Interactive Visual Dialog

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

Abstract

In this paper we propose a paradigm called the Interactive Visual Dialog (IVD) as a means of facilitating a system's ability to recognize objects presented to it by a human. The presentation centers around a supermarket checkout scenario in which an operator presents an item to be tallied to a stationary television camera. An active vision approach is used to provide feedback to the operator in the form of an image (or images) depicting what the system thinks the operator is most likely holding, shown in a viewpoint that suggests how the object should next be presented to improve the certainty of interpretation. Interaction proceeds iteratively until the system converges on the correct interpretation. We show how the IVD can be implemented using an entropy-based gaze planning strategy and a sequential Bayes recognition system using optical flow as input. Experimental results show that the system does, in practice, improve recognition accuracy, leading to convergence to a correct solution in a minimal number of iterations.

1 Introduction

In this paper, we explore how a visual dialog between a person and a machine can be used to facilitate interpretive tasks such as recognizing objects. The presentation will center around a supermarket checkout scenario in which an operator sweeps an object to be identified in front of a stationary television camera. Instead of a bar code, the system must recognize the object from the sequence of images generated as a result of its motion in front of the camera. This task is difficult because imaging conditions cannot be precisely controlled (e.g. object pose, distance to camera, illumination, etc.) so that it is likely that the system will fail to correctly identify the object in a significant number of instances. In previous work, [1] we showed how an active vision approach could be used to solve a similar recognition problem in the context of a mobile robot in a stationary environment. There, ambiguity of recognition in the form of entropy measures was used to calculate gaze trajectories that minimized the uncertainty of interpretation. The present problem is more difficult for two reasons i) motion is induced by a human instead of a robot and ii) the requested motions must somehow be communicated to the human.

Since the variability of human motions cannot be controlled, they must be treated as noise and accounted for by the recognition process. Rather than attempting to base recognition on a single measurement, a hypothesis filtering strategy is applied to the entire sequence so that evidence for different object hypotheses can be accumulated over time

BMVC 2000 doi:10.5244/C.14.76

until a clear assertion can be made. The strategy is implemented using a sequential Bayes estimator [1] and is also required to regularize the recognition process. This is because recognition is based on the appearance of the optical flow patterns induced by the object's motion in front of the camera. Although somewhat analogous to approaches employed for gesture recognition [10, 4, 7, 3], confounding of motion and structure makes it difficult (if not impossible) to make a confident assertion from a single viewpoint. However the likelihood of different objects giving rise to similar appearances from different viewpoints diminishes in the number of viewpoints (temporal regularization). Hence it becomes possible to make a confident assertion once sufficient data have been gathered.

Generating the appearance manifold for each object [12, 11] presents another problem as it must account for all of the expected motions. We argue elsewhere that this process can be made tractable by invoking constraints on the local appearance of rigid-body motion and camera to object distance [1]. However it is not always feasible to expect human generated motions to lie within the permissible range. This is where the active vision approach comes in by means of a paradigm we call the Interactive Visual Dialog (IVD). In addition to the appearance manifold, we construct a second representation called an entropy map which relates the ambiguity of recognition associated with different viewpoints (or poses of the object). The idea behind the IVD is for the system to present to the operator an ordered set of images corresponding to the most plausible hypotheses for what is being presented to the system, but shown in a different pose - the one that suggests how the object should next be presented to improve the certainty of interpretation. The operator then moves the object along a trajectory that mimics the optimal motion sequence. The process repeats until the confidence of interpretation exceeds a prescribed threshold, generally within 1 or 2 iterations.

The remainder of the paper is organized as follows. Section 2 describes the appearance manifold used in our scheme, along with the details of how it is constructed and the recognition strategy used to identify it using subspace methods. The sequential recognition approach used to regularize the interpretations follows next along with an overview of the entropy maps which serve as the basis for active vision. Section 3 lays out the structure of the IVD and the details of our particular implementation, with experimental results presented next in Section 4. Finally we conclude in Section 5 with a brief discussion about the results and pointers to further work.

