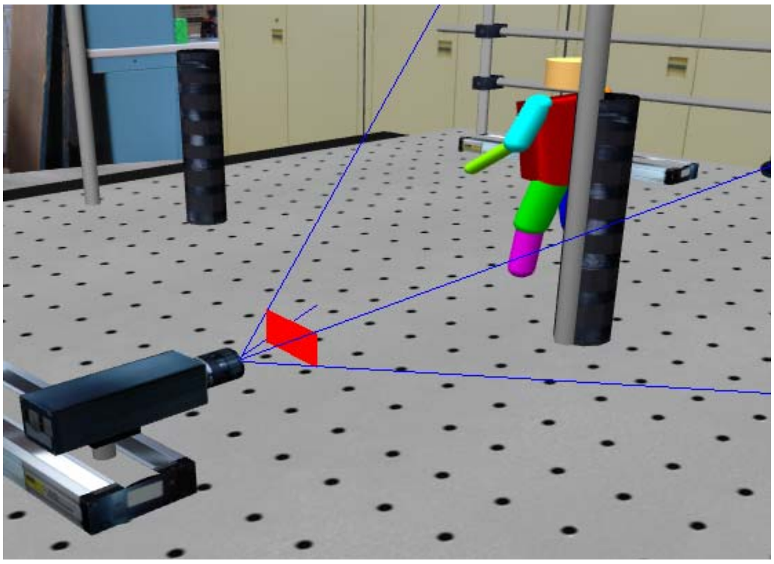


Figure 2 - (Top) Example of 3-D simulation showing projection plane

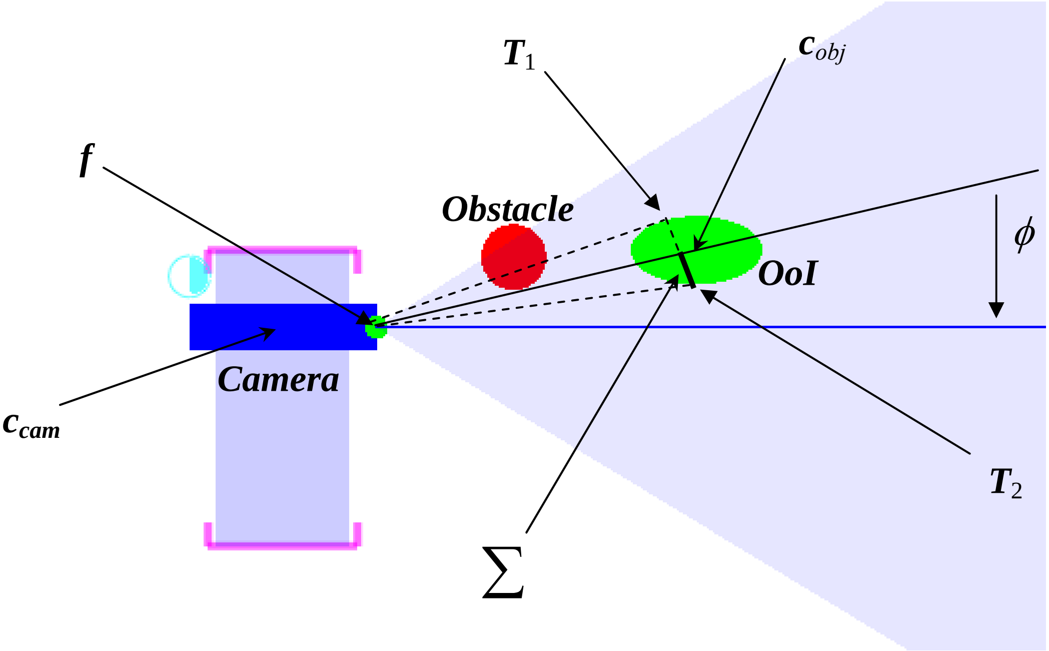


Figure 3 - Top-down view (different scene) of projection of the virtual OoI/obstacle cylinders onto camera plane.

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 maximum can be found as in (15). The values of $L_{\phi+}, L_{\phi-}, L_{\theta+}, L_{\theta+}$ are the upper and lower limits of the camera pan and tilt angles, respectively. The choice between the two normalization factors is not critical, as the difference can be corrected by increasing the corresponding weighting value.

The optimization process is summarized below:

```
For each vertex $\boldsymbol{p}^{i}_{\text{bound}}$, $i$=1, to $n_p$, on the workspace bounding polygon:
        Maximize:     $d = \left\| f(x_{cam}, y_{cam}, z_{cam}, \phi_{cam}, \theta_{cam}, \Psi_{cam}) - p^{i}_{bound} \right\|$      (16)
        Subject to:   $L_{x-} \le x_{cam} < L_{x+}$      (17)
                      $L_{y-} \le y_{cam} < L_{y+}$      (18)
                      $L_{z-} \le z_{cam} < L_{z+}$      (19)
                      $L_{\phi-} \le \phi_{cam} < L_{\phi+}$      (20)
                      $L_{\theta-} \le \theta_{cam} < L_{\theta+}$      (21)
                      $L_{\Psi-} \le \Psi_{cam} < L_{\Psi+}$      (22)
End of Loop.
```

Central Planning Agent

The use of a central planning agent is proposed to accept the sorted visibility evaluations from each sensor agent, as well as a list of the discretized achievable poses, in order to generate camera assignments and select their final poses for the demand instant at hand. The visibility metric list may be depth limited and the discretized range represented only by its outer limits, to save overhead. A simple set of rules could be used to select a subset of the cameras to service the OoI at this demand instant. For example:

- Cameras with a visibility metric less than a minimum, V_{min}, at all poses, are unassigned.
- The M highest visibility cameras are assigned and all others unassigned.
- Agents for unassigned cameras are asked to re-evaluate the visibility metric for additional demand instants. They are moved in anticipation of potentially optimal viewpoints at instants farther into the future.
- For assigned cameras, a weighted sum of metrics is evaluated. This sum must include the base object visibility (13) and a measure of importance of the view (in terms unique data about the object). Namely, feedback from the form-prediction agent on

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.

