Model-Based Feature Tracking

Karen Chaconas
Marilyn Nashman

National Institute of Standards and Technology
Robot Systems Division
Gaithersburg, MD 20899

Abstract

The use of data-driven and model-driven processing provides us with the ability to track a moving object. We are able to update the position and orientation of the modeled object by correlating camera data with a predicted object model. This paper describes such processing and also describes sensory processing and world modeling algorithms designed to make use of a high temporal sampling rate with respect to spatial changes. The algorithms operate in real-time in a multi-processing environment.

1. Introduction

The ability to visually track an object during arbitrary motion is an important part of interacting with the environment. Humans adeptly recover three-dimensional structure of an object from its rigid-body motion in order to accomplish manipulation, locomotion, and object recognition [6]. Motion and edge information is known to be an important basis by which to recover object structure. Understanding the relative motion between an object and an observer aids not only in the recovery of object structure but also provides useful information required to perform these tasks. In order to visually track an object, the object's six-dimensional position must be rapidly updated. In robotic applications, real-time camera images provide a dense stream of data from which to extract object features and recover rigid body motion.

An object's three-dimensional structure can be reconstructed from the projection of its motion onto a two-dimensional image [22]. Two methods which measure visual motion are detection of spatio-temporal intensity changes and feature tracking [18]. The measurement of spatio-temporal

intensity changes can be accomplished using correlation models, energy filters [1] [4], or gradient techniques [2] [13].These algorithms provide a description of two-dimensional image motion as a result of changes in intensity in the image. They cannot differentiate between intensity differences due to changes in viewer motion and intensity changes produced by object motion with respect to a light source. Thus, they are not suited to the task of tracking a moving object. The class of algorithms which measure visual motion by tracking features provides a means to directly measure physical motion in the world. These algorithms, however, have the disadvantage of requiring feature correspondence between images.

By combining the feature tracking approach with model-driven techniques, the feature tracking process is constrained and the correlation of features between frames is simplified. This results in a computationally inexpensive and accurate system. This paper describes an approach designed to achieve these goals and the implementation of this approach in the Intelligent Controls Group (ICG) laboratory at the National Institute of Standards and Technology. The next section discusses model-based feature tracking methods and, in particular, the two-dimensional tracking method we use. Section 3 details the implementation of the algorithm in our lab. The fourth section quantifies the accuracy and speed of this algorithm in tracking a planar target, and the final section discusses the implications of these results.

2. Model-based Feature Tracking

Model-based feature tracking correlates image features with object model features to take advantage of model information. Correlation between extracted features and an object model can be performed in either a two-dimensional [9] or three-dimensional [12] [19] frame-of-reference. A useful survey of work done using these methods can be found in [20]. Three dimensional tracking involves comparing a set of two dimensional features extracted from an image to a three dimensional model. In the most general case, this means solving the three-dimensional recognition problem for each successive image frame. The three-dimensional position and orientation of the feature are computed by analyzing images taken at different positions. Correspondence of features be-

tween image frames is determined, and the three-dimensional feature position is computed and matched to the model. This process is time consuming computationally expensive.

A more efficient method of using a three-dimensional frame of reference for model matching involves computing only the changes in structure position and orientation between image frames [20]. The set of extracted two-dimensional features that must be matched with a three-dimensional model is thus reduced. By projecting the model into image frame coordinates, the search space is further reduced since the approximate location of each model feature in the two-dimensional image frame is known. The information extracted and used to update the model is usually quite accurate, and since all surfaces of the object model are available, problems of changing viewpoints or occlusion are handled. However, the computational complexity of these algorithms prevent their use in real-time applications.

The model-based feature tracking approach where the matching occurs in two-dimensions is a less complex alternative to the tracking problem. It assumes the ability to process a continuous sequence of two-dimensional images in real-time [9]. One or two-dimensional features such as vertices, centroids, or edge segments are extracted and matched to a two-dimensional representation of the model. Analogous to the three-dimensional case, the process is simplified by computing motion features which represent the changes in position and orientation between image frames. Since the temporal sampling rate is high, there is little change in position and orientation between successive frames and the correlation between observation and model is simplified. The two-dimensional model is continuously updated based upon the most recent observation. Tracking in two dimensions continues exclusive of additional information as long as the object motion is continuous. A model update obtained from three-dimensional information is required if there is occlusion or a change of direction causing loss of two-dimensional information. In general, two-dimensional feature tracking is an inexpensive method of correlating observations with model information and is well-suited for real-time applications [20].