2 Representation and Recognition

For motions in relatively close proximity to a stationary camera, as is the case here, a significant component of the induced flow will correspond to the shape of the moving object provided that there is a sufficient rotational component. For the scenario depicted in Section 1, we can ensure that the following constraints are met: (a) The distance between camera and object is bounded, and (b) rotations are limited to axes that are approximately parallel to the image plane. Objects are swept in front of the camera along a shallow arc with rotations about the wrist, in quite natural fashion. Under these conditions, kinematic depth effects will modulate the magnitudes of the corresponding optical flow vectors according to the shape of the object in motion. The idea is to use this signature as the basis for recognition. One can invoke a general position assumption to argue that the likelihood of several objects giving rise to the same sequence of signatures over different viewpoints will diminish in the number of viewpoints.

2.1 Appearance manifold

This suggests that an appearance-based scheme could be used in the following manner. Since the class of motions is restricted, it is feasible to consider training on a set of expected motions for each object in the database. By adding the further assumption that motion can be locally partitioned into a set of curvilinear trajectories relative to the viewer, it becomes possible to consider an automated training process. Figure 2(a) shows the robot controlled camera system we use to automatically generate precise trajectories on a tessellated viewsphere surrounding each object. A complete basis for each object is generated by moving the camera to each canonical viewpoint and generating a sequence of curvilinear arcs that typify the motion of the object relative to the camera [1]. The resulting image sequences are then fed to an optical flow algorithm [2] yielding a second sequence corresponding to the time-varying optical flow field. Only the magnitudes are retained, for the reasons cited earlier. Let $\mathbf{x}$ correspond to the $m \times n \times 1$ vector corresponding to a single instance of the flow field magnitude, where m and n are the image dimensions. For brevity we will use the term "flow image" to refer to $\mathbf{x}$ since the latter refers to a scalar field. The resulting set of flow images, acquired from all viewpoints for each object, is used to determine a basis for representation using Principal Components Analysis (PCA) [12].

Next, an appearance manifold is built for each object by projecting its corresponding $\mathbf{x}_j$ onto the PCA basis and parameterizing the resulting set with an appropriate functional representation. Let $\mathbf{m}$ denote the parametric representation in this basis. A multivariate normal distribution with density $p(\mathbf{m}|O_i)$ was found to be sufficient for purposes of recognizing the test objects in our database. The manifold represents the physical theory predicting possible variations in parameters given each object in the database.

2.2 Sequential recognition

On-line, a human operator moves the the object in front of a camera according to the constraints outlined earlier. For each of the flow image in the sequence generated, the recognition strategy computes a degree of likelihood in matches with each of the objects in the database. This leads to the formulation of a Bayesian recognition strategy whose goal is to represent the posterior beliefs over the entire set of n object hypotheses, $\{O_i\}$ where $i = 1 \ldots n$, given a single flow image, $\mathbf{x}$, by a posterior probability distribution of the form $P(O|\mathbf{x})$, with discrete (conditional) probability density function $p(O_i|\mathbf{x})|_{i=1\ldots n}$. In order to attain this goal, the image, $\mathbf{x}$, corresponding to the unknown object is projected onto the basis determined during training, resulting in the parametric description $\mathbf{m}_x$. Using standard Bayesian techniques to determine the data support for each object hypothesis gives:

$$p(O_i|\mathbf{x}) \propto \; p(\mathbf{m}_x|O_i)\, p(O_i), \quad 1 \ldots n \tag{1}$$

where $p(O_i)$ defines the prior probability for each object hypothesis, O_i, $p(\mathbf{m}_x|O_i)$ is the multivariate normal distribution derived during training evaluated at the location in space defined by $\mathbf{m}_x$ (For details on the recognition strategy, see [1].). The result is a discrete conditional probability density function describing the belief in each of the models in the database, given the flow data.