4. SIMULATED EXPERIMENTS

In this paper, a single-target, dynamic environment is considered, where the OoI is a simple articulated model simulating a human walking. The simulation environment developed is capable of generating synthetic images, as would be viewed by each camera, based on detailed 3D models of all physical objects, (i.e., the OoI and obstacles) as well as a cluttered background. Although obstacles can be represented by simple models derived from real-world objects, the time-varying geometry (human) OoI is more difficult to simulate [42]. A simple segmented figure created from geometric primitives was used in our work. A system of managing and creating realistic library form data for the OoI was implemented, [43]. We also accounted for the real-time nature of the proposed system.

4.1 Experimental Set-Up

The proposed methodology was simplified for the basic verification of its principal operation: The prediction agent was replaced with a 'perfect' prediction simulator (thus, not shown on the figure) – namely, the current poses of all obstacles and the OoI are assumed to be known exactly (with no uncertainty) for this ideal case.

A total of four sensor agents (for four cameras) were implemented – each is responsible for solving its own local-optimization problem directly. The central planner accepts the highest ranked solutions, and utilizes a simple set of rules to select a subset of the cameras to service the OoI at the current demand instant, and to select their final poses. An overview of this process is shown in Figure 4. A fixed time horizon of three demand instants is used. From the current time, the system has until the first instant in the horizon, t_0, to make a final decision on camera assignment and placement.

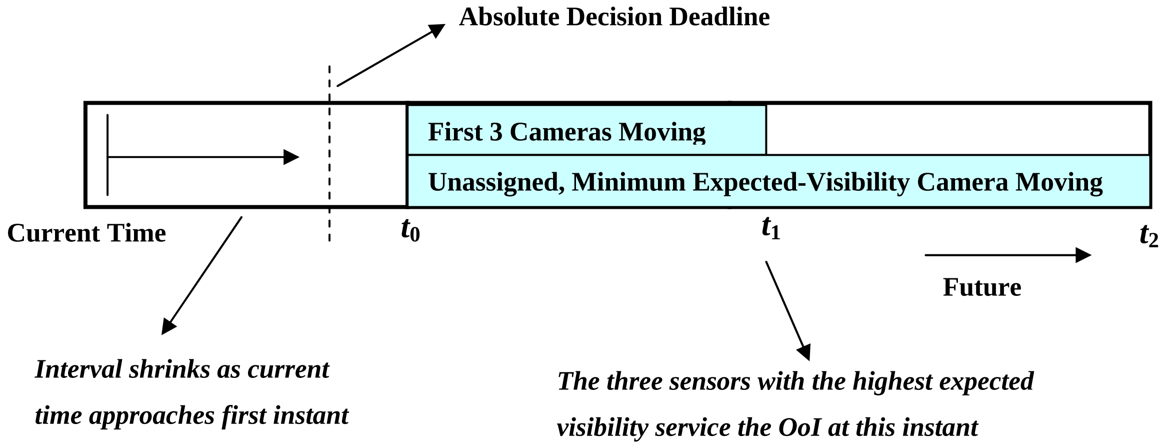

Absolute Decision Deadline
First 3 Cameras Moving
Unassigned, Minimum Expected-Visibility Camera Moving
Current Time
t_0
t_1
t_2
Future
Interval shrinks as current time approaches first instant
The three sensors with the highest expected visibility service the OoI at this instant

Figure 4. Overview of sensor assignment in simulated system.

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.

Experiment 1 – System Reconfigurability

An initial simulated test, consisting of three runs, each showing 100 frames of a subject walking, was performed to determine how sensing-system re-configurability affects form-recognition performance. Static cameras were used for the first trial (i.e., no system re-configurability). The remaining two runs were conducted using a velocity-constrained and an ideal (velocity-unconstrained) reconfiguration system, respectively. In the two dynamic-camera experiments, the initial camera poses were the same as for the static-camera case: two of the four cameras were given linear-translation ability, and all cameras could pan up to ±90° from their initial pose.

In each trial, the subject maintained a constant velocity of 100 *mm/s* on a straight-line path through the center of the workspace. For the velocity-constrained system, the maximum velocity was 450 *mm/s*, with a maximum acceleration of 9000 mm/s^2. The ideal system is considered to have unlimited velocity and acceleration (i.e., instantaneous repositioning of cameras). A virtual image acquisition rate of 10 *frames-per-second* (fps) was assumed for the quasi-static simulations.

In the proposed algorithm, each sensor uses a weighted combination of three sub-metrics to rank poses: the area of the target bounding cylinder that is visible, the angle to the target, and the distance to the target. An upper limit of 0.25 was selected for the overall error metric (the lower the value, the better the match probability). Any frame with an error value above this limit is not considered to yield a positive match. This value was determined through statistical analysis of multiple previous runs, resulting in at least 95% true positive matches for un-rejected frames, and less than 2% false negative.