The approach in our lab is based on model-based feature tracking in two dimensions. Figure 1

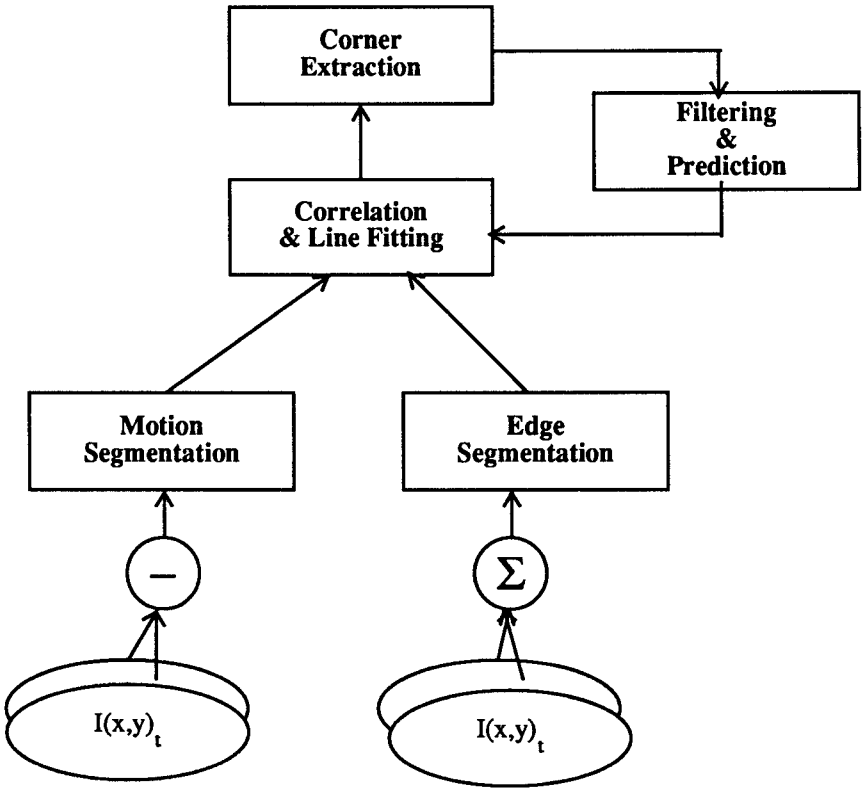

Figure 1. Model Based Feature Tracking Algorithm

depicts an overview of the algorithm. $I(x,y)_t$ refers to the intensity function at a pixel located at position (x,y) at time t. Motion and edge features from a sequence of images are correlated with model information. Model based tracking involves segmentation and correlation of observed data with model data and the prediction of the model position at the next time interval. .

During the segmentation phase, optical flow and edge orientation are extracted from an incoming sequence of images. The image flow results from changes in intensity between frames and a temporal differencing algorithm is used to measure these changes. Incoming images are smoothed using a Gaussian convolution, G*, to diminish the effects of spurious noise in the image. Two temporally-consecutive, smoothed images are subtracted from each other in order to detect any change in intensity due to motion between the frames (Equation 1). All non-moving features in the image "disappear" in this difference image since the grayscale value of a pixel in the second frame is being subtracted from the identical grayscale value in the first frame. The resulting optical flow image

$$\frac{\partial}{\partial t} I(x,y) = G^*(x,y)_t - G^*(x,y)_{t+1} \qquad [1]$$

is thresholded to produce a binary image. This operation results in a segmented scene caused by changing intensity values between successive images. Segmentation occurs whether the changing intensity is due to relative motion between the camera and the object or between the object and a light source.

Edges are also extracted during the segmentation process. Spatial orientation of edge points are computed directly from the sequence of input images. This information is used to determine the spatial orientation of the image motion points. Since the position of the edge points on the object change between frames, edge points are extracted from the sum of two temporally-consecutive, smoothed images. A two-dimensional spatial gradient operator is applied to all points (x,y) in the image. The actual direction, θ, of each point in the image is defined to be perpendicular to the direction of the gradient of the intensity function f(x,y) at that point:

$$\theta = \text{atan}\frac{(\nabla_y f)(x,y)}{(\nabla_x f)(x,y)} + \frac{\pi}{2} \qquad [2]$$

The edge extraction and the motion extraction operations are performed in parallel on the same input images.