However, recognition from a single optical flow image can be ambiguous and even erroneous. We therefore formulate the problem as a sequential estimation problem, where

a more robust solution is attained by accumulating evidence in the various object hypotheses over time, as each of the flow images in the sequence is presented to the system during motion. This is accomplished efficiently at the level of the probabilities, by using a Bayesian chaining strategy that assigns the posterior probabilities at time t, $p(O_i|\mathbf{x}_t)$, as the priors at time $t+1$:

$$p(O_i|\mathbf{x}_{t+1}) \propto\ p(O_i|\mathbf{x}_t)\ p(\mathbf{m}_{x\,t+1}|O_i), \quad 1 \ldots n \tag{2}$$

where $\mathbf{x}_t$ is defined as the data set at time t, and $\mathbf{m}_{x\,t+1}$ is the parametric description of the measured flow, $\mathbf{m}_x$, at time $t+1$. Using this framework, a human operator can move an object in front of a camera, and the system will accumulate evidence in the competing hypotheses until a satisfactory confidence level is attained.

2.3 Building entropy maps

In the context of an automated supermarket checkout scenario, it is essential that the recognition strategy minimize the chances of the system of arriving at an incorrect recognition result. In addition, time is limited, and the goal is to converge in the shortest possible number of steps. For these reasons, we propose a strategy that takes maximal advantage of a priori information available in order to attain a fast and reliable on-line solution. Specifically, we propose building *entropy maps* off-line during training to relate recognition ambiguity to viewing position. Once a map is built for each object in the database, the system can store the locations that are maximally informative in terms of disambiguating between the objects in the database. This information can then be made available to an operator on-line to aid in recognition.

To build these maps, we first derive a metric based on Shannon entropy [6] to predict the likelihood of ambiguous recognition results as a function of viewing position:

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

which relates the ambiguity of the posterior distribution produced by a recognition experiment. Entropy maps are then built off-line, for each object in the database as follows:

1. During training, each optical flow measurement, $\mathbf{x}$, is stored along with its coordinates of acquisition on the viewsphere which, for this type of measurement, refers to three parameters: latitude, longitude and relative angle of motion between camera and object.
2. Recognition is then performed on each training measurement, resulting in the association of the posterior distribution, $P(O|\mathbf{x})$, to each coordinate.
3. The entropy for each measurement, $H(P(O|\mathbf{x}))$ is computed and stored at its associated coordinate. In this context, this implies that several entropy values are stored at every (latitude, longitude) position, each associated with a different relative motion.

The resulting entropy map is smoothed so that the minimal entropy location on this map will correspond to an optimal location which is stable with respect to localization errors. The final map provides a *quantitative* prediction of the level of difficulty of recognizing each object in an on-line experiment. In contrast to human-generated aspect graphs,

e.g., [8, 9, 5]), by linking location and discriminability using entropy maps, a set of such characteristic views can be automatically generated off-line. In the next section, we will show how these maps can be very informative in the context of planning gaze for object recognition.

3 Interactive visual dialog

With the entropy maps built off-line, the question is how a human operator can make maximal use of this prior information on-line during recognition experiments. The proposed strategy works as follows: As the object is moved, a sequence of flow images is presented to the recognition engine, which accumulates evidence in the various hypotheses over the entire set. At the termination of a particular sweep at time t, the system is prompted for the top k most likely object hypotheses. This information is easily provided by extracting the top k *maximum a posteriori* (MAP) solutions corresponding to $p(O|\mathbf{x}_t)$. These most likely estimates are subsequently used to select the entropy maps to be used for planning the next best view. Associated with each entropy is the location and motion that corresponds to the most discriminant viewpoint of the corresponding object. This information corresponding to the top k hypotheses is then provided to the operator visually, in the form of a sequence of images depicting the objects moving in at their optimal viewpoints.[1] The operator, having located the object among the the set, moves the object along a similar trajectory at the appropriate position. This can be verified visually by examining the live video input image on the computer screen. Should the object of interest not be among the choices presented by the system, the operator has the choice of selecting the next k hypotheses, or simply moving the object along a new arbitrary trajectory. The system iterates until a sufficient level of confidence in one of the hypotheses is attained, as determined by its on-line entropy level. Over time the expectation is that confidence in an incorrectly chosen hypothesis will decrease as further evidence is uncovered. A flowchart of the entire system can be seen in Figure 1.

