Exploiting Feature Dynamics for Active Object Recognition

Philipp Robbel and Deb Roy
MIT Media Laboratory
Cambridge, MA 02139, USA
{robbel,dkroy}@media.mit.edu

***Abstract*—This paper describes a new approach to object recognition for active vision systems that integrates information across multiple observations of an object. The approach exploits the order relationship between successive frames to derive a classifier based on the characteristic motion of local features across visual sweeps. This motion model reveals structural information about the object that can be exploited for recognition. The main contribution of this paper is a recognition system that extends invariant local features (shape contexts) into the time domain by integration of a motion model. Evaluations on one standardized and one custom collected dataset from the humanoid robot in our laboratory demonstrate that the motion model allows higher-quality hypotheses about object categories quicker than a baseline system that treats object views as unordered streams of images.**

***Index Terms*—object recognition, active vision**

I. INTRODUCTION

In this paper we address the topic of category-level object recognition by active vision systems such as a robot equipped with a camera that can be moved to different viewing positions or a humanoid that can reach and modify the world in front of it. Confronted with a novel object, such a system must obtain an estimate over the object class as quickly as possible and be able to report the best available categorization at any point in time. The framework of maximizing recognition performance under time constraints is a general one and has a natural formulation in sequential Bayesian estimation.

We consider as a core requirement the ability to recognize object categories across viewpoint and texture variations and allow for objects without characteristic texture (for example, objects of uniform color). A large number of household items fall into this category and can not be distinguished by their texture alone. Shape-based recognition systems (e.g. [5]) are robust to such within-class texture variation but directly depend on the performance of the edge detector at every frame. Furthermore, object shape can be ambiguous across classes depending on the camera viewpoint, making classification based on a single random view inherently unreliable (see Figure 1).

In this paper, we present a system which addresses those challenges by accumulating evidence over a number of views obtained by following camera trajectories over the object. One of the core questions that we address here is what we can learn from a series of images (that stem from an "ordered sweep" over the object) without reverting to the full 3D structure of that object. In particular, we look at how low-level invariant features extracted from an object evolve over time as the active vision camera follows a set of known trajectories. We claim that the way an object transforms under specific camera motions (what we refer to as "object dynamics") allows us to disambiguate the object class faster than treating object views as independent observations. The fact that frames are connected in space and time leads to the observation that feature motion leaves a characteristic imprint of that object. To the best of our knowledge, combining this inter-frame motion information with local feature-based matching for object recognition has not been suggested before in the literature.

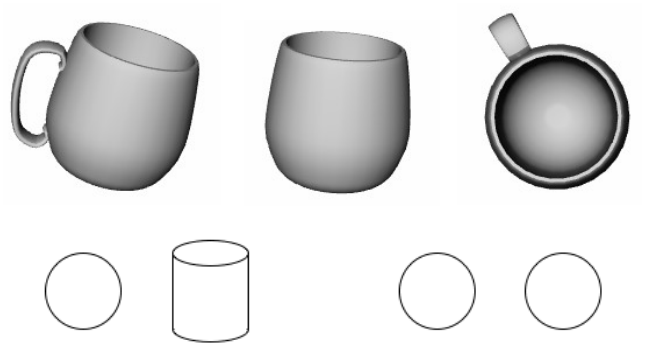

Fig. 1. Change in object appearance across viewpoints. *Above:* A cup loses its characteristic handle and approaches similarity to a cylinder (center) and a sphere (right side). *Below:* A cylinder needs at least two views to be distinguished from a sphere—both share the same top-down view.

The paper contains the following contributions. First, we develop a motion model in order to extend shape context features ([4]) into the time domain. Our model for feature motion is general and in principle compatible with other descriptors that allow to establish correspondence between successive frames such as SIFT [15]. Second, we develop a classifier that joins information from both local feature matching and motion matching on-line in a posterior class estimate. Finally, we demonstrate the performance of our system on two data sets showing that incorporating feature motion allows more certain hypotheses about the object category faster.

The paper is structured as follows. The next section discusses related work from the object recognition literature. After that, sections 2 and 3 describe our motion model and object classifier in detail. Section 4 presents experimental results before the concluding discussion in section 5.

A. Related Work

Much of the recent work in category-level object recognition is designed for single object views from constrained viewpoints. These approaches commonly combine invariant local features with methods from statistical pattern recognition to learn the distribution of appearance features for a particular view of an object class (e.g. [6], [8], [9], [13]).

