Privacy-preserving Recognition of Activities in Daily Living from Multi-view Silhouettes and RFID-based Training

Sangho Park and **Henry Kautz**
Computer Science, University of Rochester
Rochester, NY 14627, USA
{spark | kautz}@cs.rochester.edu

Abstract

There is an increasing need for the development of supportive technology for elderly people living independently in their own homes, as the percentage of elderly people grows. A crucial issue is resolving conflicting goals of providing a technology-assisted safer environment and maintaining the users' privacy. We address the issue of recognizing ordinary household activities of daily living (ADLs) by exploring different sensing modalities: multi-view computer-vision based silhouette mosaic and radio-frequency identification (RFID)-based direct sensors. Multiple sites in our smart home testbed are covered by synchronized cameras with different imaging resolutions. Training behavior models without costly manual labeling is achieved by using RFID sensing. Privacy is maintained by converting the raw image to granular mosaic, while the recognition accuracy is maintained by introducing the multi-view representation of the scene. Advantages of the proposed approach include robust segmentation of objects, view-independent tracking and representation of objects and persons in 3D space, efficient handling of occlusion, and the recognition of human activity without exposing the actual appearance of the inhabitants. Experimental evaluation shows that recognition accuracy using multi-view silhouette mosaic representation is comparable with the baseline recognition accuracy using RFID-based sensors.

Introduction and Research Motivation

There is an increasing need for the development of supportive technology for elderly people living independently in their own homes, as the percentage of the elderly population grows. Computer-based recognition of human activities in daily living (ADLs) has gained increasing interest from computer science and medical researchers as the projected care-giving cost is expected to increase dramatically. We have built the Laboratory for Assisted Cognition Environments (LACE) to prototype human activity recognition systems that employ a variety sensors. In this paper, we address the task of recognizing ADLs in a privacy-preserving manner in next-generation smart homes, and present our ongoing research on Assisted Cognition for daily living.

Our system uses multiple cameras and a wearable RFID reader. The cameras provide multi-scale and multi-view synchronized data, which enables robust visual recognition in the face of occlusions and both large and small scale motions. A short-range, bracelet form factor, RFID reader developed at Intel Research Seattle (*iBracelet*) remotely transmits time-stamped RFID readings to the vision system's computer. RFID tags are attached to various objects including furniture, appliances, and utensils around the smart home. Although we currently use commercial-quality cameras and a high-end frame grabber to integrate the video feeds, the decreasing cost of video cameras and the increasing power of multicore personal computers will make it feasible in the near future to deploy our proposed system with inexpensive cameras and an ordinary desktop computer.

Previous approaches to recognizing ADLs have depended upon users wearing sensors (RFID and/or accelerometers) (Patterson et al. 2005), audio-visual signals (Oliver et al. 2002) or using a single camera vision system (Abowd et al. 2007; Mihailidis et al. 2007). Recently, (Wu et al. 2007) employed a combination of vision and RFID. The system was able to learn object appearance models using RFID tag information instead of manual labeling. The system is, however, limited by a single camera view, which entails view dependency of the processing. The system also did not attempt to model or learn the *motion* information involved in performing the ADLs. We propose a multi-sensor based activity recognition system that uses multiple cameras and RFID in a richer way.

Understanding human activity can be approached from different levels of detail: for example, a body transitioning across a room at a coarse level, versus hand motions manipulating objects at a detailed level (Aggarwal and Park 2004). Our multi-camera based vision system covers various indoor areas in different viewing resolutions from different perspectives. RFID tags and reader(s) detect without false positives the nearby objects that are handled by the user. The advantages of such a synergistic integration of vision and RFID include robust segmentation of objects, view-independent tracking and representation of objects and persons in 3D space, efficient handling of occlusion, efficient learning of temporal boundary of activities without human intervention, and the recognition of human activity at both a coarse and fine level.

System Architecture Overview

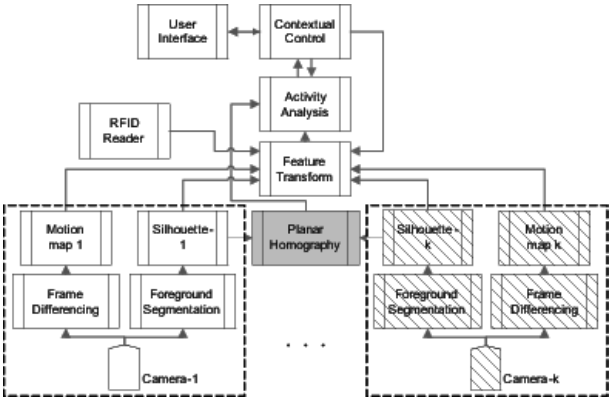

Figure 1: The overall system architecture of the Multi-scale multi-perspective vision system.

Fig. 1 shows the overall system architecture. The clear modules compose the basic single-view system, while the hashed modules compose the multi-view functionality. Activity analysis is performed with the features transformed from the input modules. The planar homography module locates persons for activity analysis. The dashed enclosing boxes indicate that the silhouette and motion-map generation processes could be performed at hardware level with infrared or near-infrared cameras to ensure the privacy.

In multi-view mode, the foreground image is redundantly captured to represent the three-dimensional extension of the foreground objects. Using multiple views not only increases robustness, but also supports simple and accurate estimation of view-invariant features such as object location and size.