The results presented in Figure 5 show a clear tangible overall reduction in error values with the use of a velocity-constrained dynamic sensing system versus the static-camera case. However, some problematic instances are still noticeable. For example at Frames 31 to 41, the system was unable to move the cameras to their optimal poses due to movement-time constraints. An example frame analysis is provided in Appendix B.

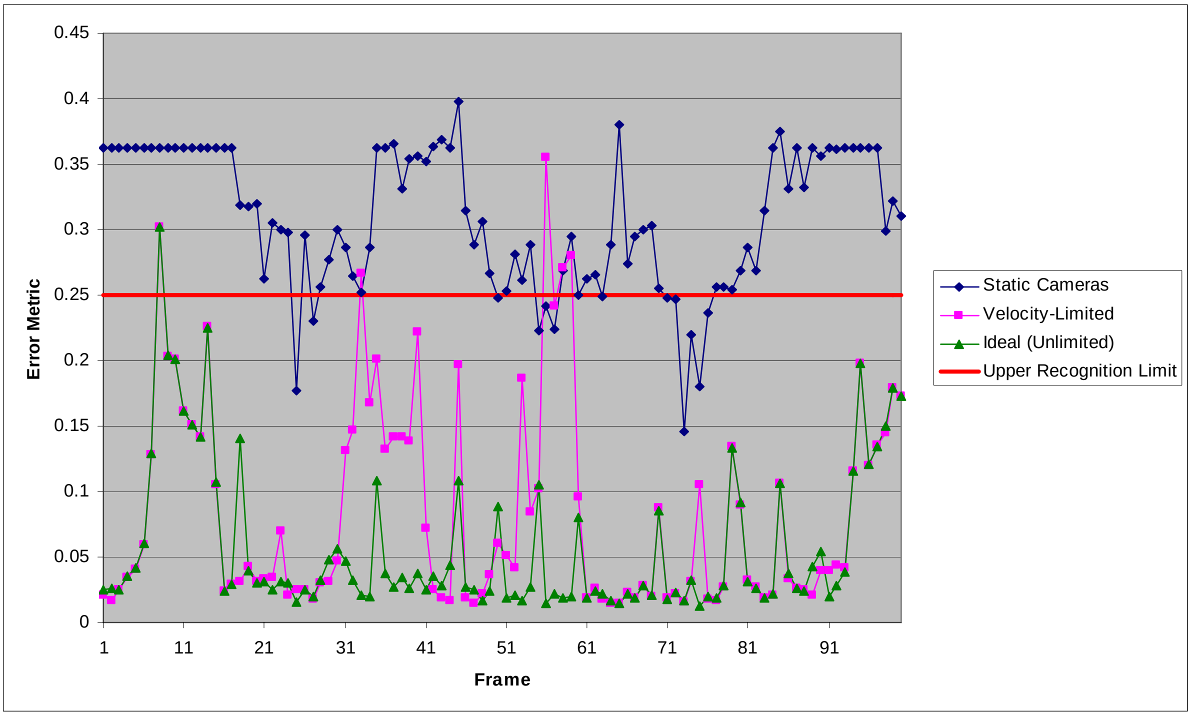


Figure 5. Comparison of error metric over three trials of 100 frames each, with walking action performed by the target.

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

Experiment 1), (2) Ideal Prediction, Dynamic Obstacles, (3) Real-world prediction, Static Obstacles, and (4) Real-world prediction, Dynamic Obstacles.

Figure 6 shows the ideal paths of the obstacles and subject for these trials. The same 100 walking frames are used for the articulated model. However the demand instant spacing is now 4 frames of separation, to highlight the effects of prediction on the system. Thus, a total of 25 data points are presented for each run. The Obstacles 1 and 2 followed their linear paths with constant velocities of $v = [100 \quad 0 \quad 100]$ *mm/s* and $v = [-100 \quad 0 \quad -50]$ *mm/s*, respectively.

All the trials, as expected, yielded results which are very similar to those in Experiment 1, confirming the robustness of the system to noise in OoI motion prediction, Figures 7 and 8.

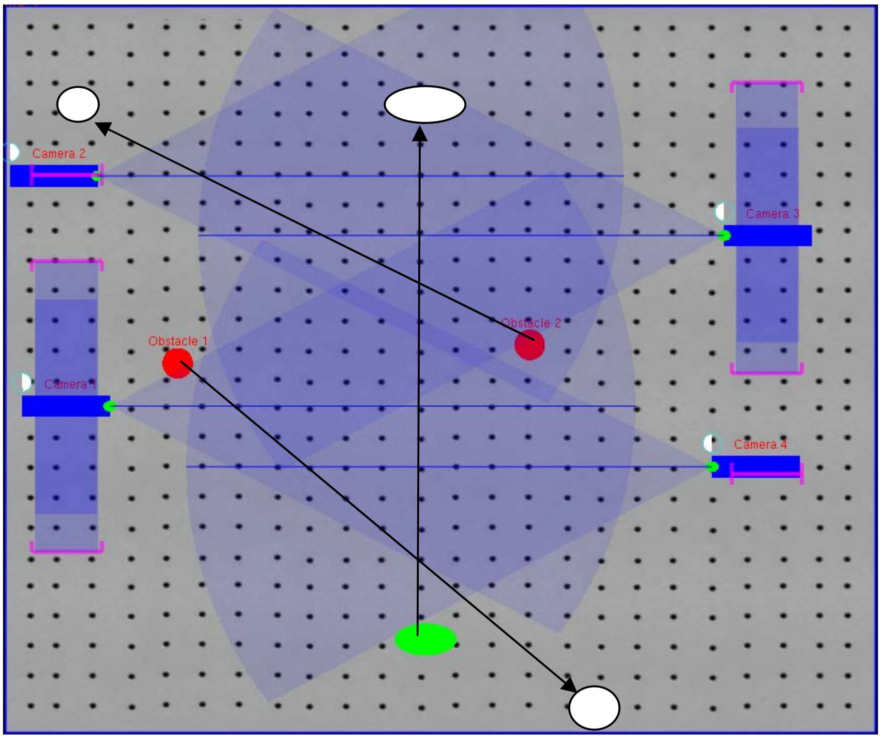


Figure 6. Initial positions and obstacle/subject paths for dynamic obstacle trials.