Work to achieve recognition from multiple viewpoints is based on object geometry [10], global appearance [16], local appearance features [14], or both geometry and appearance [18]. These approaches are suited for object instance recognition but do not trivially extend to category-level recognition.

Our system uses a view-based approach to 3D object recognition based on local shape context features [5]. The shape context descriptor is robust to small variations in shape and is used here to capture the within-class variability of the different object categories. Storing and matching against multiple views of the same object is a commonly used method to realize 3D object recognition and is supported by psychophysical evidence in human object recognition [17].

In addition to classical view-based matching, we introduce a feature motion model and associated classifier. Structure from motion algorithms estimate a 3D model of the object when observed across multiple images [12]. Our interest here is instead how to combine some of the recent feature-based object recognition schemes with motion information. Unlike traditional optical flow algorithms [11], we compute a coarse approximation of the 3D motion field at a select number of points. The usefulness of motion features has recently been demonstrated for pedestrian detection in [20].

Perhaps closest in spirit to our idea of exploiting object motion for recognition is the work of Arbel and Ferrie in [2]. It is demonstrated that changes in velocity in the optical flow field as the camera is moved along the view sphere reveal structural information about the object that can be used for recognition. Our implementation, on the other hand, joins local feature-based matching with flow information (both direction and magnitude) obtained at the same feature points.

Lastly, there also exists a large amount of literature on classifying motion (motion recognition) or using motion for recognition of higher-level behaviors (motion-based recognition). Examples are as diverse as action recognition in video [7], sign language interpretation [19], or general gesture recognition. It has to be noted that the focus of this work is different, however. The observed change from frame to frame is due to ego motion of the camera and the characteristic structure of the object. Whereas we hope to detect and exploit structural change over time (such as the handle of a cup rotating into view) for object recognition, the work above focuses strictly on classifying motion, analogous to making statements about the movement of the camera rather than the object itself.

II. MOTION MODEL

A. Local Features

Shape context is a robust local feature descriptor introduced by Belongie et al. in [5]. It is utilized to find a similarity measure between a model and a (potentially deformed) target shape. Calculation of this measure requires to establish correspondence between both shapes and to assess the warping that the model points undergo to arrive at the target shape.

Fig. 2. Two frames (*left* and *right*) from two distinct trajectories over the same object. Displayed as well is the contour motion information that links both frames (*center*).

The shape context algorithm represents shapes by a finite set of N sample points from the object contour as obtained from a standard edge extraction procedure. For reliable matching, one associates with each point a shape context descriptor (a log-polar histogram centered at that point) that encodes the global layout of the shape relative to that point. At its base, the algorithm then computes the permutation π that assigns to every point p_i on the model shape the most similar point $q_{\pi(i)}$ on the target shape by minimizing the total assignment cost:

$$H(\pi) = \sum_i C(p_i, q_{\pi(i)}) \tag{1}$$

Here, C denotes a cost function based on the χ^2 test measuring the similarity between the respective points' histograms.

The shape context algorithm has been shown to perform well in scenes with low background clutter. Our system uses background subtraction to obtain an initial segmentation of the camera image.

B. Feature Motion

The crucial observation for the extension we present here is that images stem from an ordered sweep over an object rather than being sampled independently from the view sphere. Feature change across consecutive images depends on the ego motion of the camera and on the characteristic structure of the object (assuming that the object is stationary during the observation cycle). This is demonstrated in Figure 2 for two different trajectories over the identical object. For both examples, the middle picture shows the recorded change between the left and the right frames in the sequence.

Given an ordered set of images $I_1, I_2, \ldots, I_k$ from a trajectory, we compute feature motion for all pairs of successive

frames, i.e., $(I_1, I_2), (I_2, I_3), \ldots, (I_{k-1}, I_k)$. The approach that we pursue here to compute feature motion relies on the local feature descriptors to establish correspondence between both frames in each such pair.