The next step, correlation of the extracted motion points with the model edges, makes use of the segmentation information. Each motion pixel is either labelled or discarded depending on its similarity to the model. The labelling process is based on two criteria. The first criterion is the two-dimensional spatial proximity of a motion point to the model lines.The second criterion is the similarity of direction of the edge at the location of this motion with the angular direction of the model line.

The description of each line in the model includes the slope-intercept form of the line, the coordinates of the endpoints of each line segment, and the angular direction of the line. The first step

in the correlation process compares the angular direction of the model line with the edge direction of the candidate motion point to determine if the angular disparity is within an acceptable range:

$$|\theta_{model} - \theta_{data}| < \delta \qquad [3]$$

If this condition is satisfied, the distance d is computed between the point at image coordinate (x_i, y_i) and each of k lines used to define the object model using equation [4]:

$$d = \frac{A_k x_i + B_k y_i + C_k}{(A_k^2 + B_k^2)^{1/2}} \qquad [4]$$

where $(A_k x + B_k y + C_k)$ is the general form for the k^{th} line in the model.

The minimum distance between the image motion point and each of k model lines is used to determine if that point is less than a distance threshold, ζ, from the line. The point is labelled as belonging to the model line segment it best matches when both the spatial proximity and orientation conditions are satisfied.

After the motion points are segmented, a line-fitting technique is used to update the two-dimensional position of each model line. The line fitting technique uses a least-squares linear regression which minimizes the squared error in either the x or y direction. The equations that compute slope, m, and the y intercept, b, of the best fitting line through the n points (x_i,y_i) when minimizing the x direction error are:

$$m = \frac{n(\sum_i (x_i y_i)) - (\sum_i x_i)(\sum_i y_i)}{n\sum_i (x_i^2) - (\sum_i x_i)^2} \qquad [5]$$

$$b = \frac{(\sum_i y_i)(\sum_i x_i^2) - (\sum_i x_i)\sum_i x_i y_i}{n(\sum_i x_i^2) - (\sum_i x_i)^2} \qquad [6]$$

Minimizing the least square error in the x direction presents a problem as the line being fit approaches vertical. As this happens, the denominator in equation [5] approaches zero, and the fit becomes less accurate. A more accurate fit takes this into account and minimizes errors in both x and y. Comparison of the standard deviation of the x coordinates to that of the y coordinates gives a measure as to whether the line is more horizontal or more vertical. A larger standard deviation in x means the line tends toward the horizontal. When the standard deviation of the y coordinates is larger, a linear regression of x on y is used since the line is more vertical. In this case, the equation for the slope and intercept of the fitted line is given by

$$m = \frac{n\left(\sum_i y_i^2\right) - \sum_i (y_i)^2}{n\left(\sum_i x_i y_i\right) - \left(\sum_i x_i\right)\left(\sum_i y_i\right)} \qquad [7]$$

$$b = \frac{-\left(\left(\sum_i x_i\right)\left(\sum_i y_i^2\right) - \left(\sum_i y_i\right)\left(\sum_i x_i y_i\right)\right)}{\left(n\left(\sum_i x_i y_i\right) - \left(\sum_i x_i\right)\left(\sum_i y_i\right)\right)} \qquad [8]$$

If the standard deviation of x coordinates is less than a predefined threshold value, the line is considered to be vertical, and therefore the slope is undefined.

