

Exploiting Human Actions and Object Context for Recognition Tasks

Darnell J. Moore[†], Irfan A. Essa^{†‡}, and Monson H. Hayes III[†]

[†]School of Electrical and Computer Engineering, [‡]College of Computing
Georgia Institute of Technology, Atlanta, Georgia 30332. USA.*
djmoore@ece.gatech.edu, irfan@cc.gatech.edu, mhh3@ece.gatech.edu

Abstract

Our goal is to exploit human motion and object context to perform action recognition and object classification. Towards this end, we introduce a framework for recognizing actions and objects by measuring image-, object- and action-based information from video. Hidden Markov models are combined with object context to classify hand actions, which are aggregated by a Bayesian classifier to summarize activities. We also use Bayesian methods to differentiate the class of unknown objects by evaluating detected actions along with low-level, extracted object features. Our approach is appropriate for locating and classifying objects under a variety of conditions including full occlusion. We show experiments where both familiar and previously unseen objects are recognized using action and context information.

1. Introduction

This paper proposes a novel approach to human activity recognition that uses context information of particular objects in the scene. We define classes that contain object-specific information, including associated properties, appearance-based descriptions and actions. Objects provide a means to focus attention on an individual interaction while maintaining awareness of other interactions in the scene. By tracking hands, contact with known objects is detected. Once contact has been established, context is used to suggest specific hidden Markov models (HMMs), if any, that may provide more explicit descriptions of action associated with the object. Interactions captured over time are aggregated using Bayesian statistics, producing summaries of activity.

Additionally, we show that the relationship between human actions and objects can be exploited to detect and classify objects. Object classification is inferred, in part, by detecting learned actions. Prior knowledge about object categories and image analysis provides additional discrimination.

A naive Bayesian classifier evaluates detected actions along with recovered region features to differentiate the class of the unknown object. We demonstrate detection and classification of both rigid and deformable objects under a variety of conditions including full occlusion.

This work has many practical applications where passive, non-intrusive processes are needed to capture interactive experiences, such as smart spaces or video surveillance. Our proposed framework offers a pragmatic design approach for a real-time vision system intended for use in multiple domains.

Related Work: The hidden Markov model (HMM) is widely used to model complex motions for action recognition tasks [2, 9, 12]. Our work combines HMMs with context to formalize the relationship between actions and objects.

Architectures that leverage hierarchical inheritance and context have demonstrated their utility for managing and cataloging information [4, 8] as well as for solving image understanding problems. Pinhanez and Bobick have provided a *Past, Now, and Future* (PNF)-network, based on interval algebra, to describe the temporal structure of actions, sub-actions, and events [11]. Mann and Jepson integrate information about object properties and abilities (primarily force/dynamic) to develop representations of activity [7]. They attempt a bottom-up approach that infers physical descriptions of the actions depicted in image sequences.

There has been extensive research on object detection and classification (see [13] for review). Most of these proposals require models or templates as the basis for constraining and mapping extracted features. However, only limited progress has been made towards the use of context and inference as the *primary* means of recognizing objects.

Bayesian methods provide a formal means to reason about partial beliefs under conditions of uncertainty [10]. The framework proposed by Buxton *et al.* uses Bayesian Networks to perform surveillance and evaluate evidence in well understood scenes [3]. Yi and Chelberg assert the appropriateness of these networks for selecting probable objects based on discriminating features and domain-specific knowledge [14]. We incorporate similar concepts in our framework

*This research was funded, in part, by Texas Instruments and the National Science Foundation, Grant #EIA-9806822.



Figure 1. Structure for a book article.

for summarizing activity and labeling unknown objects.

2. Representational Framework

We propose a hierarchical framework for representing prior knowledge about image contents. This architecture, called *ObjectSpaces*, uses familiar object-oriented constructs like *classes* and inheritance to manage object context. *ObjectSpaces* uses an adaptive bottom-up and top-down architecture with three, integrated layers. The *Extraction layer* is responsible for finding, extracting, and tracking people and articles in the scene using simple, low-level techniques. It also provides facilities for characterizing motion using HMMs. The *Object layer* contains objects that represent people and articles in the scene. The *Scene layer*, which contains domain-specific context, monitors physical contact between people and articles, then examines these interactions for patterns of behavior.