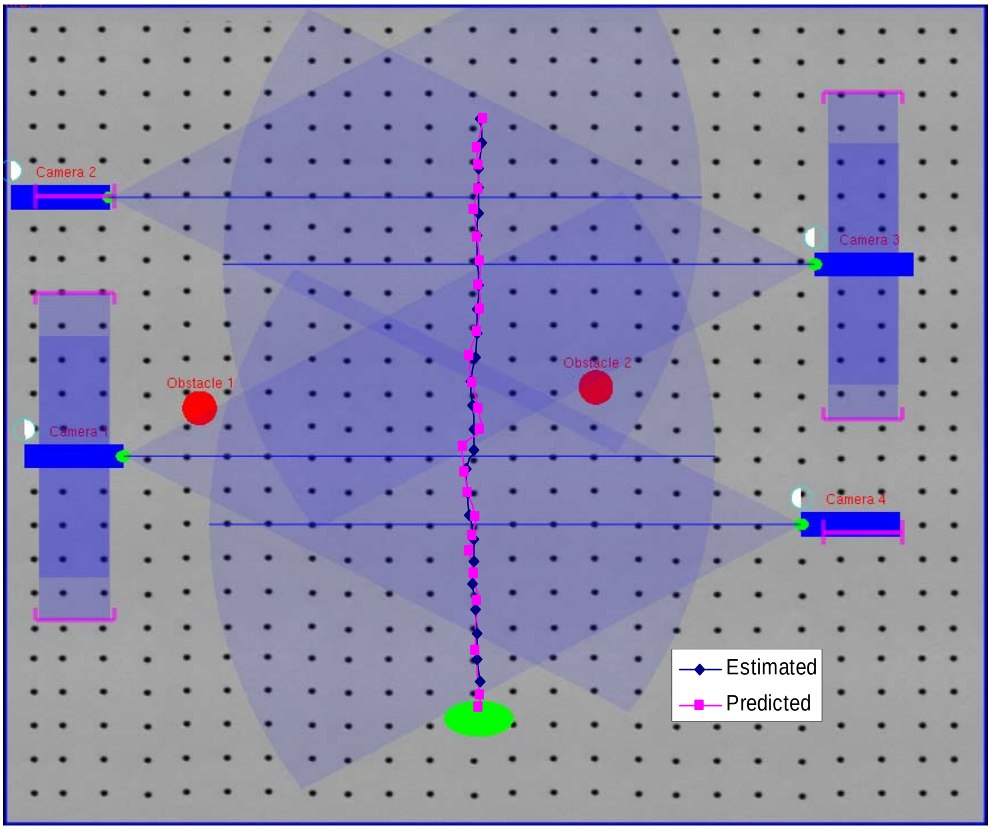


Figure 7. Subject pose estimation for Trial 3.

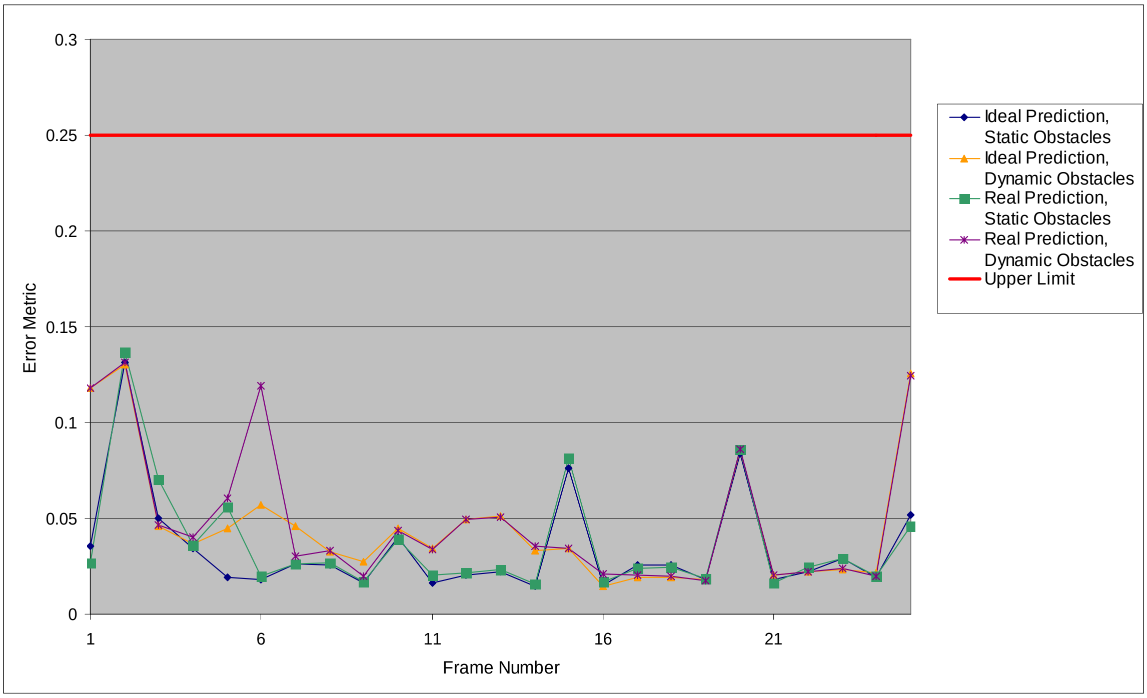


Figure 8. Error metric comparisons.