The computed lines are intersected to determine the two-dimensional corner positions of the object model. Each corner point (x_c, y_c) is computed by solving for the intersection of each set of fitted lines. The distance between the computed corners and the model corners is computed to determine the proximity of the results with the prediction. If the resulting distance is within an acceptable limit, ε, then a successful match has been detected (Equation [9]):

$$\sqrt{(x_c - x_{model\,k})^2 + (y_c - y_{model\,k})^2} < \varepsilon \qquad [9]$$

When the distance exceeds ε, tracking is lost; distances less than ε indicate tracking. When track-

ing is successful, the corners are filtered using an exponential smoothing filter [17] to predict the corner positions at the next time interval. Each corner is filtered independently of the others since the object motion isn't necessarily parallel to the image plane and the corners will move at different rates in the image plane. The filtering of each corner is done using a weighted average of the current and all past positions of that corner with exponentially decreasing weights (Equation 10).

$$(x, y)_t' = \alpha + (1 - \alpha)\ (x, y)_t \qquad [10]$$

The smoothing constant, α, is in the range $0 < \alpha \leq 1$ and, in our case, is chosen to be 0.2. The corner is filtered again using equation 11.

$$(x, y)_t'' = \alpha (x, y)_t + (1 - \alpha)\ (x, y)_t' \qquad [11]$$

After the filtered values are determined, the predicted corner position $(x,y)_{t+1}$ is computed using equation [12].

$$(x, y)_{t+1} = (2 + \frac{\alpha}{1 - \alpha})\ (x, y)_t' - (1 + \frac{\alpha}{1 - \alpha})\ (x, y)_t'' \qquad [12]$$

The predicted model information is used to segment the extracted image flow information at time t+1. Predictions are computed each execution cycle regardless of whether updated observed corners are available. This allows predicted positions to be sent to the correlation process continuously. Figure 2 shows the relative frequency with which predicted data is generated as compared to data extracted from incoming images over a period of about 1 second.

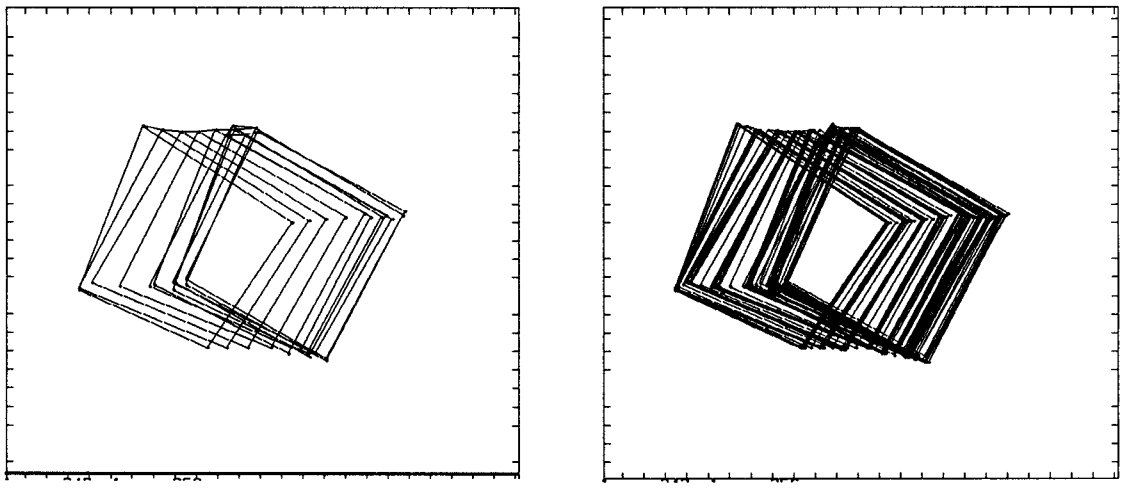

Figure 2. (a) Observed Object Positions (b) Predicted Object Positions

Since the results of the tracking algorithm are used as feedback in a real-time system, the implementation of this algorithm must be optimal. Section 3 describes the approach we use.

3. Implementation

Processing in the integrated vision testbed in the ICG laboratory is accomplished using a real-time pipelined image processor, the Pipelined Image Processing Engine (PIPE)[1] [5] and a multiprocessor system as shown in Figure 3. In this figure, the large grey rectangles represent how the software processes are distributed on this hardware. The incoming images from a CCD camera are digitized by PIPE to provide 8-bit grayscale images that are 242x256 pixels in size. The images are processed by lookup tables, neighborhood operators, and arithmetic logic units that are configured on PIPE using microcode. Smoothing, temporal integration, and edge and motion detection are performed on the grayscale images as described in equations [1] and [2]. The ISMAP stage of PIPE converts the binary motion image into a list of pixel positions segmented by the presence of motion. In addition, the corresponding edge direction values are stored in the ISMAP iconic buffer where they are memory mapped onto one of the microprocessor's memory via a specialized PIPE-VME interface board. Figure 3 displays these pipelined processes as black parallelograms.