A class is a container for properties and methods needed for holding context and implementing tasks. Objects pointing to scene articles and people are instantiated from two parent classes, *Article* and *Person*, respectively. The object representing domain-specific context is derived from a third parent class called *Scene*.

Instantiated objects provide a means to focus on motion near a specific object without directing attention away from any other object in the scene. By tracking the hands, contact between known articles and people can be detected. Once contact has been established, articles attempt to gather more explicit descriptions of activity by comparing observed hand motion to pre-trained actions that are indigenous to the object. For example, consider a *book* object with two associated actions described by HMMs, i.e., *left-to-right* motion $\rightarrow \lambda_{ff}$: “flip forward” and *right-to-left* motion $\rightarrow \lambda_{fb}$: “flip backward”. After the hands penetrate the image area where the book is located (establishing contact), sweeping the hand from left to right will indicate flipping a page forward. Several of the properties in this class are shown in Figure 1. Articles also maintain histories of such interactive events with people, which can be helpful for summarizing activities.

Class Inheritance and Reuse: Derived classes, or “children”, possess the same properties and methods as the parent class, but can be extended by adding additional properties and

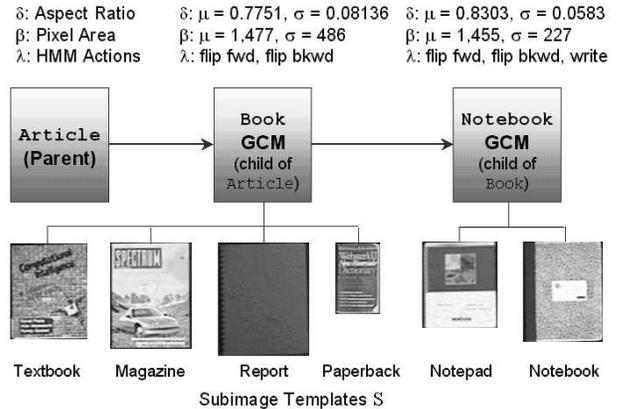


Figure 2. GCM for book class derived from instantiations.

features. Inheritance provides a natural hierarchy that we take advantage of to organize or disambiguate classes during object and action recognition.

ObjectSpaces manages context in layers of abstraction. In this way, objects (articles and people) and their behaviors can be defined intrinsically, without regard for the domain in which they will appear. Once developed, the class database is available for reuse in multiple domains without retrofitting the entire framework for a specific application. Behaviors between objects can be better specified at the scene-level, placing task dependencies at the highest level of abstraction. For more information of *ObjectSpaces*, see [8].

Generalized Class Models: To organize our class database, we develop *generalized class models* (GCMs) for articles. This model contains region- and image-based descriptions that are representative of all instantiations of that particular class. A GCM is created for every child level within the class hierarchy. For example, the *book* GCM points to stored, initialized *book* objects, as Figure 2 illustrates. The associated class actions, in this case λ_{ff} and λ_{fb} , and Gaussian probability distribution functions \mathbf{P} that describe region features, i.e. $P_{pixelarea}$, are also referenced. The *notebook* class extends *book* by adding an action model for “write,” λ_{wrt} . Because the *book* GCM is a parent, \mathbf{P}_{book} is influenced by the contribution of its children, which includes *notebook*, although $\mathbf{P}_{notebook}$ bears no dependence on \mathbf{P}_{book} . The entire set of GCMs forms model space \mathbf{M} .

3. Human Motion Analysis

Features of people and objects must be automatically identified and extracted to build meaningful representations that accurately capture activities. We employ simple approaches for tracking in real-time although a variety of more sophisticated techniques could have been used. Our main focus is

on building higher-level meanings so we have designed our framework to be invariant to these techniques.

Tracking the Hands: To determine which items in the surroundings are handled, the location of a person's hands is recovered. Color input frames are analyzed in the YUV color space. To detect new people entering the scene, we use background segmentation (Y-channel) to identify regions that belong to people. Then we segment skin-colored blobs from the image (YUV-channels) using the color table $\mathbf{C} = [\mathbf{y} \ \mathbf{u} \ \mathbf{v}]$, which is manually initialized. Blobs that do not match the view-based model supplied by the person object are eliminated. A tracking algorithm, which selects the centroid of each hand blob $\mathbf{x} = [x \ y]^T$ from the remaining candidates, leverages scene context to deal with occasional occlusions and tracking failures. To assist in tracking future locations of each hand, the linear, estimated hand position is given by $\hat{\mathbf{x}}_{t+1} = \mathbf{x}_t + d\mathbf{x}$, where $d\mathbf{x} = \mathbf{x}_t - \mathbf{x}_{t-1}$.