4 Experimental Results

The IVD framework described in this paper was tested through a series of real recognition experiments, where the task of the system was to recognize an object in a minimal number of iterations, based on a series of interactions with a human operator. We will illustrate the system's ability to (a) resolve recognition ambiguities resulting from tests with single flow images by accumulating evidence over time, and (b) have the operator plan the next gaze position and corresponding motion trajectory during on-line experimentation based on precomputed entropy maps.

4.1 Building entropy maps

Off-line during training, images of 25 household products were gathered at equally spaced locations around a coarsely tessellated viewsphere (see Figure 2(b)). Specifically, each object was placed on a rotary table. At each position on the viewsphere, a gantry robot

[1] Alternately, the system can present a *series* of optimal locations and motions provided this information was stored *a priori*.

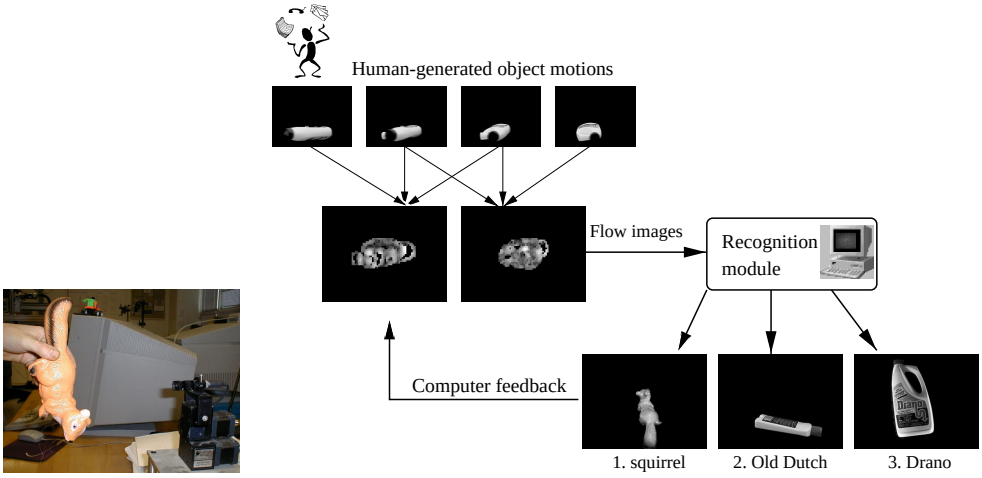


Figure 1: IVD system. (a) Setup for data acquisition, (b) system overview.

arm moved along a horizontal and along a vertical arc at fixed distances ($\approx 50cm$) from the object. A CCD camera, mounted on its end-effector (see Figure 2(a)), gathered three images in sequence along each trajectory from which optical flow was computed (using a strategy as in [2]). This served to create a local basis for flow. The expectation was that other on-line motions could be inferred from this basis. Speed normalization was achieved by normalizing the optical flow magnitudes to lie between 0 and 255. Flow was used to localize the object of interest within the images.

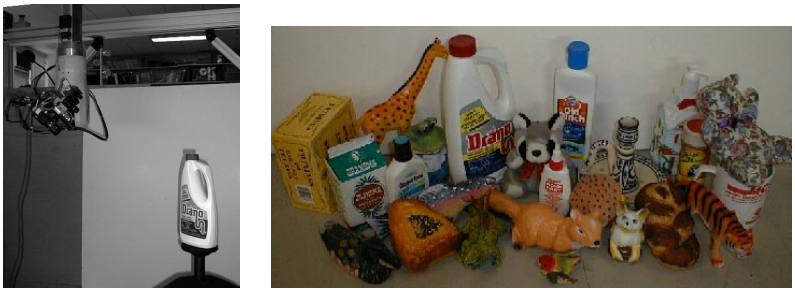


Figure 2: (a) Training Setup, (b) Database of Objects.

