Recognizing Objects From Curvilinear Motion

Tal Arbel, Frank P. Ferrie, and Marcel Mitran
Department of Electrical and Computer Engineering
McGill University, Center for Intelligent Machines
Montréal, Québec CANADA H3A 2A7
{taly,ferrie,mmitran}@cim.mcgill.ca

Abstract

This paper introduces an object recognition strategy based on the following premises: i) an object can be identified on the basis of the optical flow it induces on a stationary observer, and ii) a basis for recognition can be built on the appearance of flow corresponding to local curvilinear motion. Unlike other approaches that seek to recognize particular motions, ours focuses on the problem of recognizing objects by training on an expected set of motions. A sequential estimation framework is used to solve the implicit factorization problem, which is itself simplified in that the task is to discriminate between different objects as opposed to recovering motion or structure. Training is accomplished automatically using a robot mounted camera system to induce a set of canonical motions at different locations on a viewsphere. Experimental results are presented to support our contention that the resulting motion basis can generalize to a fairly wide range of motions, leading to a practical method for recognizing moving objects.

1 Introduction

In this paper, we consider the problem of recognizing a moving object from the optical flow pattern it induces by its motion in front of a stationary camera. Our approach is appearance-based, i.e., appearance manifolds are created for each object in a database by inducing flows corresponding to the set of expected motions. The resulting set of optical flow fields is used to build an appearance manifold using standard methods [10, 9]. What makes this problem difficult is the confounding of motion, structure, and imaging geometry such that the appearances of different objects are often indistinguishable from one another. Similar problems are inherent in recognizing particular motions, e.g., gestures [8, 5, 7, 4], where the problem is often made tractable by the dominance of the motion component of the optical flow field. In fact, this is precisely what we wish to avoid in the object recognition context, hence some control of imaging parameters is required to ensure that a reasonable component of the flow corresponds to structure. Unfortunately, this leads to a much more difficult factorization problem.

To get around this difficulty we invoke a temporal variant of the general position assumption:

A1. The probability of several objects giving rise to similar appearances diminishes with shifting viewpoint.

In the case of recognition a reasonable strategy is to maintain a list of plausible hypotheses and accumulate evidence for each one through a sequence of observations, which we

BMVC 2000 doi:10.5244/C.14.77

refer to as *temporal regularization*[1]. We will show later on how this can be precisely formulated in probabilistic terms using Bayesian chaining strategy to effect regularization. Experimental results will confirm the validity of the general position assumption by showing that in most cases the list of plausible hypotheses quickly diminishes to a single confident assertion.

The foregoing begs the important question of whether it is possible to create an appearance manifold for sufficiently general motions in the first place. Clearly this is possible in restricted contexts such as gesture recognition. What we seek is a means of generating a set of canonical motions, i.e. a *motion basis*, in different viewpoints that can generalize to a sufficiently wide range of appearances to be of use in practical recognition tasks. Again, we invoke a second general assumption regarding the motion of physical objects:

A2. The general trajectory of a solid object moving in 3-D can be locally approximated by a curvilinear arc tangent to the viewing direction.

This suggests that a motion basis can be constructed by presenting an object to a stationary observer and sweeping it in different directions in the plane tangent to the current viewing direction. The basis is constructed be repeating this process for each of the object's canonical viewpoints. (Equivalent motions can be induced by moving the camera about a stationary observer.) Additionally we need to assume that the object moves at constant velocity relative to the camera, which is in general not a severe constraint.

To test this hypothesis we have constructed in our laboratory a robot-mounted camera system that can generate the requisite sensor trajectories on a viewsphere surrounding the object of interest. This apparatus is used to automatically generate motion bases (training) for a set of approximately 25 standard household objects. On-line, objects from the database are presented to a stationary camera by subjecting each to a set of curvilinear motions generated by a precessing pendulum (using the object as the mass). This approach allows for a wide range of sample trajectories that are clearly outside of the motion basis used for training. In addition, we have also performed testing on free rotations and translations generated by hand. The results presented later in Section 4 show that the system identifies the object correctly from the on-line (novel) trajectories 80% of the time. This lends support to our contentions regarding the generalizability of our motion basis and disambiguation of competing hypotheses via temporal regularization.