Action Characterization using HMMs: The hidden Markov Model (HMM) allows us to use established stochastic processes to characterize deliberate, repeatable hand motion (see [6]). To model actions of duration T that take place throughout the scene, hand position alone is used to construct the observation feature $\mathbf{O} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$. Our approach assumes a fixed, overhead camera so scale variation is insignificant because perspective projection distortion is small. We anticipate that actions can occur any place within the scene. To normalize motion displacement, objects supply affine transformations based on the location of their bounding boxes to deal with translation and rotation. The normalized observation feature becomes $\hat{\mathbf{O}} = \mathbf{R}(\theta)\mathbf{O} + \mathbf{T}$, where \mathbf{R} is a 2x2 rotation matrix about θ and \mathbf{T} is a displacement vector from the object's centroid to the center of the image. As the hands transverse through space during some action, they pass through certain normalized areas in the image space that correspond to the HMM's states. Hand transitions from area to area generate a sequence of states, which characterize an action. We assumed all actions are single-handed motions. During training, roughly 20 examples for each action captured by the same person are manually segmented. Deliberate rest states are used as delimiters to parse individual actions during testing. A 6 state, continuous HMM was empirically selected to optimize recognition.

4. Evidence for Recognition

We collect image-, object-, and action-based evidence to label and summarize activity as well as to identify unknown objects and people.

Object-based Evidence: The statistical history of interactions between articles can be helpful for predicting future

events. To begin, we establish the known articles in the environment by deriving classes from `Article` for each type and identifying its location in the scene by a bounding box. After the scene is initialized, articles are denoted as $\mathbf{g} = \{g_1, g_2, \dots, g_q\}$. When human activities involving these objects are observed, the scene layer computes conditional probabilities between every two articles, storing them in a $q \times q$ matrix \mathbf{A} , such that $\mathbf{A} = \{o_{ij}\}$, where $o_{ij} = P(g_j | g_i)$. Here o_{ij} represents the probability that the j th article is handled given that the i th article was previously handled. This information is particularly valuable for inferring an unknown object by exploiting its relationship with known articles.

To benefit tracking and the detection of objects that are fully occluded or a part of the background, we exploit the spatial proximity of known articles. Within every scene, *activity zones* are designated by the bounding boxes of known articles to indicate image regions where there is frequent hand traffic. As the hands are tracked during activity, additional zones are created when hands spend time in undesigned areas. We use object transition matrix \mathbf{A} and activity zones as a measure of object-based evidence.

Image-based Evidence: After background initialization ($I(0)$), changes in the background, presumably new, unknown articles or people, are segmented from subsequent frames $I(t)$ based on an empirically determined threshold α , producing binary image $\hat{B}(t)$. Connected component analysis produces n regions labeled Z_i , $1 \leq i \leq n$. Our correspondence algorithm analyzes and tracks each Z_i over consecutive frames, using motion and pixel area as metrics for deciding which regions belong to a person or to an object. A parametric model for people is invoked to label regions that belong to humans and hand tracking starts as previously mentioned.

Regions that appear to be unlabeled articles are examined to identify evidence that can be used for recognition. The foreground subimage $F_i(t)$ of Z_i is given by $F_i(\mathbf{x}, t) = I(\mathbf{x}, t)$, $\forall \mathbf{x} \in Z_i$, where $\mathbf{x} = [x \ y]^T$. To determine Z_i 's orientation with respect to the angle of the principle axis, we calculate ϕ given by

$$\phi = \frac{1}{2} \arctan \left(\frac{2 \sum_{(x,y) \in Z_i} xy}{\sum_{Z_i} x^2 - \sum_{Z_i} y^2} \right). \quad (1)$$

We apply a rotation matrix to obtain the normalized foreground subimage, i.e., $\hat{F}_i = \mathbf{R}(-\phi)F_i$, so that Z_i can be compared to stored article templates from our class database, which have already been modified such that $\phi = 0$ (illustrated in Figure 3). We also recover other features of Z_i , including pixel area β_i , aspect ratio δ_i , and perimeter edge count ϵ_i . A bounding box is constructed around Z_i , which helps to detect future hand overlap and to recover actions performed around this region. To minimize noisy measurements, these parame-