5. CONCLUSIONS

In this paper, an agent-based sensing-system reconfiguration methodology is proposed for the surveillance of time-varying-geometry objects. A central planner agent, using ranked visibility metric evaluations from multiple sensor agents, is used to select sensor assignments and poses for a system of sensors tracking a single, dynamic, and articulated subject. A set of rules, and a referee agent, ensure the correct aggregate system behavior. Experiments have shown that, overall, a tangible improvement in form-recognition performance can be attained over static sensor placement. The proposed methodology was also shown to be capable of dealing with uncertainty in target pose estimation in the presence of multiple, dynamic obstacles.

Appendix A: Form Recognition

Key interest points are recovered through a simple implementation of a color cues algorithm – each limb segment is assigned a unique color. From the center lines, 2D intersections are determined. These intersections, along with the camera calibration matrices, are then used to determine a line equation for each different view of the same key-point (e.g., in this case, the left elbow), Figure 6. A robust solver finds the intersection of these lines in 3D space and, thus, the world coordinates of the point. The first point recovered is a reference point, the head center. The other contiguous regions of the OoI are then identified, and key-points determined. A simple fitting method is used to form the feature vector via four form-invariant reference points on the OoI.

A measure of the uncertainty of the fit is defined as:

$$E \propto C_1 \sum \left(\left\| \mathbf{x} - \mathbf{x}_m \right\| \right) + C_2 n_{ua} + C_3 n_{ms} \qquad \text{(A1)}$$

For each recovered keypoint location (relative to the model reference center), ***x***, and each corresponding model location, $\boldsymbol{x_{m,}}$, C_1, C_2, and C_3 are proportionality constants, n_{ua} is the number of unassigned points, n_{em} is the number of missing points, and the sum is over all assigned points. This method of fitting is performed on each model in the database and a minimum uncertainty fit is determined (subject to some upper limit for recognition).

In order to recognize the current OoI action, a post-process method is used. The search begins by comparing each recovered form to a set of 'start poses' and 'end poses' from the database. Whenever a near match (subject to a lower limit for recognition) to both a start and end frame is found, the frames between these are time-normalized. A metric of distance to 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.

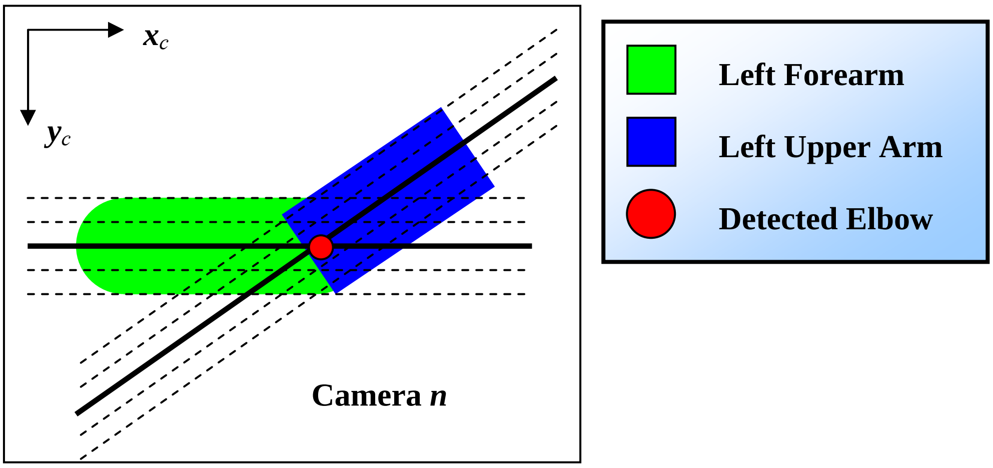


Figure A1. Simplified implementation of key-point detection using color cues.