The remainder of the paper is organized as follows. Section 2 describes the generation of the appearance manifolds from the set of canonical motions and particular details about the assumptions employed. Section 3 describes the strategy used to compute support for object hypotheses as probabilities given the appearance manifold and instantaneous optical flow measurements acquired on-line. It further shows how evidence for different hypotheses can be accumulated over time using Bayesian chaining and how this serves to regularize the interpretation of the instantaneous flow fields. Next in Section 4, we describe a set of experiments used to confirm our key assertions and present the results obtained from a series of on-line experiments. Finally, we conclude in Section 5 with a discussion of our results and future work.

2 Building an appearance flow manifold

Flows corresponding to the expected motions of a mobile object moving about a stationary observer are induced during training by moving the observer (a television camera) on a tessellated viewsphere surrounding the object according to the the following set of constraints:

1. **Camera Constraints.** Camera to object distances are bounded and scaled orthographic projection is assumed.

[1] Another example of a sequential strategy can be found in [11].

2. **Motion Constraints.** The same motion model can be used to account for an object moving about a fixed observer and vice-versa provided that rotations are limited to axes that are approximately parallel to the image plane.
3. **Motion Decomposition.** Assumption A2 applies equally to the case of the mobile observer.

Constraint 1 is required to ensure that a sufficient component of the optical flow magnitude is due to the structure of the object. Together with Constraint 2, it then becomes possible to associate the magnitude of the optical flow vector with distance to the camera. Taken as a whole, the magnitude of the optical flow field can be viewed as a rough *kinetic depth map*. This is by no means adequate for quantitative recovery of structure due, among other things, to confounding with motion, but it can provide a basis for recognition according to general position assumption A1. Constraint 3 implies that the structure component induced by camera motion about a stationary object is indistinguishable locally from that induced by the motion of the object about a stationary observer. This permits the construction of a motion basis using the more tractable approach of a mobile observer moving about a stationary object. The question of how to generate this basis from sensor motion is discussed next.

By assuming that objects can be differentiated on the basis of their local structure over time, the range of motions that needs to be generated during training is reduced significantly (Assumption A2). However this range is still considerable encompassing 4 degrees of freedom: viewsphere position, direction of motion (component parallel to the plane tangent to the surface), height above the surface, and curvature of the arc (direction orthogonal to surface). Prior knowledge about how objects are likely to interact with the observer can be used to further restrict the range of motions that need to be sampled. For example, for the experiments presented in this paper, the object was kept at a fixed canonical pose with respect to the camera (i.e. the camera was kept upright) during the acquisition of training images. Each object was sampled using only 184 trajectories corresponding to 2 orthogonal sweeps at a fixed distance from the object with the same radius of curvature as the surrounding viewsphere at each of 92 viewsphere directions. As the experimental results presented later in Section 4 will show, this basis is sufficient to generalize to a very broad range of novel motions.

Each of the 184 trajectories gives rise to a distinct motion sequence, $\mathbf{s}_{i,j,t}$, where the indices i, j, and t are used to reference a particular object, trajectory, and image within the sequence respectively. Figure 1 shows a sequence of l images resulting from motion in a vertical arc on a viewsphere surrounding one of the test objects (a horizontal arc can be seen as well). An optical flow algorithm is used to estimate for each $\mathbf{s}_{i,j,t},\ t = 1 \ldots l$, a vector field, $\vec{\mathbf{v}}_{i,j,t},\ t = 2 \ldots l-1$, corresponding to the optical flow field induced by the camera motion. Only the magnitudes of $\vec{\mathbf{v}}_{i,j,t}$ are of interest here (for the reasons cited earlier). For the particular optical flow algorithm chosen [3], three images were required to estimate a single optical flow vector. As three samples along each trajectory were sufficient to characterize each curvilinear sweep, each object i gives rise to a set of 184 scalar images. For brevity we refer to the latter as flow images. Standard PCA (principal component analysis) techniques [10] are used to construct an eigenbasis for the entire training set of 4600 flow images.