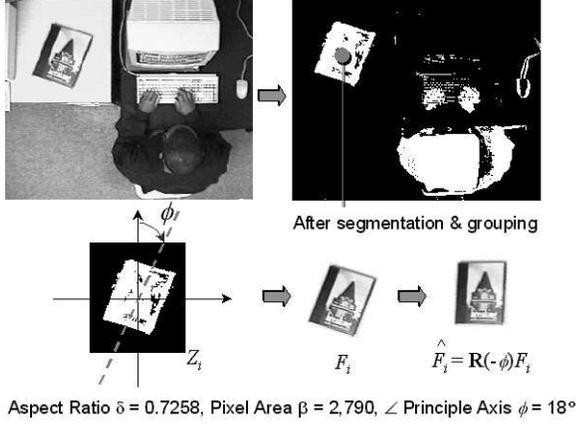


Figure 3. *Image-based Evidence:* Segmentation and analysis of newly introduced objects.

ters are averaged over several frames and used to instantiate Z_i as an *unknown Article* object.

To leverage prior knowledge, we first attempt to identify each unknown Z_i by comparing its rotationally invariant image, \hat{F}_i , to each stored subimage template S_j that shares a similarly-sized bounding box, such as those shown in Figure 2. The *mean square error* (MSE) η , used to quantify matching of an $N_1 \times N_2$ pixel template, is defined as

$$\eta(i, j) = \frac{1}{N_1 N_2} \sum_{\mathbf{x} \in S_j} [S_j(\mathbf{x}) - \hat{F}_i(\mathbf{x})]^2. \quad (2)$$

The parameters $\beta_i, \delta_i, \epsilon_i$, and η form Υ_i , which represents imaged-based evidence.

Action-based Evidence: The simple region and template metrics discussed above may not provide the sufficient evidence necessary to classify unknown objects. Moreover, the unknown object may be part of the background $I(0)$, which leads us to exploit motion to infer the object's class. Let the entire set of HMMs used by the entire set of GCMs \mathbf{M} form action model space Γ of dimension v , such that $\Gamma = \{\lambda_{ff}, \lambda_{fb}, \lambda_{wrt}, \dots\} = \{\lambda_1, \lambda_2, \dots, \lambda_v\}$. We determine the most likely sequence of states given model λ_γ using the Viterbi Algorithm [6]. So for every action model in Γ , we calculate $p_\gamma = P(\mathbf{O}|\lambda_\gamma)$, which we refer to as a measure of action-based evidence.

Evaluating Evidence: Bayes' theorem weighs the strength of belief in a hypothesis against prior knowledge and observed evidence. In addition, it provides attractive features, including: (1) its ability to pool evidence from various sources while making a hypothesis, and (2) its amenability to recursive or incremental calculations, especially when evidence is accumulated over time [1]. These features motivate

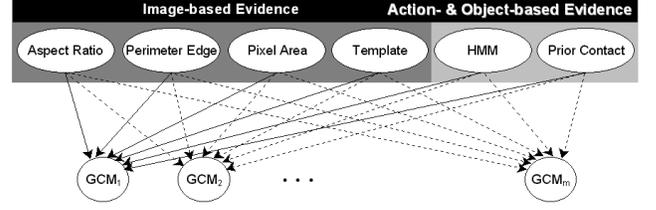


Figure 4. Belief network corresponding to a naive Bayesian classifier for selecting the most likely GCM.

our application of Bayesian classification to summarize activities and resolve unknown objects.

To summarize the object activities, human-object interactions are modeled using the Markovian assumption, i.e. *current activity influenced by previous activity*, which is a weak assumption but one that maintains computational efficiency. If hand motion is too subtle to be reliably characterized, an “action” event ϵ is used.

