A Scalable Approach to Activity Recognition based on Object Use

Jianxin Wu[1], Adebola Osuntogun[1], Tanzeem Choudhury[2], Matthai Philipose[2], and James M. Rehg[1]
[1] College of Computing, Georgia Institute of Technology [2] Intel Research Seattle
{wujx,osuntogu,rehg}@cc.gatech.edu {tanzeem.choudhury,matthai.philipose}@intel.com

Abstract

We propose an approach to activity recognition based on detecting and analyzing the sequence of objects that are being manipulated by the user. In domains such as cooking, where many activities involve similar actions, object-use information can be a valuable cue. In order for this approach to scale to many activities and objects, however, it is necessary to minimize the amount of human-labeled data that is required for modeling. We describe a method for automatically acquiring object models from video without any explicit human supervision. Our approach leverages sparse and noisy readings from RFID tagged objects, along with common-sense knowledge about which objects are likely to be used during a given activity, to bootstrap the learning process. We present a dynamic Bayesian network model which combines RFID and video data to jointly infer the most likely activity and object labels. We demonstrate that our approach can achieve activity recognition rates of more than 80% on a real-world dataset consisting of 16 household activities involving 33 objects with significant background clutter. We show that the combination of visual object recognition with RFID data is significantly more effective than the RFID sensor alone. Our work demonstrates that it is possible to automatically learn object models from video of household activities and employ these models for activity recognition, without requiring any explicit human labeling.

1. Introduction

We address the problem of recognizing human activities in indoor environments such as the kitchen or office. This problem has broad applications in computer-assisted care and workflow optimization, human-centered computing, and automated surveillance. To a first approximation, activities can be characterized by a set of "verbs", the actions performed by a human actor, and a set of "nouns", the objects or places which are the target of the action. While much work in activity recognition has focused on recognizing the verbs, our goal is to characterize activities by sensing the nouns. In other words, we recognize activities by identifying the objects which are being used in the scene.

An activity recognition approach based on *object use* can be particularly useful in domains such as cooking, which involve a relatively small number of repeated actions such as chopping, pouring, spreading, etc. Object use information can help discriminate between activities such as making toast and making a sandwich, which may be similar from the standpoint of the actions alone. Such distinctions can be important for application domains such as health monitoring or memory aids. A significant issue in the development of an object-based approach is its *scalability*, given the potentially large number of objects that must be discriminated, and the difficulty of obtaining labeled training data for each object under realistic conditions. Potential users are unlikely to be willing to spend a significant amount of time training a recognition system by presenting it with individually-labeled object instances. However, given video of everyday household activities, it is possible that object models could be extracted automatically if a sufficiently informative training signal was available.

We describe a method for activity recognition based upon *automatically-acquired* models of activities and the objects that they involve. Our learning approach is based upon two primary sources of information. The first is an RFID-based sensor which generates sparse and noisy object use events when a tagged object is manipulated during an activity. Correlations between events in the RFID and video streams are modeled using a Dynamic Bayesian network (DBN) and used as a training signal to automatically acquire object models. The second source of information is how-to websites such as about.com, which could be mined automatically to extract activity models [27]. Our DBN representation encodes this common-sense knowledge about which objects are used in various activities, *e.g.* making a cup of tea involves a teacup, a teabag, and hot water.

Our method builds upon recent techniques in the object recognition community for clustering object descriptors across a set of static images that have been labeled with the objects that they contain [23]. In our approach, RFID tags are attached to standard kitchen objects to obtain

sparse and noisy object-use information. An RFID reader bracelet worn on the user's wrist indicates whenever the user's hand is in close proximity to a tagged object. Simultaneously, video frames of the scene are acquired and segmentation techniques are used to generate candidate regions in the frame which may correspond to manipulated objects. SIFT features [13] are extracted from these regions and each object is modeled as a bag of SIFT features. The EM algorithm is used to estimate the DBN parameters that specify the distribution of the SIFT features for each object. Once the DBN has been trained, it can be used to jointly infer both activities and objects from novel video sequences.

The main contribution of this work is a scalable approach to activity recognition based on detecting object use that does not rely on explicit human labeling of sensor data. Specifically,

1. We show that activity recognition based on object-use information is viable, in that it results in good recognition performance under realistic imaging conditions.

2. We demonstrate a scalable approach to automatically learning object models from video using sparse and noisy RFID sensor data and common-sense knowledge of activities. Our object models are obtained without any form of manual labeling or intervention (either at the frame level or the region level).

3. We describe a DBN model that synergistically incorporates common-sense activity descriptions, RFID sensor events, and video data.