With every image pair (I_i, I_{i+1}) we then associate two vectors $\mathbf{v}_{i+1}$ and $\mathbf{d}_{i+1}$ of size N (the fixed number of points sampled from both image contours). Entries in these vectors respectively denote the angles and magnitudes of the displacement vectors between corresponding points in I_i and I_{i+1}:

$$\mathbf{v}_{i+1} = \begin{pmatrix} \angle(p_1 - \pi(p_1)) \\ \angle(p_2 - \pi(p_2)) \\ \vdots \end{pmatrix}, \mathbf{d}_{i+1} = \begin{pmatrix} ||p_1 - \pi(p_1)|| \\ ||p_2 - \pi(p_2)|| \\ \vdots \end{pmatrix} \tag{2}$$

Here, $\pi(p_i)$ denotes the point corresponding to p_i in the second image. For scale invariance, we normalize the magnitudes in $\mathbf{d}_{i+1}$ with the median magnitude.

We devise a cost function to compare the feature motion obtained from two trajectories at the same point in time t. Given two such motion patterns $(\mathbf{v}_t^{(1)}, \mathbf{d}_t^{(1)})$ and $(\mathbf{v}_t^{(2)}, \mathbf{d}_t^{(2)})$, we incorporate cosine similarity and magnitude difference for each of the N entries into a joint cost term.

Cosine similarity is defined as the cosine of the angle between two corresponding entries, i.e. $\cos(v_{t,i}^{(2)} - v_{t,i}^{(1)})$ for all $i = 1, \ldots, N$ and is bounded in $[-1, 1]$. Naturally, it assumes its maximum for the case that the angle between both vanishes. To assess feature motion similarity we additionally compare the difference in displacement vector lengths $|d_{t,i}^{(2)} - d_{t,i}^{(1)}|$ which we normalize to fall into the same range $[-1, 1]$ (the maximum is assumed if both share the same length). If we denote these length differences by Δ_i, we can obtain a joint similarity score as the weighted sum:

$$s_i = \cos(v_{t,i}^{(2)} - v_{t,i}^{(1)}) + w_i \Delta_i \qquad \forall i = 1, \ldots, N \tag{3}$$

A total similarity between both motion patterns can then be computed as:

$$S = \sum_{i=1}^{N} s_i \tag{4}$$

The rationale is that one expects similar objects to result in similar contour motion, which is determined by both direction and magnitude of the individual displacement vectors. In general, we want to avoid that two displacement vectors of similar lengths but in different directions result in high similarity scores s_i and discount the Δ_i score based on the size of the angle between both displacement vectors.

III. OBJECT CLASSIFICATION

In this section we join the cost term obtained as output from the shape context algorithm together with the feature motion cost in a classifier that predicts class membership for the presented object.

We pursue standard Bayesian techniques to keep track of the current posterior distribution over object classes. For every new view, after shape matching and feature motion costs have been obtained, we execute the following three steps to update our class membership estimate:

Converting costs into probabilities. We use binary logistic regression to learn one-versus-all classifiers $P(\mathcal{C}_k|\mathbf{x}_{SC})$ and $P(\mathcal{C}_k|\mathbf{x}_M)$, denoting the probability of a particular object class $\mathcal{C}_k$ given shape matching cost and feature motion cost, respectively, for each of the k classes. During training, multiple trajectories on the viewsphere are performed over the object and object views, their shape contexts and motion information recorded. We then compute cost vectors $\mathbf{x}_{SC}$ and $\mathbf{x}_M$ for pairs in the dataset and divide the data into those of the same class and the rest. Similarly to [21] we assign a posterior of 1 to the distance vectors in the first group and 0 to the vectors in the other and use maximum likelihood to determine the parameters $\mathbf{w}$ of the logistic regression model:

$$P(\mathcal{C}_k|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})} \tag{5}$$

with σ being the logistic sigmoid function. Cost vector dimensionality depends on the sampling size along the object contours and is fixed to 50 in our experiments.

Joining shape matching and feature motion. In the previous step we derived two discriminative models, $P(\mathcal{C}_k|\mathbf{x}_{SC})$ and $P(\mathcal{C}_k|\mathbf{x}_M)$, that are now combined into a single distribution over $\mathcal{C}_k$. Our combination is based on a conditional independence assumption between $\mathbf{x}_{SC}$ and $\mathbf{x}_M$ given the class $\mathcal{C}_k$. This is the naive Bayes assumption:

$$P(\mathbf{x}_{SC}, \mathbf{x}_M|\mathcal{C}_k) = P(\mathbf{x}_{SC}|\mathcal{C}_k)P(\mathbf{x}_M|\mathcal{C}_k) \tag{6}$$

For the combination of models it then follows that:

$$\begin{aligned} P(\mathcal{C}_k|\mathbf{x}_{SC}, \mathbf{x}_M) &\propto P(\mathbf{x}_{SC}, \mathbf{x}_M|\mathcal{C}_k)P(\mathcal{C}_k) \\ &= P(\mathbf{x}_{SC}|\mathcal{C}_k)P(\mathbf{x}_M|\mathcal{C}_k)P(\mathcal{C}_k) \\ &\propto \frac{P(\mathcal{C}_k|\mathbf{x}_{SC})P(\mathcal{C}_k|\mathbf{x}_M)}{P(\mathcal{C}_k)} \end{aligned} \tag{7}$$

This yields a posterior distribution over the class based on results from both feature motion comparison and shape matching.