To begin, first consider a set Ω of k different activities Λ , such that $\Omega = \{\Lambda_1, \Lambda_2, \dots, \Lambda_k\}$. Each activity Λ contains a set of action models or events, i.e. $\Lambda_{writing} = \{\lambda_{drawing}, \lambda_{erasing}, \epsilon_{move\ pen}\}$. In order to compute the likelihood of an activity, we solve the relation

$$\hat{\Lambda} = \max_{\Omega} \{P(\mathbf{O}|\Lambda)\}. \quad (3)$$

The probability that a sequence of observations is produced by a given activity Λ_α is expressed as

$$\begin{aligned} P(\mathbf{O}|\Lambda_\alpha) &= \sum_{\text{all } \mathbf{q}} P(\mathbf{O}|\mathbf{q}, \Lambda_\alpha) P(\mathbf{q}|\Lambda_\alpha) \\ &= \sum_{\text{all } \mathbf{q}} P(\mathbf{O}|\mathbf{q}) P(\mathbf{q}|\Lambda_\alpha) \end{aligned} \quad (4)$$

where $\mathbf{q} = \lambda_1, \lambda_2, \dots, \lambda_n$ represents an n -dimensional sequence of actions. An initial likelihood of a sequence of actions occurring during an activity $P(\mathbf{q}|\Lambda_\alpha)$ is computed after training. Moreover, this likelihood can be updated during actual testing.

For recognizing unknown objects, we construct a naive Bayesian classifier [5], where each contributing portion of evidence is assumed to be independent, to determine the most likely GCM associated with Z_i (Figure 4). We allow M_k to represent the k th GCM in \mathbf{M} . Starting with our image-based evidence, we develop a measure of shape similarity that is expressed as $P(M_k|\Upsilon_i)$, or alternatively by

$$\chi_{i,k} = \frac{P(\beta_i|M_k)P(\delta_i|M_k)P(\epsilon_i|M_k)P(M_k)}{P(\beta_i, \delta_i, \epsilon_i)}. \quad (5)$$

From Equation 2, if $\min_j \{\eta(i, j)\} \geq \tau$, where τ is discovered empirically, we consider the contribution from any of the stored subimage templates S_j . To express the MSE as a

likelihood that Z_i is derived from S_j 's GCM, we approximate $P(Z_i|S_j, M_k)$ by

$$P(Z_i|S_j, M_k) \approx p_\eta = e^{-\frac{\eta}{\tau}}, \quad (6)$$

to produce an exponential distribution p_η on the range [0:1]. The contribution from observing action λ_γ is quantified by solving the relation

$$\hat{p} = \max_{\Gamma} \{P(M_k|p_\gamma)P(\lambda_\gamma|M_k)\}. \quad (7)$$

For each Z_i , we select the \hat{k} th GCM producing the highest score by

$$\hat{k} = \arg \max_{M, \Upsilon, \Gamma} \{\mathbf{w}^T \mathbf{e}\} \quad (8)$$

where \mathbf{w} represents weights and $\mathbf{e} = [\chi_{i,k} \hat{p} o_{i,j} p_\eta]^T$. As evidence is acquired, \mathbf{w} is adjusted to reflect the strongest beliefs, such that

$$w_i = \frac{e_i^2}{\mathbf{e}^T \mathbf{e}}, \text{ subject to } \sum_i w_i = 1. \quad (9)$$

Additional biases are also placed on \mathbf{w} , which have the effect of dynamically adjusting the contribution of image- or action-based influences. For example, if major changes in shape are detected over time, a deformable object is assumed and both w_1 and $w_4 \rightarrow 0$. If image-based evidence Υ_i is available, \hat{k} is used to set the observation length T of the HMM based on the most probable GCM. Otherwise, it is based on the average T used to train all models in Γ . Note that motion evaluation can not begin until the specified number of observations has been acquired.

After initial image-based evidence is acquired, future assessments of Z_i are not taken unless some noticeable change has occurred. Recall that deliberate rest states between actions are used as delimiters for parsing hand motion. While every frame is evaluated, action events are only recorded once to prevent redundant observations. No belief assessment is taken if no portion of the evidence offers at least some moderate belief, i.e. if $e_i \geq \frac{2}{3} \forall i$. Because belief is accumulated, weak evidence is not considered so that scores are not compromised over time.

5. Results and Experiments

Before we conducted experiments to test our approaches, we used ObjectSpaces to set up article classes for the office, kitchen, and automobile domains. The system has been implemented in C++ to run on the Win9x/NT platform in near real-time. Our approach uses a static view of the scene from a ceiling mounted camera position (auto interior scene captured by angled, sunroof-mounted camera).