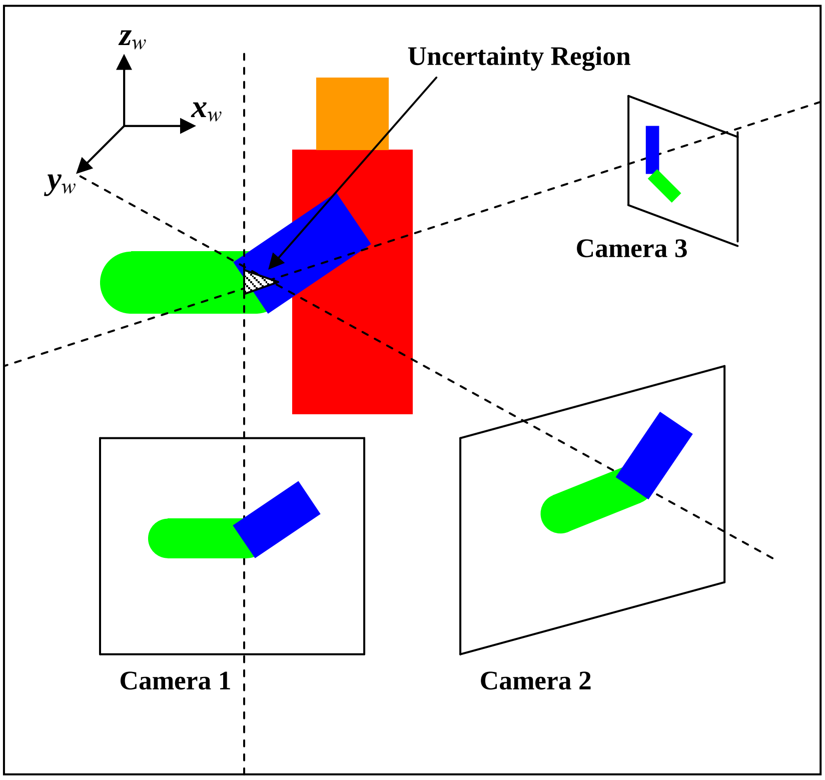
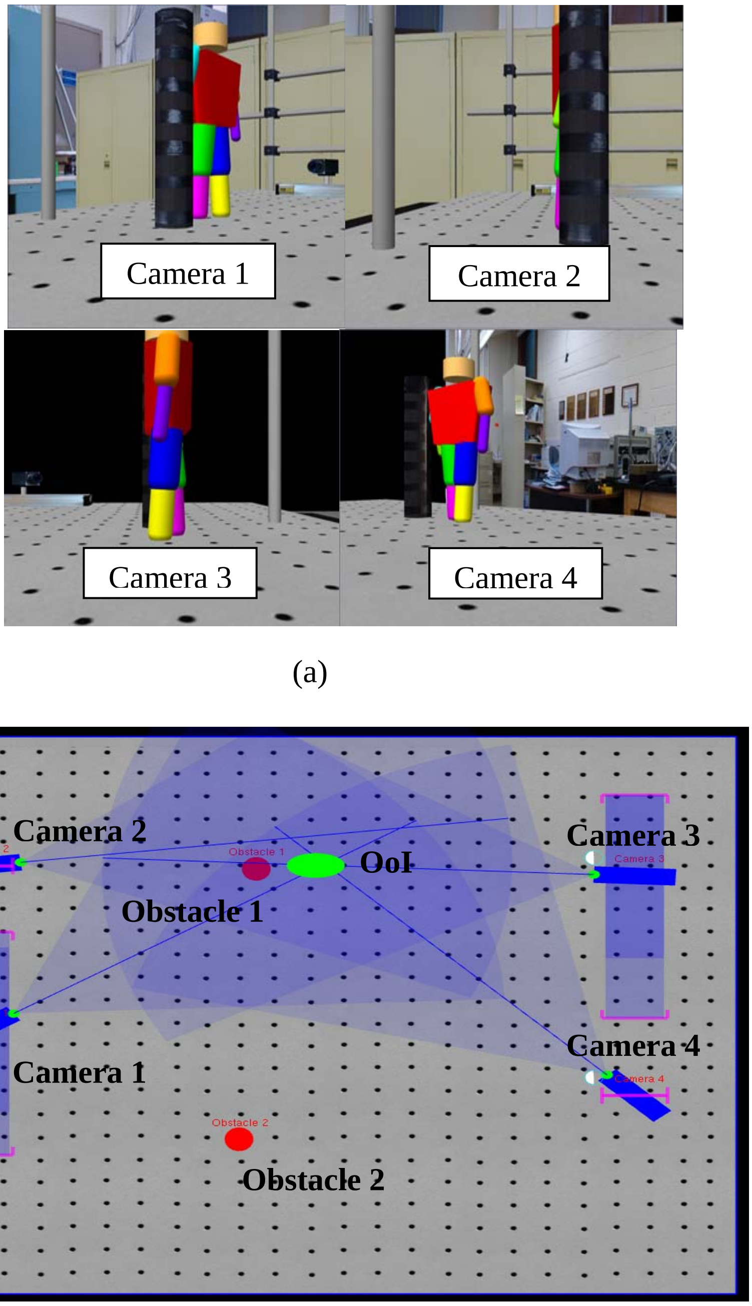


Figure A2. Implementation of algorithm to recover world coordinates of key-points.

Appendix B: An Example Frame Analysis

As an example frame, let us examine Frame 85 in Example 1 in Section 4, where a positive match was determined by the ideal algorithm, but significant error in the recovered model still exists. This is the result of the system rejecting incomplete data on one or more model sub-sections and, thus, not recovering that portion of the model. Specifically, as shown in Figure B1, the left arm of the subject is not visible. Although the images are shown without noise for clarity, all simulated images had Gaussian noise added, with distribution parameters determined from the measurement of real-world cameras.



(a)

(b)

Figure B1. (a) The four camera views and (b) sensing-system configuration at Frame 85.

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.