A low dimensional basis for flow was determined by using standard PCA techniques [12]. Empirically, it was found that 20 eigenvectors were sufficient to represent the images. Each image was then projected onto this eigenspace, and the multivariate normal distributions were computed.

Off-line, entropy maps were built from the flow images gathered. Figure 3 shows images of an object from the database and the corresponding entropy map taken from two different camera viewpoints. Each tile corresponds to a particular camera view of the object at the origin. The tiles are shaded in accordance with their entropy values:

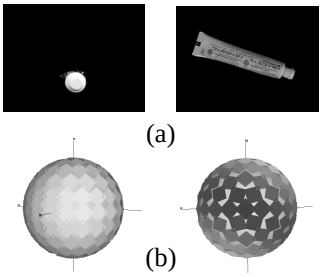


Figure 3: (a) Images of a toothpaste tube, (b) corresponding entropy maps seen from two viewpoints. The system chose the right view as the most informative (seen with darker shading on the map), and the left view as a relatively bad one (lighter tiles). This corresponds to an intuitive notion of "good" and "bad" views.

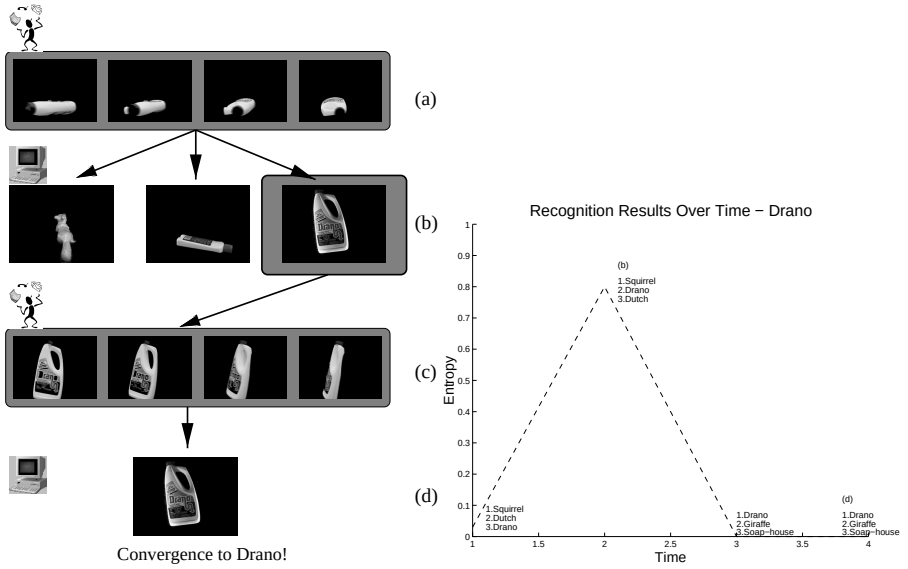
from low entropy (dark) to high entropy (light). Only the best entropy result (among all those generated from different movements at that location) is shown at each location. The system located the best viewpoint for identification of this object, one that was maximally far from the most ambiguous ones. Notice that the entropy maps match an intuitive notion of viewpoint ambiguity.

4.2 Interactive visual object recognition

Next, the entropy maps were used in a series of real, on-line experiments with a human operator to test the system's ability to converge to a fast and accurate recognition result based on the proposed IVD system. On-line, the operator was asked to move an object in front of a stationary camera along some curvilinear trajectories, loosely complying with the set of motion and camera constraints adhered to during training. This implied that the object was placed at approximately the same distance from the camera as in the training setup, and was rotated about the wrist where the axes of rotation were parallel to the image plane. Care was taken to avoid the possibility of fingers appearing in the image. Figure 1(a) shows the setup used for these experiments.

In order to maintain consistency, four grey-scale images were acquired during each trajectory. Given that three images were required to compute the flow, this implied that two iterations of recognition were performed during each trajectory. At this point, the system displayed the motion sequences corresponding to the top k object hypotheses. If the operator located the object of interest among these, an attempt was made to mimic this motion sequence. Part of the challenge was to attain quick success despite the facts that (a) the motion sequences varied considerably from those trained on, and (b) the resulting motion couldn't possibly match the optimal trajectory exactly as they were generated by a human. The system was said to have converged when an on-line entropy reached an arbitrarily set value of 1×10^{-6}.