Let each of these images be represented by an $m \times n \times 1$ vector, $\mathbf{d}_j$, with m and n corresponding to the image dimensions. We refer to the latter as an image flow vector. A representation for each object is subsequently constructed by projecting each of its corresponding N image flow vectors $\{\mathbf{d}_j\}|_{j=1\ldots N}$ onto the eigenbasis and representing the resulting set with an appropriate parameterization. We refer to the latter as the appearance flow manifold. For the experiments presented in this paper, the manifold is parameterized with a multivariate normal distribution of the form $N(\mu_{m_i}, \mathbf{C}_{T_i})$, where the sample mean,

μ_{m_i}, and sample covariance $\mathbf{C}_{T_i}$, are estimated in the usual manner.

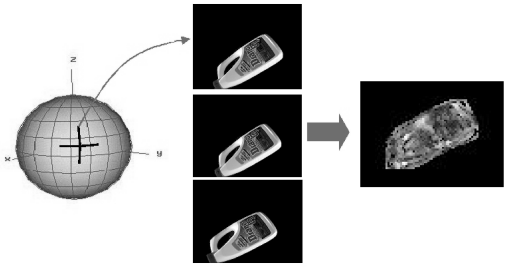


Figure 1: Flow image acquisition during training.

3 Sequential recognition strategy

On-line, a sequence of images is gathered by moving an unknown object in front of the camera via a set of curvilinear movements. Let $\mathbf{d}$ correspond to a single image flow vector computed from this sequence. The first step in the recognition process is to compute the support for each object hypothesis, $O_i|_{i=1..K}$, i.e. the probability that the single image flow vector, $\mathbf{d}$, corresponds to each object, represented by $p(O_i|\mathbf{d})$. Using standard Bayesian techniques leads to the following solution for each $p(O_i|\mathbf{d})$ [1, 2]:

$$p(O_i|\mathbf{d}) \propto p(O_i)\ N(\mu_{m_i}, \mathbf{C}_{T_i})|_{\mathbf{m}=\mathbf{m}(\mathbf{d})} \quad i = 1..K, \tag{1}$$

where $p(O_i)$ is the prior probability of the object hypothesis, and $N(\mu_{m_i}, \mathbf{C}_{T_i})|_{\mathbf{m}=\mathbf{m}(\mathbf{d})}$ is the multivariate normal distribution representation for the object hypothesis, evaluated at the projected parameters of the measurement, $\mathbf{m} = \mathbf{m}(\mathbf{d})$.

As each of the image flow vectors is computed from a set of intensity images, evidence can be inexpensively and efficiently accumulated on the level of the probabilities over time, by using a Bayesian chaining strategy that assigns the posterior probability distribution at time t, $p(O|\mathbf{d}_t)$, as the prior at time $t+1$. In this fashion, probabilistic evidence is cascaded until a clear winner emerges. Substituting the posterior density function derived from one view (derived in Equation (1)) as the prior for the next view leads to the following updating function for $p(O|\mathbf{d})$ [1, 2]:

$$p(O|\mathbf{d}_{t+1}) \propto \sum_{i=1}^{K} p(O_i|\mathbf{d}_t)\ N(\mu_{m_i}, \mathbf{C}_{T_i})|_{\mathbf{m}=\mathbf{m}(\mathbf{d}_{t+1})}\ \delta(O - O_i). \tag{2}$$

The hypothesis is that confounding information in the optical flow signatures can be resolved by accumulating support for different object hypotheses over a sequence of observations in this manner according to Assumption A1. Presenting the system with all prior evidence should resolve ambiguities and lead to a winning solution in a short number of views. The entire system can be seen in Figure 2.

4 Experiments and results

A set of recognition experiments was devised to test the extent to which the motion basis trained on can generalize to a wider range of motions. Two sets of experiments were performed: (1) database objects were presented to a stationary camera through a set of human-induced pendulum motion sequences, and (2) database objects were swept in front of the camera by hand. Through application of the sequential recognition strategy, the system was shown to converge to a correct assertion in terms of its MAP (maximum *a posteriori*) solution in the majority of cases.