The remaining software processes operate in real-time in the multiprocessing environment. They are implemented within the hierarchical sensory-interactive robot control system in our lab [7] [10] [15] [16] [21] that is defined in the NASA/NBS Standard Reference Model for Telerobot Control System Architecture (NASREM) [3]. Figure 3 shows the software processes using black rectangles and the data which is passed between processes as ovals. The processes are labeled according to their functional role in the NASREM architecture as sensory processing (SP) or world modeling (WM) modules.

The implementation is based on the concept of cyclically executing modules which serve as the

1. Commercial equipment and materials are identified in this paper in order to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the materials or equipment identified are necessarily the best for the purpose.

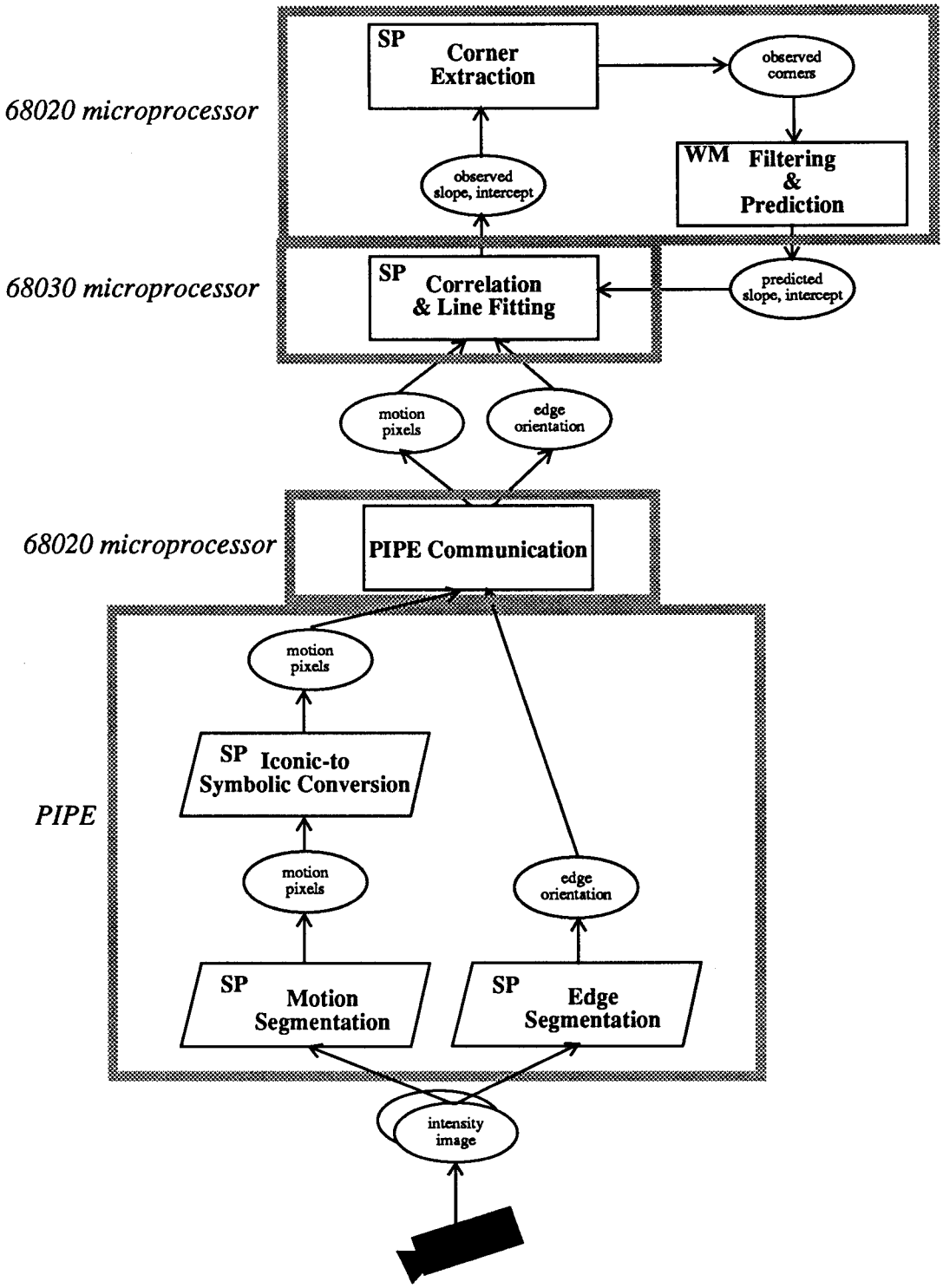

Figure 3. Implementation of Model-Based Feature Tracking Algorithm

computational units for the NASREM architecture [11]. After initialization, all computations are performed by cyclically executing processes which communicate via global read-write interfaces. Each unit acts as a process which reads inputs, performs computations, and then writes output. Such a process always reads and executes on the most current data; it does not wait for new data to arrive since reliable cyclic execution requires that a module be able to read or write data with minimal delay. Reading and writing involves the transfer of data between local buffers and buffers in global memory. System software has been written to prevent data corruption during these transfers.

Three cyclically-executing software processes execute the model-based feature tracking algorithm. These reside on two of the three microprocessor boards. The remaining software process performs communication with PIPE. The PIPE communication process is a cyclically executing process which polls PIPE status, reads the ISMAP output produced by PIPE, and writes it to the appropriate common memory locations that it shares with the segmentation process. Though the amount of data transferred is large, on the order of hundreds to thousands of pixels every 60 Hz, the direct memory accessing provides a high rate of accessibility to the symbolic data. The execution time for this process takes an average of 90 ms on a 68020 processor for about 1400 motion pixels, an average sample for our tests described in the following sections. The correlation steps described in equations [3] - [4] and the line fitting described in equations [5] - [8] are computationally intensive. The processing time for these operations depend on the number of motion points. This process requires high bandwidth, and for this reason, the correlation and line fitting process executes on a dedicated 68030 