We have conducted an extensive set of experiments on the classification of kitchen activities carried out in two realistic settings. We demonstrate that the automatically learned vision-based object models can be successfully utilized to recognize activities and objects.

2. Related Work

Much early work on the analysis of activities took place within the computer vision community and leveraged video cameras as passive and non-invasive sensors [5, 6, 9, 14, 25]. More recently, alternative paradigms based on dense sensors have emerged [1, 18, 19]. In this approach, tiny battery-free wireless sensors are attached to objects and surfaces in a space and can provide direct measurements of the user's proximity to objects and regions of the environment. For example, [18] described a system based on RFID tags for recognizing a subset of the activities of daily living, a canonical activity recognition task for computer-assisted care applications. While dense sensors are appealing due to their low cost and simplicity, they have several drawbacks in comparison to video-based analysis. First they require objects and often people to be instrumented. Second, they do not work with certain types of objects such as metallic objects, food items, and objects that are very small. In addition there are problems with signal drop out, latency, and confusion between labels during reading.

Many activity recognition methods in computer vision have focused on the representation and modeling of actions, which are the atomic units within activities. This line of work explores tracking methods and other forms of spatio-temporal video analysis in order to sense what the actor is doing. In other words, they attempt to identify the activity by sensing the *verbs*. A common theme in these works is the exploration of spatio-temporal video features [3, 5, 11, 21, 25]. Other work has addressed the use of multiple resolutions [24]. Temporal constraints on actions are addressed either through the use of probabilistic models such as HMM's [4, 16] or SCFG's [9], or through explicit temporal correlation methods [21, 26].

The recognition of actions can often be aided by incorporating context from the environment. We loosely characterize these approaches as sensing *verbs plus context*. For example, in the work of Moore *et al*. [14] and Peursum *et al*. [17], actions can be discriminated by identifying the spatial location within the scene in which they occur. Hand motions which might be ambiguous in general can be correctly classified as typing when they occur in the vicinity of the keyboard. The W^4 system used outdoor scene context in conjunction with a robust blob-tracking algorithm to analyze scenarios in which multiple people interacted and exchanged bags and other objects [6]. Other sources of task-specific context include the identification of roads and entrances/exits in parking lot surveillance [8] and tracking the components of a blood glucose monitor in [22].

In contrast to these works, we are interested in domains where the action and spatial location are of limited utility in recognizing activities. Many different cooking activities, for example, involve picking up and putting down objects within a single counter-top area. To differentiate among these activities it is necessary to identify the objects which are being manipulated. We characterize this approach as recognizing activities by sensing the *object use* (*i.e*. the nouns). The cooking domain involves a large number of different objects which are shared across multiple activities and are not restricted to any particular location in the image. A major challenge in this approach is the need for a robust general-purpose object recognition system which could reliably discriminate between hundreds of different cooking items under real-world imaging conditions.

Building models for object recognition usually requires labeled training images without a cluttered background. In order to obtain this, a significant amount of work (segmentation and labeling of objects) is required. In contrast, our work leverages temporal continuity in video frames to roughly segment the moving object. An object is modeled

as a bag of SIFT features and learning object models is equivalent to assigning the probabilities of seeing different SIFT features in an object. Our approach is similar to [23], which assigned features from independent images into 'topics' using pLSA, an unsupervised learning method. Their results showed that the revealed topics usually coincided with objects. In contrast, we use sparse and noisy RFID measurements to guide the learning process.

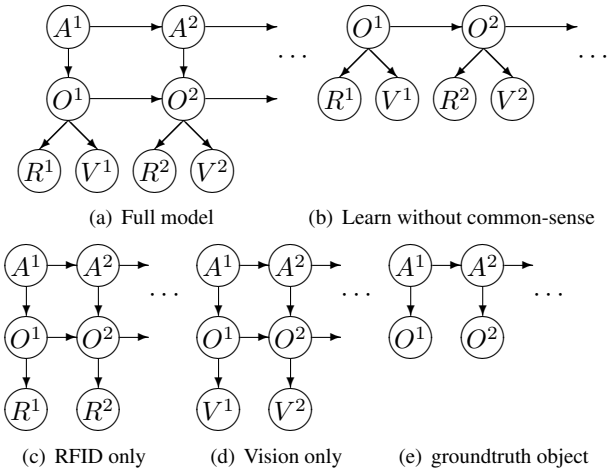
Recently, dense sensors have been proposed as an alternative to vision-based object recognition for obtaining object information in activity recognition tasks. In these works, wireless sensors attached to both humans and objects make it possible to directly measure actor-object interactions. Possible sensor data includes the identities of people and objects (*e.g*. RFID sensors) as well as their positions and velocities (*e.g*. accelerometers and audio sensors). In RFID-based systems, activities are represented as probability distributions over sequences of object-use [18] obtained from sparse and noisy RFID readings. In other approaches, information from accelerometers is used to identify actions such as walking and climbing stairs [1, 20]. RFID and vision were also used as complementary sensors. In [12], RFID and vision were used to track object and human independently, and were combined using rules. In contrast to this latter work, we utilize RFID sensors to fit vision object models in an integrated DBN framework.