Online updates of the class distribution. To continuously adapt the estimate of the class distribution as more views of the object are discovered, we evaluate the naive Bayes classifier above at every time step. Let $P_i(\mathcal{C}_k|\mathbf{x}_{SC})$ and $P_i(\mathcal{C}_k|\mathbf{x}_M)$ denote the predictions of the shape context and feature motion models for the ith object view, respectively. Then, at time t, we have that

$$P_t(\mathcal{C}_k|\mathbf{x}_{SC}, \mathbf{x}_M) \propto \frac{\prod_{i=1\ldots t} P_i(\mathcal{C}_k|\mathbf{x}_{SC})P_i(\mathcal{C}_k|\mathbf{x}_M)}{P(C_k)} \tag{8}$$

which allows us to aggregate the numerator in an efficient, recursive manner. Throughout, we assume a uniform prior over the object classes, $P(\mathcal{C}_k)$.

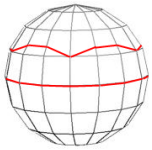

Fig. 3. Trajectories adopted for the ETH-80 data set.

Fig. 4. The three selected instances per category from ETH-80.

IV. EXPERIMENTAL EVALUATION

A. Data Sets

ETH-80. The original ETH-80 dataset consists of eight object categories with ten instances per category. Instances vary in shape and texture but otherwise appear on a uniform background and roughly share the same size. Each instance comes with 41 images that are distributed evenly over the upper view sphere at a resolution of 256×256 pixels.

The use-case suggested for this data set in [3] is leave-one-out cross-validation. Since we operate on object sweeps instead of single images, we initially have to introduce an order over the included images to simulate camera trajectories. Unfortunately, the images taken across the view sphere do not stem from smooth, curvilinear camera paths but instead from equally spaced points on an octahedron approximation to the sphere. In Figure 3 we show the approximate trajectories we adopted for the dataset, resulting in four trajectories overall (two as shown and two for the backside of the sphere).

The resulting image assortment for all four trajectories is visualized for a cup example in Figure 5. As shown, each trajectory results in a unique sweep of either eight or six images.

Our entire training set consists of the first three instances per object class from the ETH-80 data, resulting in a total number of 672 images. Prototypical views for each instance are shown in Figure 4. Note, that being a shape-based method, we discard color information from our data set.

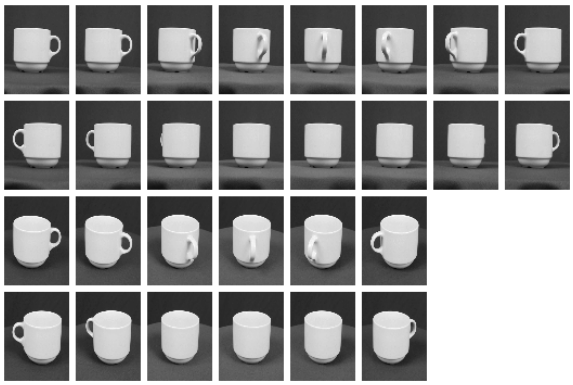

Fig. 5. The four sweeps associated with the third cup instance.

Trisk-22. The Trisk-22 data set consists of a set of 22 real-world objects collected from an active vision robot named Trisk in our laboratory. Trisk is loosely modeled after human physiology and consists of a 6-degree-of-freedom (DOF) arm, a 4-DOF actuated head with stereo vision cameras, and a three-fingered hand with 6-DOF force-torque sensing abilities at the fingertips. As shown in Figure 6, the dataset is divided into eleven categories with a variable number of instances per category. Images are generally low-resolution, of different sizes and taken under varying lighting conditions. All views are obtained during the traversal of three pre-determined camera trajectories and are sampled at regular 1 second intervals. There is a total of 1010 object views contained in the Trisk-22 database. It is available from the authors upon request.