microprocessor at an average rate of 110 ms per execution cycle. Since the execution time is greater than that of the PIPE communications process, it always has new data at the beginning of its execution cycle. The resulting lines are written to a common memory buffer shared with the process which computes the corners of the observed object. This process computes the object corners in 2.1 ms and then updates the buffer shared with the filtering and prediction process. The filtering and prediction in equations [10] - [12] are used to obtain the predicted model and execute in 3.1 ms. These processes are combined on one board, a 68020 processor since

their total execution time is small compared to the other processes.

4. Results

To determine how accurately this model-based feature tracking algorithm can track an object, a set of experiments was run for the case of simple translational velocity. In these experiments, a planar, rectangular object was mounted on a pendulum as shown in Figure 4. The pendulum is released at different heights to provide image motion at differing velocities. The only a priori knowledge of the object is the image positions of the four corners at an arbitrary point during the path of the pendulum. These points establish a crude model and are necessary to establish when the object has initially been matched. Once the observed data matches the initial model, the object is tracked by the single-camera vision system in the manner previously described.

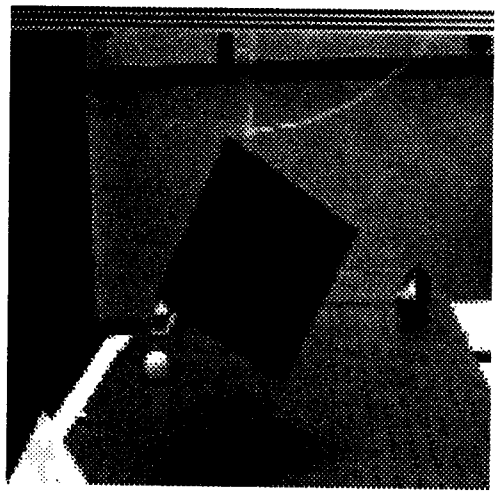

Figure 4. Experimental Scenario

In order to quantify the accuracy with which the model predicts the position of the observed data, the difference between the model corners and the actual corners is computed. This difference is used to determine the accuracy of the fit between the predicted and the observed data at varying object velocities. The accuracy of the tracking algorithm is also affected by the threshold parameters used. The threshold used in equation [4] controls the labelling of motion pixels and determines

how closely the data can be correlated to the model. The threshold used to match the model corners to the observed corners in equation [9] determines how closely the model must match the observed data before tracking is lost. Our experiments consisted of three cases of varying object velocity. Velocity is measured in the image plane as the distance that an object feature moves between camera frames. The velocities tested are 1.1 pixels per frame, 3.9 pixels per frame, and 4.5 pixels per frame. For each velocity, the correlation threshold is tested at four values, 3 pixels, 5 pixels, 10 pixels and 15 pixels. In addition, the corner matching threshold is tested at 32 pixels, 26 pixels and 20 pixels. In all, twelve sets of data were collected for each velocity. Each set of data consists of 200 error measurements.