Another aspect of activity recognition which has received significant attention is computational models of activity which can serve as a constraint on the interpretation of noisy sensor data. Complex activities such as baking a cake can be decomposed into subtasks, and constraints from the domain (*e.g*. the oven must be preheated before it can be used) result in partial orderings of these subtasks. There has been much interesting work in representing and exploiting these constraints during recognition [7, 15, 22].

3. Object-use Based Activity Recognition

Our object-use based activity model is depicted in figure 1(a) as a DBN, in which A^t, O^t, R^t, and V^t represent activity, object, RFID, and video frame respectively. The superscript t indexes video frames and standard first-order Markov assumptions are made for A^t and O^t. The DBN model is fully specified by the following parameters: the prior distribution $P(A^1)$, the observation model $P(O^1|A^1)$ and $P(O^{t+1}|O^t, A^{t+1})$, the state transition model $P(A^{t+1}|A^t)$, and the output model $P(R^t|O^t)$ and $P(V^t|O^t)$. Tables 1 and 2 list the activities and objects used in our experiment (*c.f*. page 5). For example, $A^5 = \mathbf{b}, O^5 = \mathbf{3}$ means that at frame 5, the activity is make tea, and the object is teabag.

In this section we introduce the RFID sensors, present how to learn the object model $P(V^t|O^t)$ using RFID readings without any manual labeling, and how the learned models can improve both activity and object recognition. Except for the object model $P(V^t|O^t)$, all other parameters can either be specified from domain knowledge (*c.f*. Sec. 3.5) or generated automatically by mining common-sense knowledge from web as described in [27].



(a) Full model (b) Learn without common-sense

(c) RFID only (d) Vision only (e) groundtruth object

Figure 1. Various graphical models for activity and object recognition.

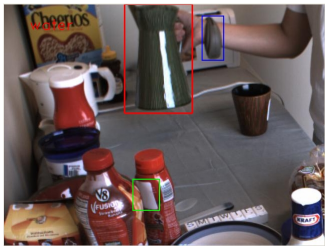


Figure 2. A typical kitchen setup. The user is manipulating a water jug (red rectangle) while wearing an RFID bracelet (blue rectangle). Some objects have RFID tags attached (green rectangle).

3.1. Sensors and setup

Figure 2 illustrates our setup. One camera overlooks the space, in this case a kitchen counter, where the activities take place. Users wear RFID bracelets on their dominant hands. A bracelet incorporates an RFID reader, battery, and radio. RFID tags are attached to some objects in the space. These tags are postage-stamp to credit-card sized, battery-free, 40-cent stickers available off the shelf. When the user handles a tagged object, the bracelet scans ID of the tag and sends it wirelessly to a nearby computer that maps the ID to an object name.

Although RFID can sense the use of objects, in practice it has several limitations which motivate us to bootstrap the RFID readings using vision. If the bracelet is close to an object by accident, it may indicate erroneously that the object is being manipulated. If a tagged object is grasped far from the tag, on the other hand, the manipulation may be missed.

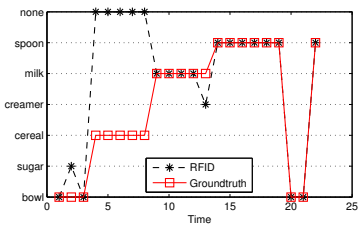


Figure 3. Erroneous object-use detection with RFID.

Consider the RFID trace depicted in figure 3, which was obtained during the making cereal activity. In this example, the RFID reader erroneously detected the use of sugar and creamer, and missed the use of the cereal box. Furthermore, some objects such as toothbrush are too small to be tagged. In many cases the manipulating hand may be bracelet-free: only some users in a space may wear bracelets, and even bracelet wearers may use non-dominant hands or wear the bracelets intermittently. We believe therefore that it is useful to have a sensing modality that can exploit RFID-based object-use data when it is (sporadically) available, but can detect objects even when it is not available.