In Figure 7 we visualize a number of views from that data set with their extracted motion information overlaid. Due to imaging artifacts, such as from specular highlighting, the contour samples and their motion vectors are more noisy than in the ETH-80 data set.

B. Results and Discussion

In this section, we compare the performance of the multi-frame shape context classifier $P(\mathcal{C}_k|\mathbf{x}_{SC})$ (referred to as SC or the baseline), the feature motion-based classifier $P(\mathcal{C}_k|\mathbf{x}_M)$ (M), and the combined classifier $P(\mathcal{C}_k|\mathbf{x}_{SC}, \mathbf{x}_M)$ (SC+M) on both data sets. All classifiers accept a sweep of incoming pictures until the entropy of the posterior reaches a threshold and then return the MAP class estimate. The general approach we take here is leave-one-sweep-out cross-validation, i.e., for each possible hold-out sweep we retain the remainder of the data for training purposes.

ETH-80. Figure 8 shows the leave-one-out error rate for the single frame-based shape context classifier on the ETH-80 subset. This classifier computes the shape context distances to each of the stored models and outputs the MAP class estimate for every frame (essentially a nearest neighbor classifier with the shape context distance metric). Contrary to the proposed sweep-based classifiers SC, M, and SC+M, this recognition system operates purely on a frame-by-frame basis and does not take the object history into account.