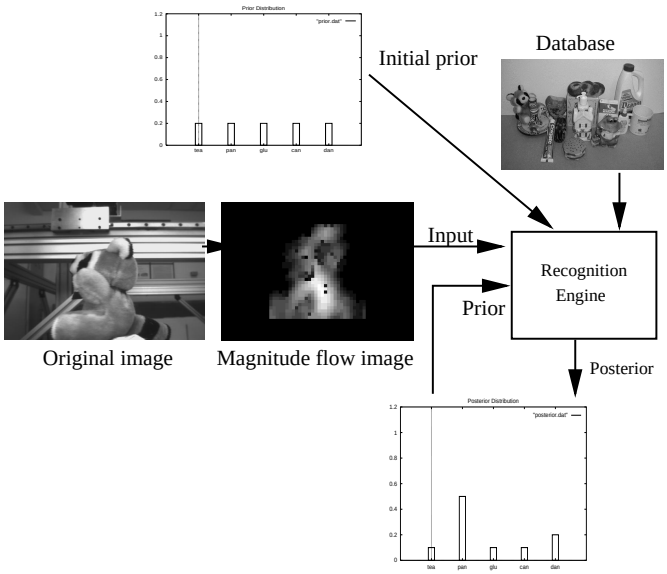


Figure 2: Sequential recognition system.

4.1 Building an appearance flow manifold

For each of these experiments, the system was trained on a set of 25 household items. A control system was developed in order to gather images at precise locations on the viewsphere. The objects were placed on a rotary table at a fixed distance from a CCD camera affixed to the end-effector of a gantry robot. The distance was fixed in order to account for size differences in cases of objects with similar shapes. With this configuration in place, all positions along the latitude and longitude of the viewsphere were accurately reachable. The viewsphere was tessellated into coarse segments of equal area. For the purposes of training, a black cloth was placed around the rotary table so as to focus the optical flow computations on the object of interest.

At each position on the viewsphere, a local description of flow was attained through a short, rectilinear basis, computed by moving the robot arm along an arc in the horizontal and vertical directions. During each sweep, three images of the object were used to compute an image flow vector by using a strategy as in [3]. During the entire training procedure, the camera was kept upright. The optical flow was instrumental in locating and centering the object of interest within the image (thus achieving position normalization). Temporal scaling was performed by normalization of the optical flow magnitudes to values between 0 and 255. This ensured that speed did not affect the appearance of the 3D structure.

The entire set of image flow vectors was parameterized using PCA techniques, where 20 eigenvectors were sufficient to account for the majority of the variance in the training data. The training images were then projected onto the resulting eigenspace and probability distributions were generated for each object in the database as was described earlier.

4.2 Recognizing objects based on novel motion sequences

For the first set of recognition experiments, the robustness of the approach was tested by moving objects along motions sequences that varied significantly from those trained on using calibrated equipment. Here, a human was asked to move a database object in

front of a stationary camera along arbitrary curvilinear motion sequences that included both translation and rotation components. In order to remove the possibility of artifacts such as a person's fingers or arms appearing in the images, objects were suspended from transparent wires and were swung along pendulum-like motion sequences in front of the camera. The motion was therefore comprised of both a rotation component about the wire, as well as a rotation about the relatively far pivot point to which the wire was attached. This motion consisted roughly of a translation in front of the camera. The experimental setup can be seen in Figure 3.

Figure 3: Setup for first set of experiments consisted of a person swinging an object in front of a stationary camera.

The context of these experiments demonstrates the benefits of using optical flow as the input to the recognition system. First, the optical flow algorithm was required in order to locate the moving object, and to center it within the tracking window. Second, using differential properties was essential as the lighting conditions varied considerably from those of training. For simplicity, the object was moved in front of the same black background as in the training setup. However, in the next set of experiments, objects were moved in more arbitrary scenes, and the optical flow algorithm was used to extract the object from the background using motion segmentation.

In order to examine the effects of violating several motion constraints from training, the following set of experiments was performed. Five database objects were moved along two different pendulum motions: (a) swinging parallel to the image plane, (b) swinging perpendicular to the image plane. The radius of the pendulum was approximately 3 feet, and the diameter of the arc it cut was approximately 8 inches. The first set of experiments introduced a translation component to the motion, and a variation from its canonical camera pose in the training set. In the second set of experiments, the constraint that fixed the camera-to-object distance was violated. Five trajectories for each type of pendulum motion were performed for each object. The system acquired up to 16 images of the object for each trajectory. An example of a sequence of images acquired from the squirrel moving in a pendulum motion towards the camera while translating can be found in Figure 4.