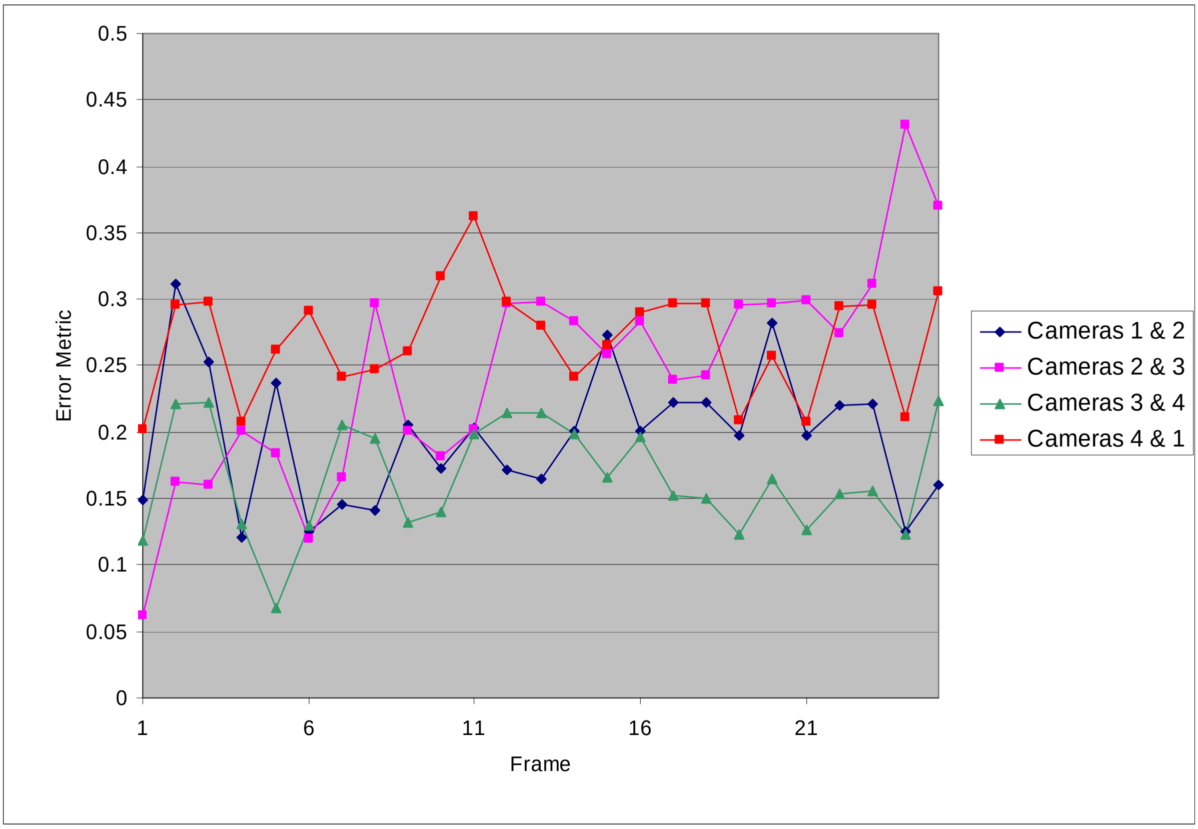


Figure C1. Comparison of error metric for four different sets of two cameras each.

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 camera pairs. Generally, those with camera pairs on the same side (1&2 or 3&4) have lower error metrics, as they are more likely to produce two views of the same object subparts. Finally, there is also a repeating pattern (with a period of approximately 2-3 frames) of an increasing, followed by a decreasing error metric spike – this is likely due to self-occlusion of the leg joint during walking. The effect of self-occlusion is greatly amplified when using a limited number of cameras. One can, thus, conclude that all viewpoints are not equal in the information they contain – while fusion from more cameras may hide this fact, it is essential that cameras be assigned to acquire as much *unique* information as possible at each instant.

Experiment – Effect of Self Occlusion

This experiment was designed to quantify the effects of self-occlusion on this particular vision algorithm. In particular, it is designed to show that without directly considering the current and predicted form of the subject, there will be partial (or full) self occlusions, and these will impact recognition performance.

The same experimental setup as above was used, with 100 frames of a linear walking motion. Both external obstacles were removed for these trials, and the reconfiguration capability of the system was limited to real-world velocity and acceleration limits. In order to simulate the effects of increasing self-occlusion, the radius of the limbs of the articulated model was increased with each successive trial. The results from the four trials are shown in Figure C2, up to a 60% increase in limb radius.

The results for this trial clearly show the negative effect of increasing self-occlusion. First of all, one can note the same periodic spikes in the error metric originally seen in the experiment above. As the limb radius increases, these spikes grow larger as more parts of the articulated model are not recovered. As such, self-occlusion can prevent model recovery and reduce performance in this manner. Also, in frames preceding and following each of these spikes, all parts of the model are considered recognized for most trials. However, let us for example consider Frames 46-49 and 51-54: For these frames, all trials show the model to be completely recovered, though one can note a clear trend of increasing error with increasing self-occlusion. Although the algorithm does not fail to recover the model, the quality of the output is clearly affected.