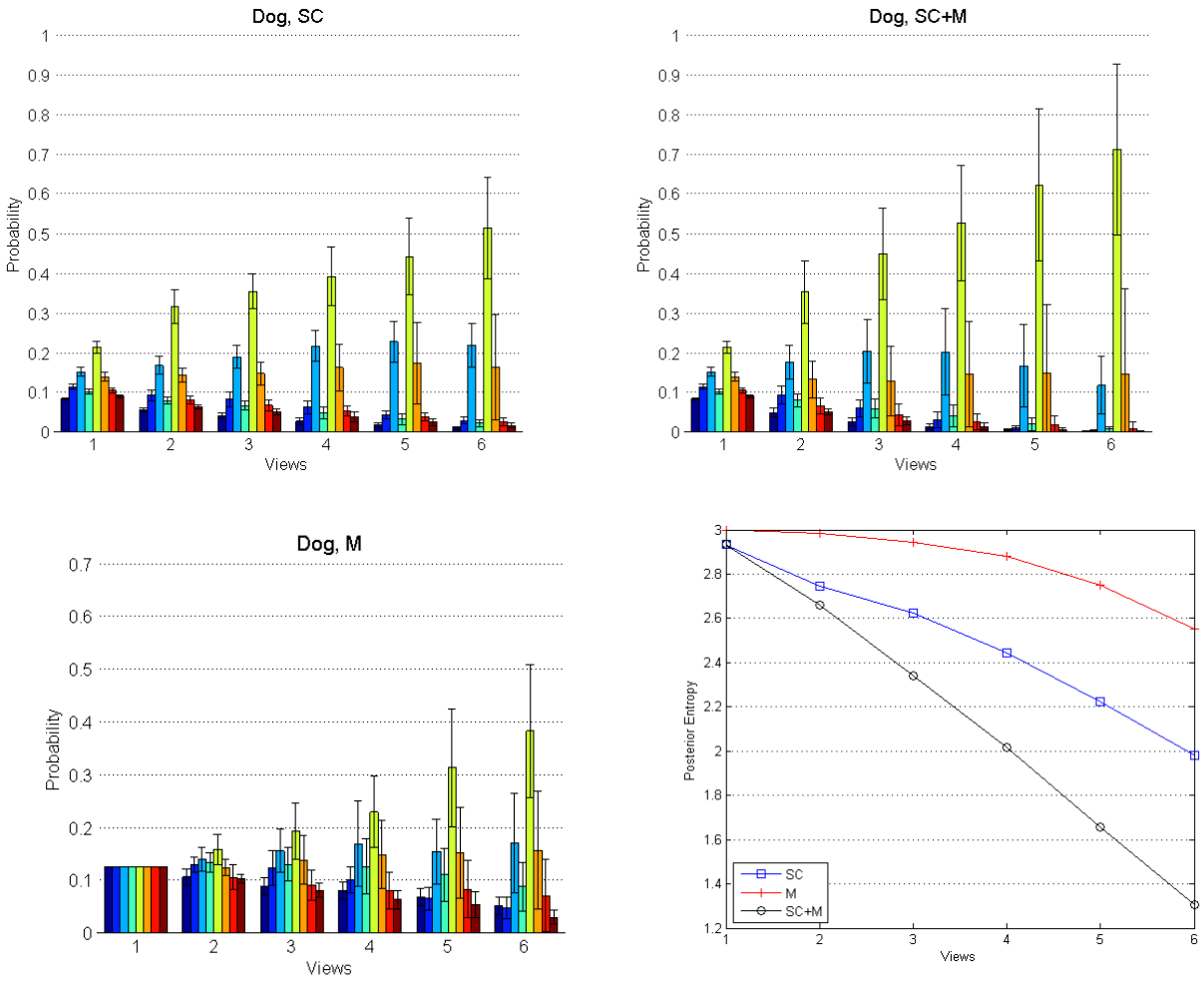

Fig. 9. Mean posterior distributions after a specified number of dog views for SC, M, and SC+M classifiers. The entropy of the distributions is shown on the bottom right.

All following experiments summarize the recognition performance of the different sweep-based classifiers. For the first experiment, we compute the average posterior distributions $P(\mathcal{C}_k|\mathbf{x}_{SC})$, $P(\mathcal{C}_k|\mathbf{x}_M)$, and $P(\mathcal{C}_k|\mathbf{x}_{SC}, \mathbf{x}_M)$ obtained over a range of 1-6 impressions of the same object. The results are averaged over all twelve sweeps per object category (3 instances times 4 sweeps per instance). For sake of space, we report the results for the dog category only; we obtain comparable results for the other categories. Figure 9 shows the mean posterior distributions together with the standard deviation error bars for the SC, M, and SC+M classifiers. We also show the reduction in entropy for each of the average posterior distributions throughout the sweep.

For both SC and SC+M classifiers, the error rate reaches zero for all categories after three consecutive object views.

We can draw the following conclusions from this experiment: 1) Using motion as an additional source of information is generally valid. For all categories the motion-based classifier produces results that boost the posterior probability of the correct category. 2) Motion information does not replace traditional local features. As seen in Figure 9, the entropy of the posterior $P(\mathcal{C}_k|\mathbf{x}_M)$ is generally higher throughout the sweep. In the same Figure one can also observe how $P(\mathcal{C}_k|\mathbf{x}_M)$ evolves from an uninformative (first view where no motion data has been obtained yet) toward more informative distributions as more object views are obtained. 3) Joining both classifiers (SC+M) generally leads to a desirable posterior distribution (as judged by both the correctness of the MAP estimate as well as the posterior entropy) quickest. We experience significant gains in posterior probability over the baseline system for all classes in the database. For the dog class in Figure 9, we achieve a mean posterior probability $P(\mathcal{C}_{dog}|\mathbf{x}_{SC}, \mathbf{x}_M)$ of 0.72 versus 0.52 for the shape only classifier after six views.

Trisk-22. For this data set, we again look at the average posterior class distributions resulting from the classification of the hold-out sweeps. The presentation is analogous to that of the ETH-80 results, except for the fact that we now evaluate the posterior distributions after up to 10 views (the smallest sweep in the data set).