3.2. Segmentation using change detection

We assume that the object currently being moved is always the object in use, and utilize change detection to segment the object. Since some kitchen utensils are textureless, simple change detection by subtracting pixel values will have difficulty handling these objects. We group pixels into 8x8 square superpixels, thus a 640x480 frame is converted into 80x60 superpixels. The difference between two superpixels $\mathbf{x}$ and $\mathbf{y}$ at the same position in two frames is calculated as one minus the normalized dot product $1 - \frac{\sum_{i=1}^{64} \mathbf{x}_i \mathbf{y}_i}{\sqrt{\sum_{i=1}^{64} \mathbf{x}_i^2} \sqrt{\sum_{i=1}^{64} \mathbf{y}_i^2}}$. A frame t is compared with both frames $t+3$ and $t-3$. Superpixels that have both differences bigger than a predefined threshold is classified as have changed and belong to the object. One failure mode of this approach is the user's hands and arms. They always move along with the object, and should be excluded. We use skin color detection to detect skin colored pixels and remove them [10]. Using change detection in the superpixel mode, we can get a roughly correct segmentation for the object currently being used, which constitutes the set V^t for frame t. As appeared in figure 4, segmentation is difficult for some objects, *e.g.* the spoon.

3.3. Object Model Representation

Given a video frame, SIFT features that are within the segmented area are extracted. The object in this frame is modeled as a bag of these SIFT features. In this model, a video frame V^t consists of a set of SIFT features $v_1^t, v_2^t, \ldots, v_{n^t}^t$, where v_i^t and v_j^t are supposed to be conditionally independent for $1 \leq i, j \leq n^t, i \neq j$. Thus,[1]

Figure 4. Examples of segmentation results, with the left and right row being original frame and segmented result, respectively. Only pixels in the biggest connected component (depicted by the rectangle) are used. The three rows are examples of good, moderate, and bad segmentation results, respectively.

$$P(V^t|O^t) = \prod_{i=1}^{n^t} P(v_i^t|O^t) = \prod_{i=1}^{n^t} \mathbf{h}_{O^t}(v_i^t) \qquad (1)$$

in which $\mathbf{h}_{O^t}(v_i^t)$ is the probability of observing a SIFT feature v_i^t when the object is O^t, *i.e.* a histogram of SIFT features. Thus, learning object model requires specifying the values $\mathbf{h}_{O^t}(v_i^t)$ for all possible values of O and v.

Since the number of possible SIFT features is vast. A vector quantization procedure same as that used in [23] is used. 2000 clusters are constructed using the k-means algorithm from about 600k SIFT features collected from images of the objects of interest. A SIFT feature is then represented by the closest cluster center.

3.4. Learning object models w/o human labeling

The maximum likelihood estimates for histogram parameters are the counts of features falling into a bin, divided by the total number of features. Thus, if we are given the object identities O^t, learning is simply a counting procedure. However, one of our goals is to avoid any manual labeling and we have no direct access to O^t. Instead, the RFID readings are used in the EM algorithm to learn the object models.

RFID readings are sparse in comparison to video frames and very noisy, *i.e.* we only have a few RFID labels and these labels are not reliable (*c.f.* figure 3). The idea is to use the common-sense knowledge and temporal continuity of

[1] In practice we use $P(V^t|O^t) = \sqrt[n^t]{\prod_{i=1}^{n^t} \mathbf{h}_{O^t}(v_i^t)}$ in order to be fair to objects containing different numbers of features.

object use to infer object identity in each frame. Precisely, we build a DBN that models the relationship between activities, objects, frames, and RFID readings, as shown in figure 1(a). RFID readings are viewed as evidences to a small subsets of the R^t nodes, and the noise in RFID is modeled by $P(R^t|O^t)$. In learning the object model $P(V^t|O^t)$ using the EM algorithm, all other parameters are specified from domain knowledge and fixed in the learning process.

The E-step is then to estimate the marginal distribution of O^t given the R^t evidences, the video frames V^t, and the current $\mathbf{h}_{O^t}(v_i^t)$. The standard junction tree algorithm is used to infer the marginal distributions of O^t for every t. Since V^t is independent of A^t, given the marginal distribution of O^t, the M-step is a simple counting procedure that updates $\mathbf{h}_{O^t}(v_i^t)$. The E-step and M-step are iterated until the log-likelihood of the DBN converges.

We should highlight that object models are learned on the fly which makes it very easy to install our system in a new environment without requiring any user-specific training. However, it is also possible to use previously learned object models directly in a new video (*c.f*. Sec. 4.2).