Figure 4 illustrates an example of a typical result of experiments with objects from the database. Here, the system was presented with a non-informative viewpoint of the Liquid Drano bottle. The operator identified the bottle among the top three hypotheses presented on the screen, and attempted to move the bottle in the most informative manner. The result was that the system converged to the correct solution in the next iteration. Figure 4(e) shows the corresponding on-line entropy values as the object was moved in



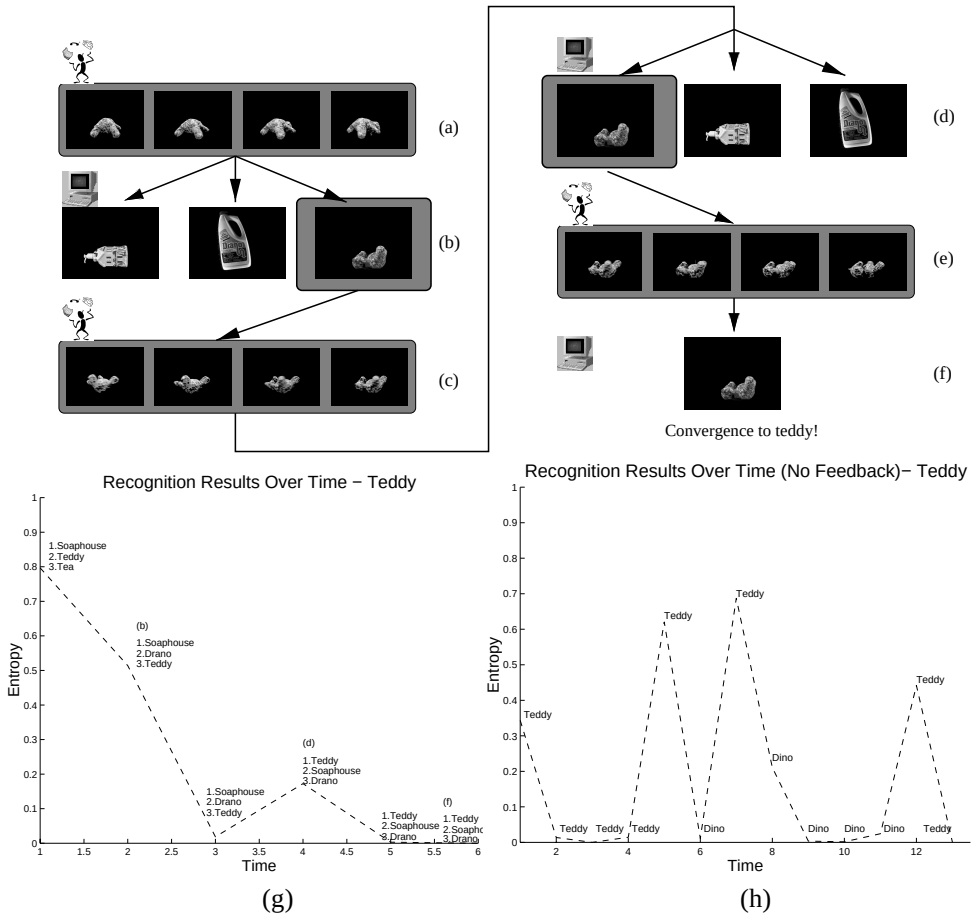
Above one can see results of an IVD experiment with the Drano bottle: (a) First motion sequence generated by operator. (b) Top 3 object hypotheses generated by system ordered from most (left) to least (right) likely. The best locations and motions are displayed to the operator (Here, only one image in the sequence is shown.) (c) The operator located the object of interest and performed the second motion sequence. (d) The system converged to the correct object. (e) On-line entropy of recognition results over time. The top 3 object hypotheses are displayed at each recognition iteration, with a letter above the iteration corresponding to the appropriate stage on the left (Recall that each trajectory consisted of 2 iterations of recognition.)