In order to determine a suitable convergence criterion, we derive a metric that predicts the likelihood of ambiguous recognition results as a function of the measurement based on Shannon's entropy [6]:

$$H(P(O|\mathbf{d})) = \sum_i p(O_i|\mathbf{d})\ log\ \frac{1}{p(O_i|\mathbf{d})}, \tag{3}$$

which is a measure of the ambiguity of the posterior distribution produced by a recognition experiment. Higher entropies reflect greater ambiguity. Using this metric, an on-line

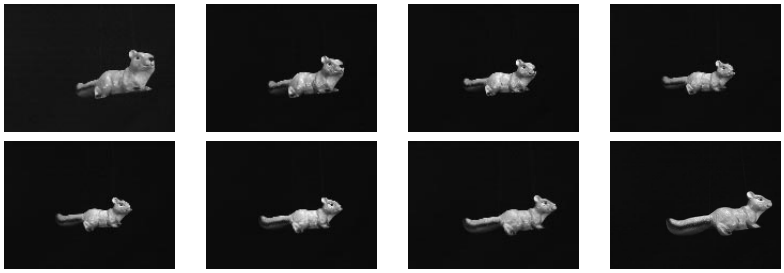


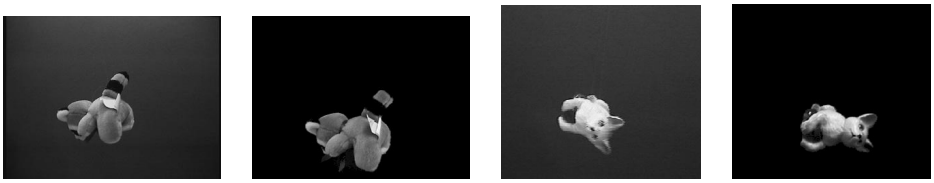
Figure 4: Typical sequence of images of object moving in pendulum motion towards camera while translating. Every second image in the trajectory is displayed.

entropy threshold establishing convergence was arbitrarily set to 1×10^{-6}. The percentage of correct MAP results at convergence for each type of experiment can be seen in Table 1.

Object	Cat	Teddy	Fish	Panda	Squirrel
Pend1	60%	100%	60%	60%	80%
Pend2	100%	80%	60%	100%	100%
Total	80%	90%	60%	80%	90%

Table 1: Recognition results in terms of MAP assertions at convergence for each type of trajectory: *Pend1*: swinging parallel to the image plane, *Pend2*: swinging perpendicular to the image plane.

The results indicate that the system was able to correctly identify the object from novel motion sequences 80% of the time. The system was particularly adept at recognizing objects from pendulum motions perpendicular to the image plane (i.e. Pend2), converging to the correct solution 88% of the time. This type of experiment violated the scale constraint in that the objects differed in size from the training set. Figure 5 demonstrates this effect through an example depicting the panda swinging in a pendulum motion away from the camera (one image in the sequence can be seen in Figure 5(a)) . The system was able to recognize the object despite substantial variations in scale from the training set. The closest training image can be seen in Figure 5(b).



(a) (b) (c) (d)

Figure 5: (a) Panda swinging in pendulum motion perpendicular to image plane during experiment Pend2. (b) Image of panda in database. (c) Cat swinging in pendulum motion parallel to image plane during experiment Pend1. (d) Image of cat in database.