3.5. Specify parameters from domain knowledge

Parameters beside $P(V^t|O^t)$ can easily be specified by a human expert or automatically mined using techniques proposed in [27]. We specify these parameters as follows, based on domain knowledge and our assumptions.

The prior $P(A^1)$ is set to be uniform. $P(O^t|A^t)$ is defined as follows

$$P(O^t|A^t) = \begin{cases} \frac{2}{n_O} & \text{if } O^t \text{ is used in } A^t, \\ \frac{1-2n_{A^t}/n_O}{n_O - n_{A^t}} & \text{otherwise} \end{cases}$$

where n_O and n_{A^t} are the number of total objects and the number of objects used in A^t, respectively. This choice of parameter values ensures that an object used in an activity will have higher probability than an object not involved in the same activity.

$P(O^{t+1}|O^t)$ is set to be θ_O if $O^{t+1} = O^t$, and $\frac{1-\theta_O}{n_O-1}$ if otherwise. θ_O is usually set to a large number, in order to reflect the fact that an object is usually used in consecutive frames. The CPT $P(O^{t+1}|O^t, A^{t+1})$ is then specified as

$$P(O^{t+1}|O^t, A^{t+1}) \propto P(O^{t+1}|A^{t+1})P(O^{t+1}|O^t) \quad (2)$$

$P(A^{t+1}|A^t)$, the state transition model, is set to be θ_A if $A^{t+1} = A^t$, and $\frac{1-\theta_A}{n_A-1}$ if $A^{t+1} \neq A^t$, where n_A is the number of possible activities. Again, θ_A is usually large since an activity will span multiple frames.

The final CPT models RFID noise as $P(R^t|O^t) = \theta_R$ if $R^t = O^t$, and $\frac{1-\theta_R}{n_O-1}$ if $R^t \neq O^t$. The fact that $\theta_R \neq 1$ encodes noise in RFID readings.

It is also possible to learn all the above parameters from data, using the same EM algorithm. However, our experiments showed that learning these parameters lowers activity and object recognition rates. This is likely due to the fact that the initial estimates based on domain knowledge is a good starting point and in the absence of a large amount of training data these priors achieve better generalization performance during testing.

a.boil water	**e**.cheese sandwich	**i**.tend plants	**m**.make popcorn
b.make tea	**f**.buttered toast	**j**.take medicine	**n**.drink juice
c.make cereal	**g**.peanut butter sand.	**k**.make salad	**o**.wipe counter
d.make coffee	**h**.pack lunch	**l**.make TV dinner	**p**.phone call

Table 1. List of 16 activities.

1.water jug	**8**.cereal	**15**.knife	**22**.plant	**29**.microwave
2.kettle	**9**.bowl	**16**.toaster	**23**.plant care	**30**.popcorn
3.teabag	**10**.coffee	**17**.plate	**24**.watering	**31**.juice
4.cup	**11**.creamer	**18**.butter	**25**.pill box	**32**.cloth
5.spoon	**12**.sugar	**19**.peanut but.	**26**.salad tosser	**33**.phone
6.milk	**13**.cheese	**20**.jelly	**27**.dressing	
7.honey	**14**.bread	**21**.lunch bag	**28**.meat in box	

Table 2. List of 33 objects.

4. Experimental Results

The proposed framework was tested using videos containing 16 daily kitchen activities involving 33 objects. The list of activities and objects are shown in tables 1 and 2. Videos 1 and 2 were collected at Intel Seattle lab on different days. An additional Video 3 was collected at the Aware Home in Georgia Tech. Video 3 contained 13 activities and 28 objects, in which the objects were different than those used in Videos 1 and 2.

For testing purposes, we manually labeled the activity and object in every frame.[2] After the most probable sequence of A^t and O^t are inferred, they are compared to the groundtruth. The activity and object recognition rates are then computed as the percentage of frames in which the activity or object is predicted correctly.

The viability of object-use based activity recognition was verified in three ways. First it was tested empirically by learning object models from Videos 1. The learned models were then used to recognize activities and objects in the same video (Sec. 4.1). The same experiments were repeated using Videos 2 and 3. Second the models learned from Video 1 were used to recognize activities and objects in Video 2 without fitting the models again, *i.e*. generalization was tested in Sec. 4.2. Third, the approach was evaluated under ideal conditions, when groundtruth object labels are provided in every frame of the test video (Sec. 4.3). These results demonstrate the viability of object-use based approach and our automatic object model acquisition method. Furthermore, the proposed approach exhibited tolerance to adverse conditions, *e.g*. in case when some RFID tags were missing (Sec. 4.4). Finally, we showed that the common-

[2]We assumed there was one active object in each frame. In the small fraction of frames where multiple objects were used, we only chose one.