Article	Action Recognition Accuracy
Office Environment	
bookcase	grab book - 100%, return book - 90%
book [†]	flip forward - 94%, flip backward - 90%
chair	no action, event only - 100%
cup	drinking - 100%, stir - 90%
desk	drawing - 93%, erasing - 86%, open drawer - 100%, close drawer - 80%
keyboard	no action, event only - 97%
mouse	no action, event only - 95%
notebook [†]	flip forward - 94%, flip backward - 92%, write - 80%
phone	pick up - 90%, put down - 60%, dial num. - 80%
printer [†]	open cover - 95%, close cover - 92%
table	feeding - 88%, stirring - 93%, cut - 85%
Kitchen Environment	
bowl	stir - 90%
cabinet	no action, event only - 100%
cut board	cut - 88%, scrape off - 93%
can opener	no action, event only - 100%
disposal	no action, event only - 100%
microwave [†]	open door - 100%, close door - 90%
pot/pan	stir - 85%
refrig.	open door - 100%, close door - 90%
stove [†]	clean surface - 79%, open oven - 90%, close oven - 77%, adjust temp controls (n.a.) - 95%
salt	shake - 72%
sink	adjust water flow (no act) - 100%, wash dish - 85%, grab rinse nozzle (no action) - 100%
toaster	no action, event only - 95%
Automobile Environment	
cup	drink - 85%
gearbox [†]	gear changes: neut. → 1 - 100%, 1 → 2 - 90%, 2 → 3 - 100%, 3 → 4 - 90%, 4 → 5 - 100%, neutral → reverse - 100%
lock	no action, event only - 100%
prkg brake	pull brake - 80%
radio	adjust (no action, event only) - 100%
strg wheel	turn left - 100%, turn right - 95%
temp ctrl	adjust (no action, event only) - 100%
window	roll up - 100%, roll down - 100%

Table 1. *Experiment I:* Office, kitchen, & automobile objects with associated actions and recognition accuracy, respectively. [†]Class with multiple or deformable shape states.

Experiment I: Capturing activities Office, kitchen, and automobile (interior) scenes were configured with articles and a person was invited to perform many common tasks and activities. We manually segmented and labeled 597 action examples from video for training the HMMs. Scripted sequences for each domain were used for testing. Accuracy of interactive events captured by the system are shown in Table 1. Percentages for objects with no associated actions were based on detectability alone; otherwise, recognition precision is shown. The following activities were summarized with the respective accuracy: **office** - reading (95%), coffee break (90%), computer-work (87%), taking phone message (60%), and counting documents (93%); **kitchen** - washing dishes (95%), cooking stir fry (87%), and cleaning kitchen (60%); **auto** - accelerating (shifting up) (95%), drinking-and-driving

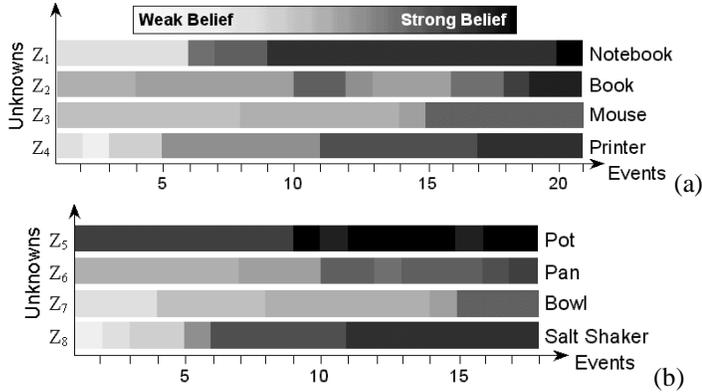


Figure 5. From experiment II, the strength of belief (proportional to grayscale) as image and action events are registered to one of four unknowns Z_i . (a) office (b) kitchen

(87%), winding road (93%), roll window down (100%), roll window up (100%), and parking car (85%). Throughout these interactions, 90% of all actions were recognized (assessed by hand-generated ground-truth observations). Using time-stamped logs, the system was also able to measure duration of hand-object contact. For the auto domain, a different person was used for testing than for training, illustrating person independent recognition of action (since only hand position is used).