Figure 4: (a)–(d) IVD experiment with Drano bottle, (e) on-line recognition results.

front of the camera. Notice that the system converged after the second trajectory (Recall that each trajectory consisted of two iterations of recognition).

Figure 5 illustrates an example of the system working in the case of the teddy bear. The first trajectory in (a) presented a rather uninformative viewpoint of the object, resulting in the system having the most confidence in the wrong model. However, the correct object was among the top three in (b). The operator then performed a motion sequence in (c) that was close to the one suggested for the bear, resulting in it moving to the position of most likelihood in (d). However, the motion did not match the suggested one closely enough to lead to convergence in one step. This is realistic given that the motion was induced by a human operator. The operator then moved the object in a motion sequence that was close to the one suggested in (e), and the system converged to the correct solution in (f). The corresponding on-line entropy values for this experiment can be found in (g).

The advantages of the IVD system can be fully appreciated by comparing these results to cases where no feedback was provided to the human operator. Figure 5(h) illustrates the results of an experiment where the operator moved the teddy bear in front of the camera with no feedback from the system as to where and how to proceed. The operator was instructed to begin the experiment in roughly the same pose as in the experiment shown in Figure 5(a)-(g). By examining the entropy over time, one can see that the system took several more iterations to converge. In fact, the system was presented with viewpoints of the object that rendered it indistinguishable from the dinosaur doll. This emphasizes



(g) (h)

Figure 5: IVD experiment: recognition of teddy bear. From (a)–(f), one can see the IVD system. Displayed below it are the on-line entropy results over time for the teddy bear with (g) feedback,(h) no feedback (only the MAP is displayed here). Notice that the system took longer to converge in (h).

the other benefit of using IVD: ambiguous viewpoints are avoided, thus minimizing the possibility of convergence to an incorrect hypothesis.

The experiment was performed with several other objects from the database, and the results were quite similar to those presented above. These preliminary results are quite promising in that the system converged to the correct solution quickly using the IVD framework in all cases.

5 Conclusions

In this paper we have demonstrated how view-based entropy measures can be effectively used to steer an appearance-based recognition system towards viewpoints or poses that minimize ambiguity. The specific case studied was object recognition through the optical flow patterns induced by an operator sweeping an object by hand in front of a stationary

camera. A paradigm called the Interactive Visual Dialog was used to provide feedback to the operator in the form of images depicting how the object should be presented to improve the certainty of interpretation. Experimental results demonstrated the effectiveness of this feedback, both for keeping the on-line presentation within the training set and for quickly taking the operator through a sequence of poses by which the correct hypotheses could be identified with certainty. Although motion-based appearances were used in this study, the approach is applicable to any view-based method.

Future work will involve more extensive testing of the algorithm, particularly over a much broader range of permissible motions. This will be possible by using chromatic keying to remove the presence of moving hands from the images. Although it is highly unlikely that this approach will soon be found in your local supermarket, it does point out the possibilities for other applications of machine vision.

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] 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 August 1996.

[3] 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.

[4] 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.

[5] K. Bowyer and C. Dyer. Aspect graphs: An introduction and survey of recent results. In *Close Range Photogrammetry Meets Machine Vision*, volume 1395, pages 200–208. SPIE, 1990.

[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] D. W. Eggert, K. W. Bowyer, C. R. Dyer, H. I. Christensen, and D. B. Goldgof. The scale space aspect graph. In *Proceedings, Conference on Computer Vision and Pattern Recognition*, pages 335–340, Champaign, Il., June 15-18 1992. IEEE.

[9] D. J. Kriegman and J. Ponce. Computing exact aspect graphs of curved objects: Solids of revolution. In *PROC. of IEEE Workshop on the Interpretation of 3-D Scenes*, pages 116–122, Austin, Texas, November 27-29 1989. IEEE.

[10] N. Takeda 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.

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

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