Common sense used	Testing sensors	Activity	Object
Yes	RFID only	64.31%	63.00%
Yes	RFID+Vision	80.67%	72.36%
	Vision only	80.97%	73.30%
No	RFID+Vision	60.84%	74.68%
	Vision only	62.76%	74.72%

Table 3. Activity and object recognition rates using different sensors, and different learning methods.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
a	35	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	96	0	4	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	99	0	0	0	0	1	0	0	0	0	0	0
f	0	0	0	0	0	58	41	0	0	1	0	0	0	0	0	0
g	0	0	0	0	0	41	59	0	0	0	0	0	0	0	0	0
h	18	45	0	0	0	0	0	37	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
l	0	0	0	0	4	0	0	0	0	0	0	96	0	0	0	0
m	0	0	0	0	6	0	0	14	0	0	0	0	79	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
o	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p	0	91	0	0	0	0	0	0	0	0	0	0	0	0	0	9

Table 4. Confusion matrix for activity recognition using Vision only. Numbers are shown in percentage.

sense knowledge incorporated in the DBN framework was essential for recognizing activities (Sec. 4.5).

4.1. Effect of learning object models

The first set of experiments examine the effect of the object models bootstrapped from RFID readings. After fitting object models, we test the performance under three conditions when inferring activity and object labels: using both sensors (figure 1(a)), RFID alone (figure 1(c)), and vision alone(figure 1(d)). Both activity and object recognition rates are shown in the first three rows of Table 3. In this set of experiments object models are learned with commonsense knowledge (figure 1(a)).

We learn object models using our unsupervised approach described in Sec. 3. We then use the learned models to infer activity and object labels. Video 1 is used in both stages. Although the object models are fitted from the noisy and sparse RFID readings without any manual labeling, the object models alone can recognize 80.97% activities and 73.30% objects, which is significantly better than the RFID only results (64.27% and 63.00%). The improvement in recognition rates with the addition of the vision sensor demonstrates the effectiveness of the learned object models. It is also worth noting that the vision only results are indistinguishable from the results when both RFID and vision are used (80.67% and 72.36%). This observation suggests that our DBN framework effectively utilized all the useful information in RFID readings. After the object models are learned, RFID readings no longer provide any useful information.

The detailed activity recognition confusion matrix is shown in table 4 for the vision only case. Table 4 reveals that the activities **a**.boil water, **o**.wipe counter, and **p**.phone call have very low recognition rates. This is not surprising since objects involved in these activities (water jug, cloth, and phone) are all white and textureless, thus indistinguishable. Another group of activities, **e**,**f**,**g**, all deal with different sandwiches and are inherently error prone. In addition the lunch bag is almost textureless. Furthermore, the RFID readings are sparse for these examples: there is only one RFID reading of kettle and none of cloth. This may explain why these activities had low recognition rates.

We repeated the same experiments using Videos 2 and 3. The RFID only activity and object recognition rates on Video 2 are 80.33% and 66.54% respectively. By learning object models from Video 2, RFID+Vision improves the rates to 82.52% and 80.16%. The Vision model alone achieves 82.53% and 80.29% recognition rates.

Activities in Video 3 are relatively easier to recognize than those in Videos 1 and 2. Using RFID alone, the activity and object recognition rates in Video 3 were already 88.55% and 67.49%. Many objects in Video 3 are textureless, and make them difficult for vision based object modeling. RFID+Vision had lower recognition rates (83.10% and 60.34%). However, it is worth noting that even with the difficulty in visually modeling objects, the vision sensor alone could still recognize activities in 83.42% of the frames, but only 56.40% for the objects.

4.2. Generalizability of the learned models

In previous experiments, we fit the object models and then infer activity and object labels using the same video. In real applications we are interested in training models on a representative corpus of unlabeled video and then applying those models to new videos without retraining. In order to evaluate the ability of our models to generalize to new videos, we conducted a second experiment. We learn object models using Video 1, and test the models on Video 2.