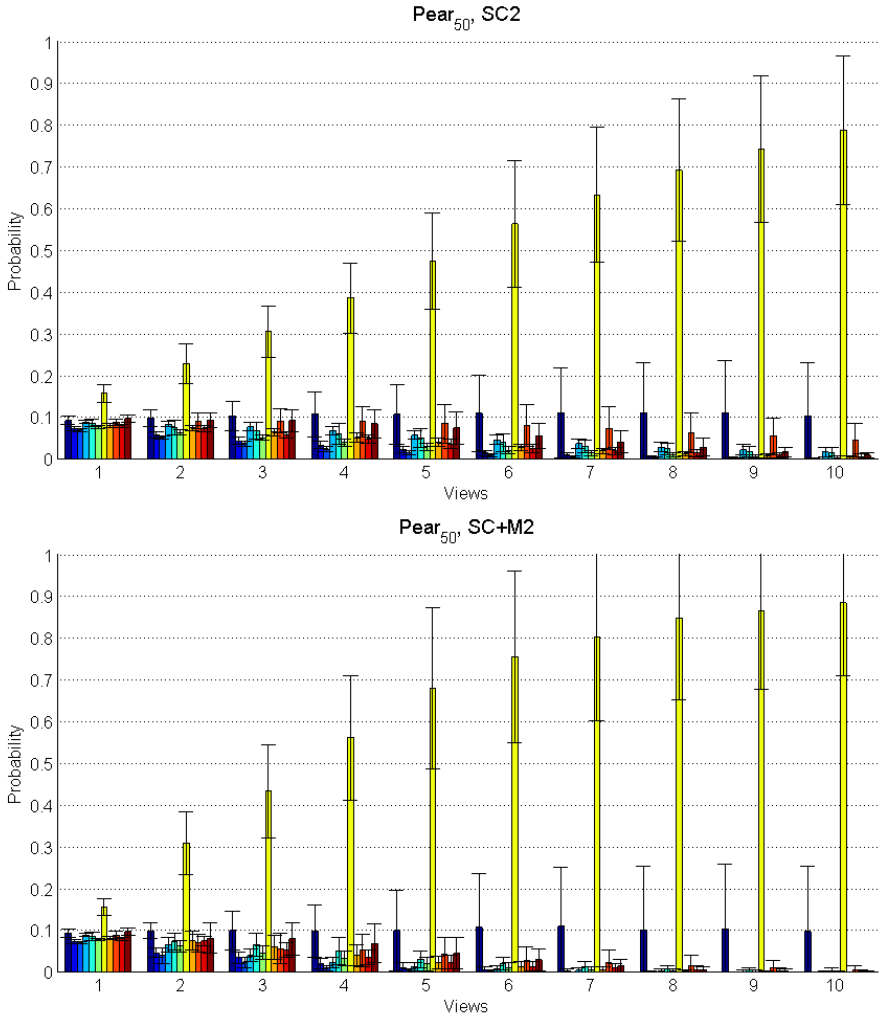

Fig. 10. Mean posterior distributions after a specified number of pear views for SC and SC+M classifiers.

The main results of this experiment are summarized in Figure 10 for the pear class in the Trisk-22 dataset. The lower of the two sets of distributions shows the improvement after including the explicit feature motion model as an additional source of information during classification. The results for the remaining classes in the data set show similar benefits from including the motion model in the classifier.

The following observations can be made about the results on the Trisk-22 database: 1) Classification arising from the MAP estimate of the motion model $P(\mathcal{C}_k|\mathbf{x}_M)$ is generally more discriminative than for the ETH-80 dataset. One can assume that the reason for this lies with the denser sampling on the object contours due to the generally smaller image sizes. In addition to this spatially denser sampling, we also have a temporally much denser sampling of object views in every sweep which may add to the robustness of the motion model. 2) As observed before, the joint SC+M model outperforms the SC baseline model significantly in terms of producing low-entropy posterior class distributions at an earlier point in time. This is more apparent with the Trisk-22 dataset than with the ETH-80 data presented previously.

V. CONCLUSION

We established on two different data sets that feature motion (or "object dynamics") is a valid principle that can be exploited for object recognition of active vision systems. In contrast to our baseline model that essentially treats every image sequence as unordered, our joint model $P(\mathcal{C}_k|\mathbf{x}_{SC}, \mathbf{x}_M)$ makes use of the additional structural information revealed by the feature motion between frames about the object under the assumption that the object is static and not undergoing self-motion.

In the course of this paper we developed three probabilistic classifiers based on the shape context algorithm (SC), feature motion information (M), as well as both shape and feature motion (SC+M). We then used a distinct (yet limited in our initial experiments) set of robot head trajectories to test our sweep-based recognition methods on two data sets. The demonstrated system is invariant to scale and in-plane object rotation. For the motion model, the latter is achieved by training the system with different object orientations.

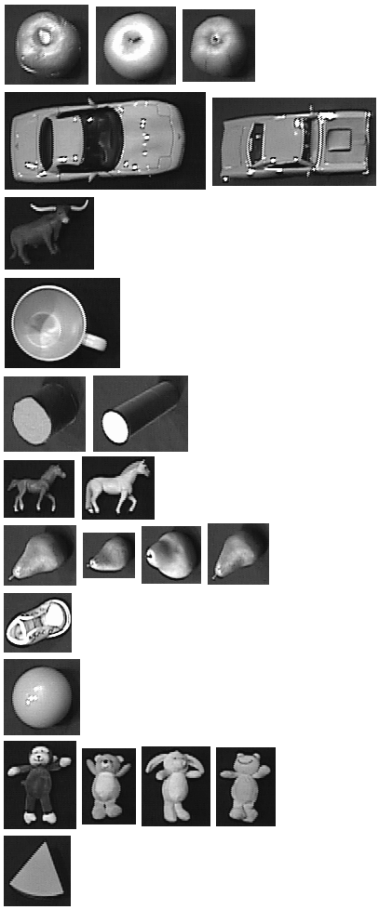

Fig. 6. The 11 object categories and object instances in the Trisk-22 database.