The threshold parameters were varied for three different object velocities measured in the image plane. Table 1 summarizes the results of experiments for the three velocities at varying correlation thresholds. The corner threshold remains constant at 20 pixels. Tables 2 and 3 are similarly organized except that the corner thresholds are set to 26 pixels and 32 pixels, respectively. Continuous tracking was performed successfully for all cases over a period of 200 iterations. By comparing the tables, it can be seen that the threshold used to determine successful tracking described in equation [9] does not play a significant role. Table 1 shows that for each object velocity, the mean error increases as the correlation threshold increases. This is a result of the fact that as the distance threshold is relaxed, there is a greater chance of misclassifying motion pixels. This effect is noticeable at faster velocities since there are more motion pixels available for processing. Also it can be seen that at distance thresholds of 3 and 5 pixels, the tracking error decreases with increasing velocity. This can be attributed to the value chosen for the smoothing constant α described in equations [10] - [12] which provides predictions more closely matching the observed data at a velocity of 4.5 pixels per 60 Hz. It is not clear that this trend would continue for higher velocities using the same smoothing constant. At distance thresholds of 10 and 15 pixels the tracking error increases as the velocity increases. This is caused by a greater number of motion pixels being present at higher velocities compounding the effect of the relaxed threshold. Similar conclusions can be drawn by analyzing Tables 2 and 3.

	Distance Threshold for Correlation			
Velocity	**$\zeta = 3.0$**	**$\zeta = 5.0$**	**$\zeta = 10.0$**	**$\zeta = 15.0$**
1.1	0.302	0.338	0.335	0.407
3.9	0.045	0.096	0.096	1.151
4.5	0.009	0.024	1.174	1.368

Table 1. Mean Data Error Using Corner Threshold $\varepsilon = 20$

	Distance Threshold for Correlation			
Velocity	**$\zeta = 3.0$**	**$\zeta = 5.0$**	**$\zeta = 10.0$**	**$\zeta = 15.0$**
1.1	0.301	0.313	0.357	0.407
3.9	0.110	0.108	0.052	0.052
4.5	0.009	0.005	1.177	1.368

Table 2. Mean Data Error Using Corner Threshold $\varepsilon = 26$

	Distance Threshold for Correlation			
Velocity	**$\zeta = 3.0$**	**$\zeta = 5.0$**	**$\zeta = 10.0$**	**$\zeta = 15.0$**
1.1	0.302	0.313	0.348	0.407
3.9	0.031	0.109	0.097	1.160
4.5	0.001	0.003	1.174	1.273

Table 3. Mean Data Error Using Corner Threshold $\varepsilon = 32$

5. Conclusions

The two-dimensional tracking algorithm we implemented combines data-driven and model-driven processing. It allows us to correlate camera data with a predicted object model in order to update the position and orientation of the object. By taking advantage of a high temporal sampling rate with respect to spatial changes, we are able to successfully track an object in real-time. The experiments conducted show that we are able to successfully track an object moving at a velocity of 4.5 pixels per camera frame. We plan to continue the experiments on model-based feature tracking with a single camera and to extend the scope of our algorithms to include processing on a stereo set of cameras [8]. Knowledge of the two dimensional position of the same feature as viewed from two cameras will enable us to determine the position and orientation of the object in world space using range from triangulation [10]. This capability will also allow us to supply feedback information to the manipulator system to aid in tasks involving tracking or grasping a moving part.

6. References

[1] Adelson, E. H., J. R. Bergen, "Spatio-temporal Energy Models for the Perception of Motion," *Journal of the Optical society of America A*, Vol. 2, No. 2, February, 1985, pp. 284-299.