Experiment II: Recognizing new objects To demonstrate detection and recognition of newly introduced objects, several objects (*book*, *notebook*, *printer*, and *mouse*) were carried into an **office** scene after the background was acquired. Table 1 shows the actions or events associated with each article. The system was already aware of several other objects in the room, including a keyboard, chair, and desk. Segmentation began immediately as initial image-based evidence of the unknown objects was acquired and initial beliefs were forged. As a person interacted with both known and unknown objects over several minutes, the strength of belief grew in proportion to the number of actions identified, as shown in Figure 5(a). During each *event* period, new object-, action- or image-based evidence is contributed from one of the four unknowns. Rough shape and size information was sufficient for establishing Z_2 and Z_3 early on (event 1). While relevant actions were able to classify Z_1 as a notebook by event 9, conflicting actions registered to Z_2 during events 12 and 13 compromised belief testimony. Although Z_3 (mouse) has no actions associated with it, moderate belief can still be established by monitoring its interaction with the keyboard (provided via **A**). In general, however, articles such as Z_3 stand a greater chance of being mis-labeled if actions associated with other GCMs are performed while interacting with it. Classification probability for Z_1 through Z_4 after 21 events (acquired over 1500

frames) was 97%, 94%, 80%, and 91%, respectively. Closer inspection reveals that 5, 8, 4, and 6 events for Z_1 through Z_4 , respectively, were needed to achieve this degree of classification.

A similar experiment was conducted in the **kitchen** domain with a *pot*, *pan*, *bowl*, and *salt shaker* added to the scene already initialized with a stove, a pan, and cabinets. Classification probability of these 4 unknowns after 18 events was 98% (resulting from a template match of the same object stored in the database), 85%, 77%, and 93%, respectively.

Experiment III: Object Recognition from action To evaluate the strength of action-based evidence, 11 action events that were acquired over 583 frames. Image evidence assisted in action recognition, but was not used to score GCMs during evaluation. Figure 6(a) shows the mean log probability of the candidate GCMs. Note that belief was shared between the GCM for notebook and book until event 7, when “write” was the most probable action observed, consequently rejecting book. Figure 6(b) shows the accumulated likelihoods of several actions as they occurred throughout this sequence. It also reveals the potential for actions to be confused. Note that some actions that never actually occurred, such as “erase,” have high, accumulated probabilities, suggesting that it is similar to several of the gestures performed. Also note that recognition is not affected by an object’s deformable structure.

Experiment IV: Recognizing objects in the background To demonstrate detection and recognition of objects in the background (full occlusion), we performed several eating actions (*stir*, *cut*, *feed*) in an undeclared space in the scene. When actions, such as stir, occur for more than one GCM, belief is shared. Without image-based segmentation, motion normalization suffers, resulting in lower action recognition rates and occasional mis-labeling. (Notice in events 8 and 9 in Figure 6(c), “open” and “erase” were mistaken for “feed” and “cut”.) The table GCM exhibits the strongest belief, as shown in Figure 6(c).

Over many of these tests, belief ranged from 5%-17% lower over the same number of action events when image information was acquired, but not used for scoring GCMs. With no image evidence acquired, action recognition suffered even more and belief estimates dropped 14%-33%. Under these circumstances, it is difficult to find an appropriate observation length to parse continuous motion, especially if rest states are not deliberate. However, these estimates can be raised by assessing belief over more action events.

6. Summary and Conclusions

We present a flexible approach for recognizing human activities as well as objects in the environment using motion