Using the object models learned from Video 1 and the RFID readings accompanying Video 2, RFID+Vision achieved activity and object recognition rates of 73.37% and 71.02% respectively. Comparing with the RFID only rates on Video 2 (80.33% and 66.54%), RFID+Vision had higher object recognition rate but lower activity recognition rate. Confusion matrices revealed that although on average RFID+Vision recognized objects better, all occurances of **2**.kettle and **21**.lunch bag were misclassified (both objects were textureless). Thus activities "boil water" and "pack lunch" were completely misclassified. Excluding these 2 activities, RFID+Vision had higher activity recognition rate than RFID only (81.91% vs. 71.72%). Furthermore, using models learned from Video 1, Vision only achieved on Video 2 recognition rates of 59.22% and 56.43% respec-

tively, which was significantly better than chance (chance probability is 6.67% and 3.33% respectively). Thus, the learned models can be used directly in new videos in case RFID is not available (*e.g*. user not wearing bracelet), and can improve recognition when RFID is available.

4.3. Limit of performance in ideal situations

The above experiments showed that the object-use based activity recognition worked well in practice. Our next experiment evaluates the viability of this approach in ideal situations, *i.e*. given the groundtruth object identity information in every video frame, how well can the system perform?

Given the groundtruth object information, the system is modeled as an HMM (figure 1(e)), and activities are correctly recognized in 90.29% of the frames. The RFID readings are sparse and noisy, which makes the RFID only activity recognition rate (64.27%) significantly lower than the recognition rate using groundtruth information (90.29%). However, by learning object models, the vision based recognition rate reached 80.67%. The gap between the real world and empirical best case performance was greatly reduced.

4.4. Dealing with missing RFID tags

Another important benefit of combining RFID and vision is to deal with missing RFID tags, *e.g*. objects that are not RFID taggable. Common-sense knowledge can be used to infer the existence of certain objects that are not tagged. For example, the combination of bowl and milk would suggest a nearby 'cereal' object, which is then reflected in the marginals computed in the E-step. The learned object models would then locate the frames with the "cereal" object.

The intuition sketched above is verified by experiments. For example, the cereal object does not have any RFID readings in our data set (*c.f*. figure 3). In the RFID only method, none of the 123 frames containing the cereal object are recognized correctly. After object models are learned using the common-sense knowledge (*i.e*. when object models are learned using figure 1(c)), all 123 frames of cereal are recognized successfully. In contrast, if the object models are learned without common-sense knowledge (figure 1(b)), no frame of cereal is recognized.

To explore this issue further we conducted an additional experiment in which increasing numbers of RFID tags were systematically removed. Our goal was to study the impact of missing tags on the overall performance. Figures 5 and 6 show activity and object recognition rates respectively, when a fixed number of RFID tags were removed.

The x-axis in figures 5 and 6 is the number of missing tags. For a fixed number of K missed tags, we removed K tags from randomly chosen objects. We experimented with $1 \leq K \leq 20$. For any fixed value of K, the experiment was repeated 10 times. Figures 5 and 6 show the average recognition rates and one standard deviation error bars.

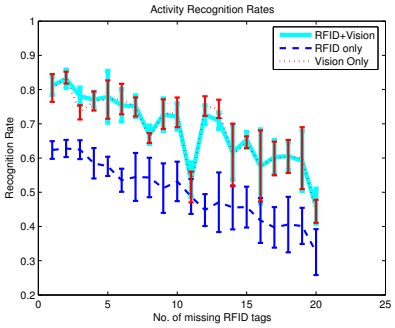


Figure 5. Activity recognition rates with different number of missing RFID tags.

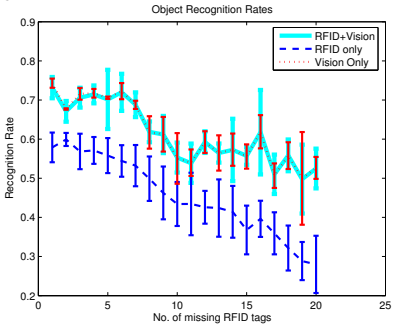


Figure 6. Object recognition rates with different number of missing RFID tags.

We observe in figures 5 and 6 that the vision only results are almost the same as RFID+Vision results. Recognition rates in both cases are significantly higher than those in the case only RFID is used. The system performance degrades smoothly when the number of missing tags increases.

4.5. Importance of common-sense knowledge

The common-sense knowledge is encoded in the edge from A^t to O^t in figure 1(a), *i.e*. which objects are likely to be used in a given activity. Its importance can be verified by comparing systems in which the common-sense knowledge is present or absent during learning object models. When common-sense knowledge is removed, the DBN in figure 1(a) becomes the HMM depicted in figure 1(b). We use the standard EM algorithm in this HMM to fit the object models. After the object models are learned, activity and object labels are inferred using figure 1(a).