[2] Albus, James S., Tsai-Hong Hong, "Motion, Depth, and Image Flow," *IEEE Robotics and Automation Conference*, Cincinnati, OH, May 13-18, 1990.

[3] Albus, J. S., H. G. McCain, R. Lumia.,"NASA/NBS Standard Reference Model for Telerobot Control System Architecture (NASREM)", NIST Technical Note 1235, Gaithersburg, MD, July, 1987.

[4] Allen, Peter K., "Real-time Motion Tracking Using Spatio-Temporal Filters," *Proceedings of the DARPA Image Understanding Workshop*, Palo Alto, TX, May 23-26, 1989.

[5] Aspex, Inc., "PIPE--An Introduction to the PIPE System", New York, 1987.

[6] Bolles, Robert C., H. Harlyn Baker, David H. Marimont, "Epipolar-Plane Image Analysis: An Approach to Determining Structure from Motion,"*International Journal of Computer Vi-*

sion, Vol. 1, 1987, pp. 7-55.

[7] Chaconas, K., M. Nashman, "Visual Perception Processing in a Hierarchical Control System", NIST Technical Note 1260, Gaithersburg, MD, March, 1989.

[8] Chaconas, K., "Range from Triangulation Using An Inverse Perspective Method to Determine Relative Camera Pose," NIST Internal Report 4385, Gaithersburg, MD, August, 1990.

[9] Crowley, James L., Patrick Stelmaszyk, Christopher Discours, "Measuring Image Flow by Tracking Edge-Lines," *Proceedings of the 2nd International Conference on Computer Vision*, 1988, pp. 658-664.

[10] Fiala, J., "Manipulator Servo Level Task Decomposition," NIST Technical Note 1255, NIST, Gaithersburg, MD, October, 1988.

[11] Fiala, J. "Note on NASREM Implementation," NIST Internal Report 89-4215, Gaithersburg, MD, December, 1989.

[12] Gennery, Donald B., "Tracking Known Three-Dimensional Objects," *Proceedings of the National Conference on Artificial Intelligence*, Pittsburg, PA, August 18-20, 1982, pp. 13-17.

[13] Horn, B. K. P., B. Schunk, "Determining Optical Flow," *Artificial Intelligence*, Vol. 17, 1983, pp. 185-203.

[14] Jenkin, M., J. K. Tsotsos, "Applying Temporal Constraints to the Dynamic Stereo Problem," *CVGIP*, 33, 1986, pp. 16-32.

[15] Kelmar, L. "Manipulator Servo Level World Modeling," NIST Technical Note 1258, NIST, Gaithersburg, MD, March, 1989.

[16] Lumia, R., Fiala, J., Wavering, A., "The NASREM Robot Control System and Testbed," 2nd Intl. Symp. on Robotics & Automated Manufacturing, Albuquerque, NM, November, 1988.

[17] Montgomery, D. C., L. A. Johnson, and J. S. Gardiner, "Forecasting & Time Series Anal-

ysis," Second Edition, McGraw-Hill, New York, 1990.

[18] Spetsakis, Minas E., John Aloimonos, "Closed Form Solution to the Structure from Motion Problem from Line Correspondences," *Proceedings of the National Conference on Artificial Intelligence*, 1987, pp 738-743.

[19] Thompson, D. W., J. L. Mundy, "Model-based Motion Analysis - Motion from Motion," *Robotics Research: The Fourth International Symposium*, R. C. Bolles and B. Roth, eds., The MIT Press, Cambridge, MA, 1988, pp. 229 - 235.

[20] Verghese, Gilbert, Charles R. Dyer, "Real-time Model-Based Tracking of Three-Dimensional Objects," Computer Sciences Technical Report #806, University of Wisconsin - Madison, November, 1988.

[21] Wavering, A. "Manipulator Primitive Level Task Decomposition," NIST Technical Note 1256, NIST, Gaithersburg, MD, October, 1988.

[22] Waxman, Allen M., Kwangyoen Wohn, "Image Flow Theory: A Framework for 3-D Inference from Time-Varying Imagery," LSR-TR-1, Boston University, January, 1986.

[23] Waxman, Allen M., Jian Wu, F. Bergholm, "Convected Activation Profiles: Receptive Fields for Real-Time Measurement of Short-Range Visual Motion," *International Conference on Computer Vision*, April, 1988.