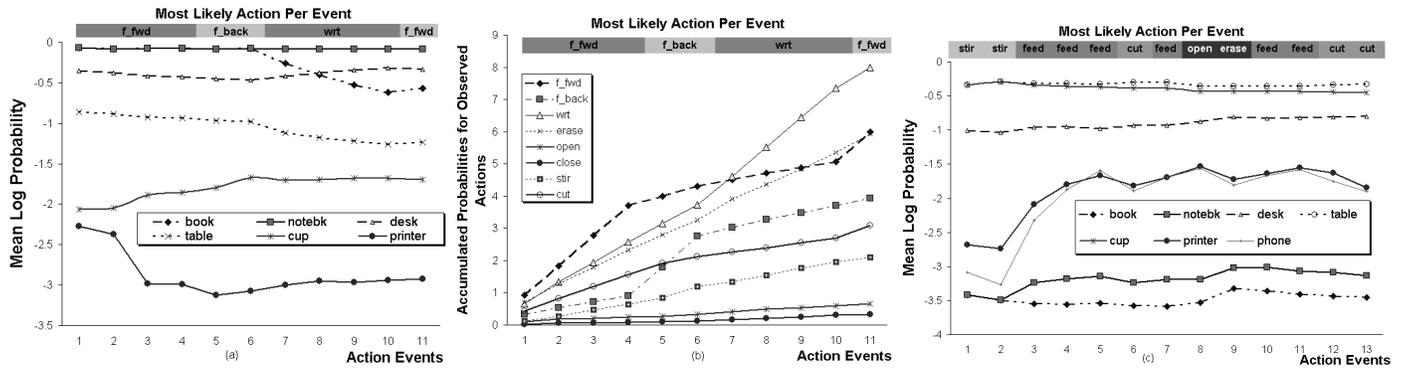


Figure 6. (a) from experiment III, mean log probability of GCM classification over several action events; (b) from experiment III, shows the accumulated likelihoods of several actions as they occurred throughout the corresponding sequence, with the most probable action per event highlighted (top); (c) from experiment IV, GCM Mean log probability of unknown object without image-based segmentation

and context. A hierarchical framework is applied to extract information about interactions between objects and people. HMMs are used for recognition of complex human actions because they can efficiently characterize motion profiles in spite of spatio-temporal variation. We use Bayesian approaches to label activities and to classify rigid and deformable objects by relying on actions and image clues, when available. We show that convergence toward proper GCM classification is contingent on the quality and consistency of the observed evidence. Overall recognition of 46 actions is 90.2%, tested on 450 sequences. Summarization accuracy of 14 activities averaged 87.6% and classification of 8 unknown objects (using all available evidence) measured 89.7%.

Although the single-camera approach is appropriate most of the time, non-planar, complex motion can become ambiguous from one perspective. Because only position information is used to model actions, single person training generalizes well for person-independent testing, barring wild variations. Although several scene articles have no actions associated with them or no hand-based motion can be adequately modeled, we have shown that exploiting the relationship between object pairs can be helpful for summarizing activities and discovering articles. Despite the scalability problems that can be caused by associating all of the possible actions with each object, we hope to limit these by working with well defined and constrained domains. We plan to extend this work by considering multitasked activities for multiple people.

References

- [1] J. K. Aggarwal, S. Shah, R. Chin, T. Pong, "Bayesian Paradigm for Recognition of Objects - Innovative Applications," *ACCV Proceedings: Third Annual Asian Conference on Computer Vision*, Vol. 2, pp. 275-82, Hong Kong, 1997.
- [2] A. Bobick, "Movement, Activity, and Action: The Role of Knowledge in the Perception of Model," *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, February 1997.
- [3] H. Buxton and S. Gong, "Advanced Visual Surveillance using Bayesian Networks," *International Conference on Computer Vision*, Cambridge, Mass., June 1995.
- [4] D. W. Etherington and Reiter, "On inheritance hierarchies with exceptions," *Proceedings AAAI-83: Third National Conference on Artificial Intelligence*, Washington, D.C., August 22-26, 1983.
- [5] N. Friedman and M. Goldszmidt, "Building Classifiers using Bayesian networks," *Proceedings: 13th National Conference on Artificial Intelligence*, Portland, Oregon, pp. 1277-1284, 1996.
- [6] X. D. Huang, Y. Ariki, and M. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.
- [7] R. Mann and A. Jepson, "Towards the Computational Perception of Action," *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 794-799, Santa Barbara, CA, 1998.
- [8] D. Moore, I. Essa, M. Hayes, "ObjectSpaces: Context Management for Action Recognition," *Proceedings of the 2nd Annual Conference on Audio-Visual Biometric Person Authentication*, Washington, D.C., March 1999.
- [9] N. Oliver, B. Rosario, and A. Pentland, "Statistical Modeling of Human Interactions," *Proceedings from the Computer Vision and Pattern Recognition*, 1998.
- [10] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc. San Mateo, California, 1988.
- [11] C. Pinhanez and A. Bobick, "Human Action Detection Using PNF Propagation of Temporal Constraints," *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 898-904, Santa Barbara, CA, 1998.
- [12] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 12, pp. 1371-5, December 1998.
- [13] S. Ullman, *High-level Vision: Object Recognition and Visual Cognition* MIT Press, 1996.
- [14] J. Yi, D. Chelberg, "Model-based 3D object recognition using Bayesian indexing," *Computer Vision and Image Understanding*, Vol. 69, No. 1, pp.87-105, January 1998.