Recognition rates of models learned using this DBN are shown in the last two rows of Table 3. When the common-sense knowledge is missing, the learned object models still have roughly the same object recognition rates. However, the activity recognition rates drop significantly to the level when only RFID is available.

5. Conclusions and Future Work

We presented a scalable approach to recognizing activities by recognizing the objects that are manipulated in these

activities. We proposed a framework that can automatically learn object models from video using sparse and noisy RFID readings and common-sense knowledge. A dynamic Bayes network was designed to systematically incorporate common-sense knowledge, the RFID sensor data, the vision sensor data, and time continuity in these sensors. Using the DBN framework, learning object models is naturally formulated as an Expectation-Maximization problem.

Our experiments validate the object-use based activity recognition approach in both realistic and ideal situations. In a realistic kitchen setup involving 16 activities and 33 objects, activities were correctly recognized in 80.97% of the video frames using the automatically learned object model, and objects were recognized in 73.30% of the frames. In addition, the activity recognition rate was 90.29% if groundtruth object labels were given during testing. Our experiments also showed that the learned object models can be used directly in recognizing activities and objects in new video if RFID is missing, and can improve recognition rates if RFID is available.

There are a few research directions that will improve the proposed approach. Faster features (*e.g*. SURF features [2]) could be used to replace SIFT features. Further recognition of human hand motion and interactions with objects in video frames may also reveal actions (*e.g*. chop, scoop), which are useful to activity analysis. With actions recognized, subtasks and partial ordering constraints could also be applied.

References

[1] L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive*, volume 3001 of *Lecture Notes in Computer Science*, pages 1–17, 2004.

[2] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *ECCV (1)*, pages 404–417, 2006.

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. PAMI*, 23(3):257–267, 2001.

[4] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *Proc. CVPR*, volume 1, pages 838–845, 2005.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. ICCV*, pages 726–733, 2003.

[6] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Trans. PAMI*, 22(8):809–830, 2000.

[7] S. Hongeng, R. Nevatia, and F. Brémond. Video-based event recognition: activity representation and probabilistic recognition methods. *CVIU*, 96(2):129–162, 2004.

[8] Y. Ivanov, C. Stauffer, A. Bobick, and W. Grimson. Video surveillance of interactions. In *2nd IEEE workshop on Visual Surveillance*, pages 82–89, 1999.

[9] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. PAMI*, 22(8):852–872, 2000.

[10] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *IJCV*, 46(1):81–96, 2002.

[11] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. ICCV*, pages 166–173, 2005.

[12] N. Krahnstoever, J. Rittscher, P. Tu, K. Chean, and T. Tomlinson. Activity recognition using visual tracking and rfid. In *WACV/MOTION'05*, pages I: 494–500, 2005.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. In *Proc. ICCV*, volume 1, pages 80–86, 1999.

[15] C. J. Needham, P. E. Santos, D. R. Magee, V. E. Devin, D. C. Hogg, and A. G. Cohn. Protocols from perceptual observations. *Artif. Intell.*, 167(1-2):103–136, 2005.

[16] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *CVIU*, 96(2):163–180, 2004.

[17] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *Proc. ICCV*, volume 1, pages 82–89, 2005.

[18] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, 2004.

[19] R. Raskar, P. A. Beardsley, J. van Baar, Y. Wang, P. H. Dietz, J. C. Lee, D. Leigh, and T. Willwacher. Rfig lamps: interacting with a self-describing world via photosensing wireless tags and projectors. In *Siggraph*, 2004.

[20] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman. Activity recognition from accelerometer data. In *AAAI*, pages 1541–1546, 2005.

[21] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. CVPR*, pages I:405–412, 2005.

[22] Y. Shi, A. Bobick, and I. Essa. Learning temporal sequence model from partially labeled data. In *Proc. CVPR*, volume 2, pages 1631–1638, 2006.

[23] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, volume 1, pages 370–377, 2005.

[24] P. Smith, M. Shah, and N. da Vitoria Lobo. Integrating and employing multiple levels of zoom for activity recognition. In *Proc. CVPR*, volume 2, pages 928–935, 2004.

[25] J. Sullivan and S. Carlsson. Tracking and labelling of interacting multiple targets. In *ECCV (3)*, pages 619–632, 2006.

[26] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury. The function space of an activity. In *Proc. CVPR*, volume 1, pages 959–968, 2006.

[27] D. Wyatt, M. Philipose, and T. Choudhury. Unsupervised activity recognition using automatically mined common sense. In *AAAI*, pages 21–27, 2005.