The system had more difficulty recognizing the objects based on pendulum motions parallel to the image plane. This was not due the translation motion itself but rather due to sensitivities to variations in camera pose. Figure 5(c) show an image from a sequence where the system converged to the wrong object. Here, one can see that the pose varied substantially from the training image in (d). This implies that motions such as these

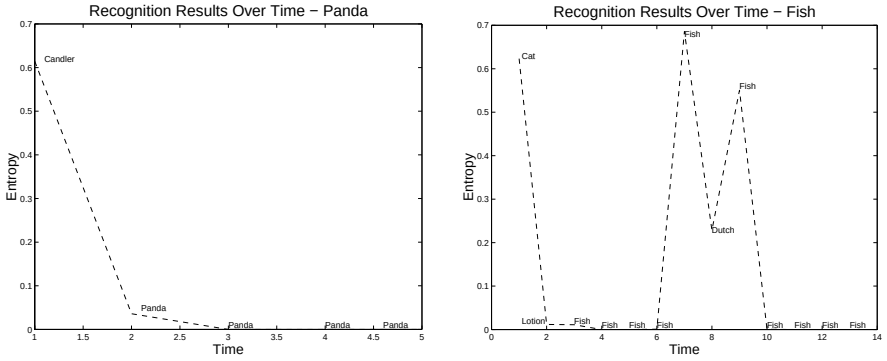
require a more comprehensive training set, one that accounts for variations in pose due to rotations in the image plane.

The relatively poor fish results can be explained through examination of Figure 6(a)–(c), where images extracted from a typical fish motion sequence are shown. Notice that, unfortunately, only the profile of the fish were exposed, providing very little structural information to the system (which often confused it with a tube of toothpaste). The database image depicting the most informative viewpoint for recognition is shown in Figure 6(d). Examining these images, it is noteworthy that recognition system was able to converge to the correct result 60% of the time.

(a) (b) (c) (d)

Figure 6: (a)–(c) Sequence of images of the fish. (d) Most discriminant view of the fish in the database.

Of course, the MAP results only depict the results at convergence. One interesting result to examine is how the on-line entropy of the distribution computed in Equation 2 varies over time. Figure 7(a) plots the entropy over time for a panda bear, pushed along a pendulum parallel to the image plane. At the beginning of the trajectory, the panda was maximally far from its training position and the system had high confidence in the wrong object. However, as it moved away from this extremal point, confidence in the correct grew and remained high, until convergence to the correct solution was attained.



(a) (b)

Figure 7: On-line entropy of recognition results over time for (a) panda bear, (b) fish swinging on a pendulum towards the camera. The MAP result can be seen above the curve at each iteration. Notice the variation in entropy as the fish swings through the scene.

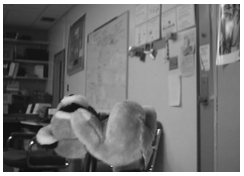
In general, pendulum motions posed a difficulty for the recognition task in that the system began in its least favorable position. This implied that it was either at the lim-

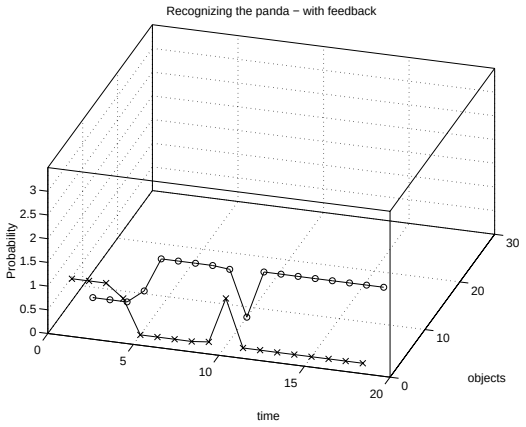
it of its ability to recognize based on variations in proximity from the camera (i.e. too close or too far from the camera), or based on variations from the camera pose trained on. As the object swung towards the center, the system generally recovered. However, Figure 7(b) illustrates a case where the object moved beyond the window of acceptable motions/positions, resulting in oscillations between hypotheses prior to convergence. However, the system converged to the correct solution, primarily due to the power of temporal regularization at quickly removing the hypotheses with little evidence.

4.3 Recognizing objects from hand-held motion sequences

A second set of experiments was performed in order to test the system's robustness at adapting to different environments. Here, a person was asked to sweep an object in front of a stationary camera with some arbitrary arc-like motion (roughly within the permissible set), at approximately the same distance from the camera as in training. The benefits of using flow are made evident as both lighting and background varied substantially from training. In this experiment, motion segmentation was used to perform figure/ground separation as well as to localize the flow to a specific window.