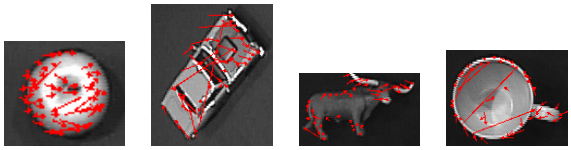

Fig. 7. Feature motion on a set of views contained in the database.

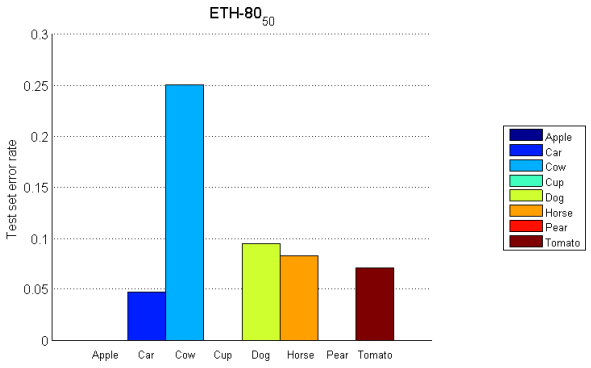

Fig. 8. The ETH-80 test set error rates for one-shot shape context matching.

We could demonstrate for both data sets that incorporating feature motion achieves a higher-quality hypothesis about the category in faster time. In particular for the real-world Trisk-22 database the joint model considerably improved on the individual models SC and M. In conclusion, feature motion-based classification appears to be a valid addition to an active recognition system built around invariant local features.

For future work, we will explore how recognition can be accelerated further. In the current implementation, training and test camera trajectories are identical. However, a simple modification would allow the camera to skip ahead on the trajectory to highly disambiguating views as determined by the entropy of the posterior class distribution for all views in the training set (e.g. [1]). Second, it is feasible to explore whether clustering in the feature evolution space can reveal part structures. Even in its current form, however, we believe to have contributed a reliable object recognition system that may particularly be useful for robotic or otherwise time constrained active vision systems.

References

[1] T. Arbel and F. P. Ferrie. Entropy-based gaze planning. *IVC*, 19(11):779–786, September 2001.

[2] T. Arbel, F. P. Ferrie, and M. Mitran. Recognizing objects from curvilinear motion. In *BMVC*, 2000.

[3] B. Leibe B. and Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*, pages II: 409–415, 2003.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, pages 831–837, 2000.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[6] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[7] J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. In *CVPR*, pages 928–934, 1997.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, 2005.

[9] R. Fergus. *Visual Object Category Recognition*. PhD thesis, University of Oxford, 2005.

[10] W. E. L. Grimson and T. Lozano Perez. Localizing overlapping parts by searching the interpretation tree. *PAMI*, 9(4):469–482, July 1987.

[11] B. K. P. Horn and B. G. Schunck. Determining optical flow. *AI*, 17(1-3):185–203, August 1981.

[12] T. Jebara, A. Azarbayejani, and A. P. Pentland. 3d structure from 2d motion. *SPMag*, 16(3):66–84, May 1999.

[13] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *British Machine Vision Conference (BMVC'03)*, pages 759–768, Norwich, UK, Sept. 2003.

[14] D. G. Lowe. Local feature view clustering for 3d object recognition. *CVPR*, 1:682–688, 2001.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[16] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. J. Comput. Vision*, 14(1):5–24, 1995.

[17] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3:1199 – 1204, 2000.

[18] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Comput. Vision*, 66(3):231–259, 2006.

[19] T. E. Starner, J. Weaver, and A. P. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *PAMI*, 20(12):1371–1375, December 1998.

[20] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. *CVPR*, pages 794–801, 2009.

[21] H. Zhang and J. Malik. Learning a discriminative classifier using shape context distances. *CVPR*, 1:242, 2003.