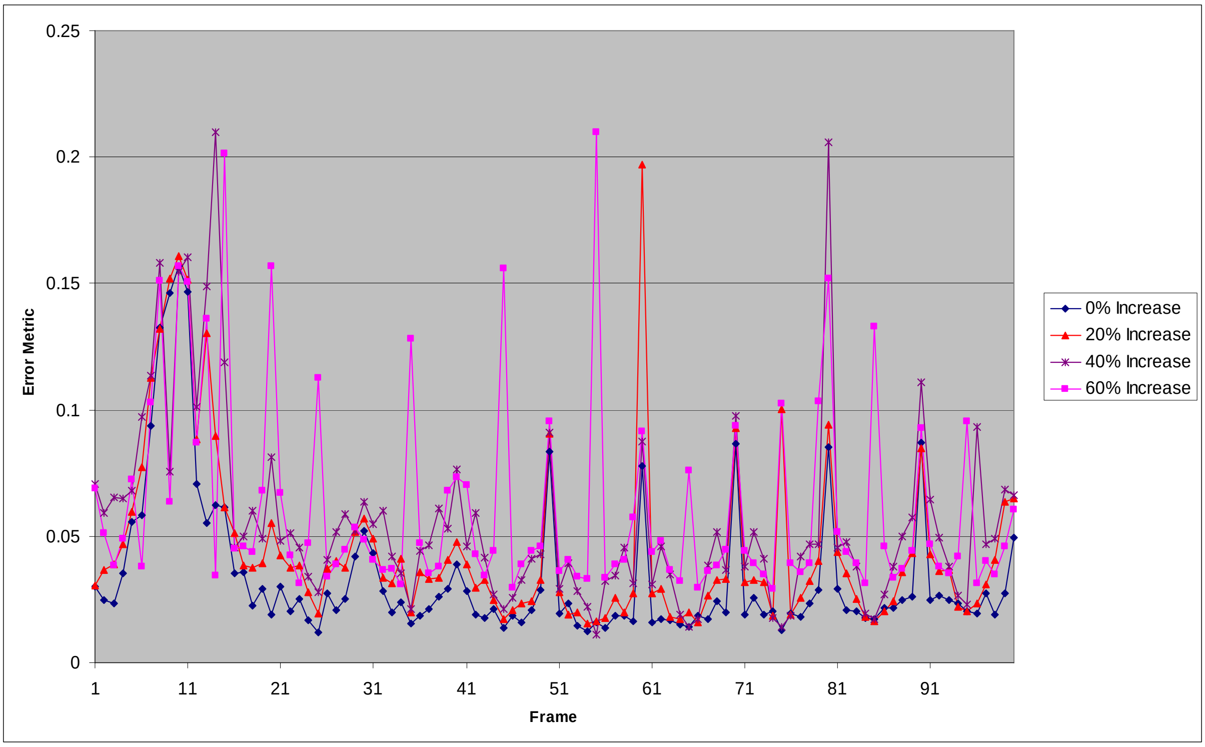


Figure C2. Comparison of error metric for varying degrees of limb size.

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.

[8] L. Hodge and M. Kamel, "An Agent-Based Approach to Multi-sensor Coordination," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 33, no. 5, pp. 648-662, Sept. 2003.

[9] D.P Anderson, "Efficient Algorithms for Automatic Viewer Orientation," *Comp. & Graphics*, vol. 9, no. 4, pp. 407-413, 1985.

[10] 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.

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

[12] D.P Anderson, "Efficient Algorithms for Automatic Viewer Orientation," *Comp. & Graphics*, vol. 9, no. 4, pp. 407-413, 1985.

[13] M.K. Reed and P.K. Allen, "Constraint-Based Sensor Planning for Scene Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1460-1467, Dec. 2000.

[14] L. Hodge and M. Kamel, "An Agent-Based Approach to Multi-sensor Coordination," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 33, no. 5, pp. 648-662, Sept. 2003.

[15] T. Urano, T. Matsui, T. Nakata, and H. Mizoguchi, "Human Pose Recognition by Memory-Based Hierarchical Feature Matching," *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 6412-6416, The Hague, Netherlands, 2004.

[16] 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.

[17] R. Murrieta-Cid, B. Tovar, and S. Hutchinson, "A Sampling-Based Motion Planning Approach to Maintain Visibility of Unpredictable Targets," *Journal of Autonomous Robots*, vol. 19, no. 3, pp. 285-300, 2005.

[18] J. R. Spletzer, and C. J Taylor, "Dynamic Sensor Planning and Control for Optimally Tracking Targets," *Int. Journal of Robotic Research*, vol. 22, no. 1, pp. 7-20, Jan. 2003.

[19] M. Kamel, and L. Hodge, "A Coordination Mechanism for Model-Based Multi-Sensor Planning," *Proc. of the IEEE International Symposium on Intelligent Control*, pp. 709-714, Vancouver, Oct. 2002

[20] J. Spletzer and C. J. Taylor, "Sensor Planning and Control in a Dynamic Environment," *Proc. of IEEE Int. Conf. Robotics and Automation*pp. 676–681, *Washington, DC,* 2002.

[21] S.G. Goodridge, R.C. Luo, and M.G. Kay, "Multi-Layered Fuzzy Behavior Fusion for Real-Time Control of Systems with Many Sensors," *IEEE Transactions on Industrial Electronics*, vol. 43, no. 3, pp. 387-394, 1996.

[22] S. G. Goodridge and M. G. Kay, "Multimedia Sensor Fusion for Intelligent Camera Control," *Proc of IEEE/SICE/RSJ Multi-sensor Fusion and Integration for Intelligent Systems*, pp. 655-662, Washington, DC, Dec. 1996.

[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.

[39] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H. Zhang. "Human Gait Recognition with Matrix Representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 7, pp. 896-903, July 2006.

[40] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, pp. 35-45, March 1960.

[41] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. of IEEE Int'l Conf. on Computer Vision*, Kerkyra, Greece, pp. 1150-1157 1999.

[42] H. Jing, C. K. Fi, and E.C. Prakash, "Component Based Human Animation Architecture – Design Issues," *Proc. of TENCON*, vol. 2, pp. 528-532, Kuala Lumpur, 2000.

[43] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, "On a Large Sequence-Based Human Gait Database," *Proc. of 4th International Conference on Recent Advances in Soft Computing*, pp. 66-72, Nottingham UK, 2002.