Initial recognition results with hand-moved objects are quite promising. The case of the panda bear can be see in Figure 4.3(a). The sequential recognition results, seen in Figure 4.3(b), show the probabilities in competing object hypotheses varying over time. The non-zero probabilities in each object are represented by curves in the graph. Here, the system started with a high belief in an object other than the panda (represented by the X's). The evidence in the panda (represented by O's) grew quickly and, despite a brief drop in confidence when presented with significant evidence in the wrong object (at time interval 10), the system remained confident in the panda throughout the experiment. All other objects were eliminated from its set of hypotheses. Several other motion sequences lead to constant high beliefs in the panda after one or two iterations. These preliminary results illustrate the robustness of the strategy at successfully identifying unknown objects through previously unseen motion sequences, generated from a wide range of sources, under varying environmental conditions.





(a) (b)

Figure 8: (a) Panda bear moved in front of the camera by a person, (b) sequential recognition results.

5 Conclusions

In this paper, we have examined the problem of recognizing objects based on optical flow signatures induced by their motion with respect to a stationary camera. We have shown how to generate a motion basis that generalizes to permit recognition based on a wider set of motions. The factorization problem was shown to be partially alleviated through a temporal regularization strategy that uses Bayesian methods to accumulate evidence in the competing hypotheses over several viewpoints until a clear winner emerges. Empirical evidence supports the hypotheses through tests with sample trajectories that were clearly different from those comprising the training motion basis: Training was accomplished using a camera mounted onto a robot arm to acquire a set of motions on a viewsphere about the object, and recognition was performed based on human-generated motions by a precessing pendulum (Figure 3) and by free hand-generated rotations and translations. The results confirm our hypotheses in that system converged to the correct object identity in the majority of the cases, even with the compounded effect of partial occlusion.

To our knowledge, this paper presents the first attempt to examine the robustness of a recognition strategy based on motion signatures. Future work will entail a more comprehensive set of experiments, systematically examining the effects of violating each of the constraints imposed during training. The results of this work indicate that the recognition with an even wider range of motions should be feasible.

References

[1] T. Arbel and F. P. Ferrie. Viewpoint selection by navigation through entropy maps. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 248–254, Kerkyra, Greece, Sept 20-25 1999.

[2] T. Arbel and F. P. Ferrie. Recognizing objects by accumulating evidence over time. In *Fourth Asian Conference on Computer Vision*, page to appear, Taipei, Taiwan, Jan 8-11 2000.

[3] S. M. Benoit and F. P. Ferrie. Monocular optical flow for real-time vision systems. In *Proceedings of the 13th International Conference on Pattern Recognition*, pages 864–868, Vienna, Austria, 25–30 Aug. 1996.

[4] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parametrized models of image motion. *Int. Journal of Computer Vision*, 25(1):23–48, 1997. Also found in Xerox PARC, Techinical Report SPL-95-020.

[5] A. F. Bobick and J. W. Davis. An appearance-based representation of action. Technical Report 369, MIT Media Lab, February 1996. As submitted to ICPR 96.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 1991.

[7] T. J. Darrell and A. P. Pentland. Recognition of space-time gestures using a distributed representation. Technical Report 197, M.I.T. Media Laboratory Vision and Modelling Group, 1992.

[8] N. T. M. Watanabe and K. Onoguchi. A moving object recognition method by optical flow analysis. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 1 of *A*, pages 528–533, Vienna, Austria, Aug 1996. International Association for Pattern Recognition.

[9] S. K. Nayar, H. Murase, and S. A. Nene. *Parametric Appearance Representation in Early Visual Learning*, chapter 6. Oxford University Press, February 1996.

[10] M. Turk and A. P. Pentland. Eigenfaces for recognition. *CogNeuro*, 3(1):71–96, 1991.

[11] S. Vinther and R. Cipolla. Active 3d object recognition using 3d affine invariants. In J.-O. Eklundh, editor, *In Proc. 3rd European Conf. on Computer Vision*, volume II of *LNCS 801*,, pages 15–24, Stockholm, Sweden, May 1994. Springer–Verlag.