Multiple View Scene Modeling

Contrary to single camera systems, our multi-camera system provides view-independent recognition of ADLs. Our vision system is composed of two wide field-of-view (FOV) cameras and two narrow FOV cameras, all synchronized. The two wide FOV cameras monitor the whole testbed and localize persons' positions in the 3D space based on a calibration-free homography mapping. The two narrow FOV cameras focus on more detailed human activities of interest (*e.g.*, cooking activities at the kitchen countertop area in our experiments).

Currently, four cameras are used (Fig. 2) to capture the scene from different perspectives. Two wide-FOV cameras (in darker color) form the approximately orthogonal viewing axes to capture the overall space (views 1 and 2 in Fig. 3), while the two narrow-FOV cameras (in lighter color) form the approximately orthogonal viewing axes to capture more details of certain focus zones such as the kitchen (views 3 and 4 in Fig. 3.) The four synchronized views are overlaid with a virtual grid to compute scene statistics such as pixel counts in each grid cell. Both the track and body-level analysis can be used for the activity analysis depending upon the analysis tasks. In this paper, we focus on multi-view silhouette mosaic representations of detected foreground objects for privacy-preserving recognition of activities.

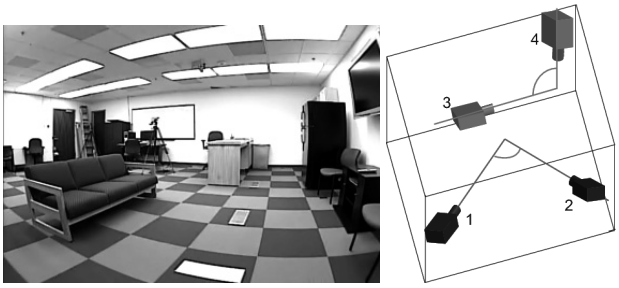

Figure 2: Multi-camera configuration. Two wide-FOV cameras (in darker color) and two narrow-FOV cameras (in lighter color) form the approximately orthogonal viewing axes, respectively.

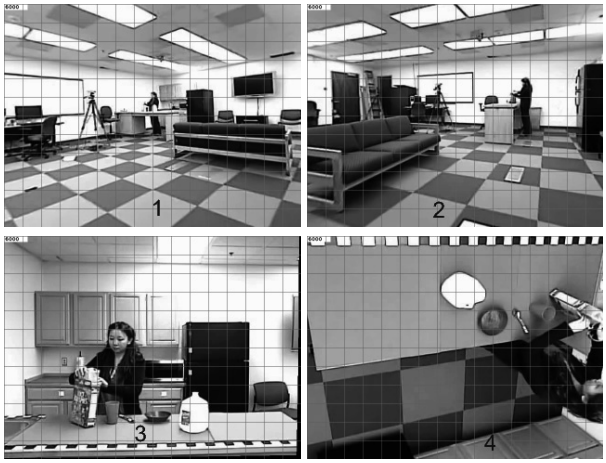

Figure 3: Distortion-compensated multi-view images 1 to 4 of the ADLs experiment in which a person performs the *prepare cereal* activity. The virtual grid is overlayed to compute scene statistics.

In Fig. 1, dynamic contextual control with optional user involvement can be incorporated with activity analysis, and provides constraints to other processing modules as feedback. The top-down feedback flows in the system are marked as red arrows.

Appearance-Based Segmentation and Tracking

ADLs may involve multiple objects moving simultaneously (Fig. 3), which can create challenges for a vision system — for example, changing backgrounds and object occlusion.

We adopt a dynamic background model using K-means clustering (Kim et al. 2005). The background model is updated with a memory decay factor to adapt to the changes in the background, and foreground-background segmentation is achieved at each pixel.

Silhouette representation of the foreground regions may obscure object appearances to preserve privacy (Fig. 4), but it also loses the useful information about object identities. RFID reading provides a smart way for identifying the ob-

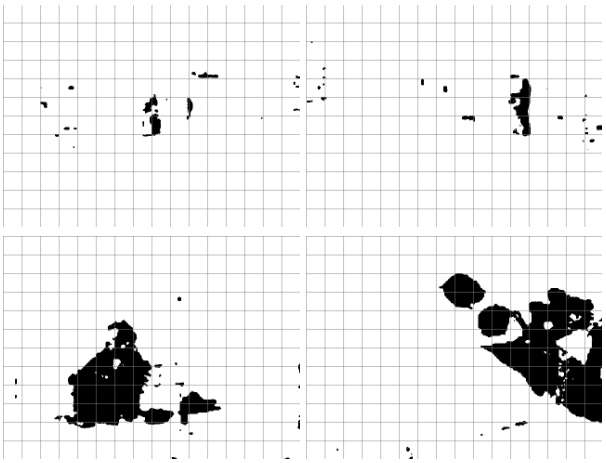

Figure 4: Binary foreground maps corresponding to the multi-view images in Fig. 3. The whole image forms the *super foreground map* Γ_t. Black areas represent effective foreground regions (*i.e.*, inverted for visualization only.)

ject types. The onset and offset of a specific RFID label stream may provide a useful clue for the onset/offset of a certain activity that typically manipulates the corresponding objects.

Representation of Scene Statistics

Figs. 3 and 4 show the process of representing scene statistics. We denote the m-th camera image and its foreground map at time t as I_m^t and F_m^t, respectively, ($m \in \{1, 2, 3, 4\}$, See Fig. 3). A *super image* ϑ^t and its associated *super foreground map* Γ^t are obtained by juxtaposing the individual images $I_1^t, \ldots, I_4^t$ and $F_1^t, \ldots, F_4^t$, respectively. Therefore if I_m^t is of size $W \times H$ pixels, ϑ^t and Γ^t are of size $2W \times 2H$. (In our implementation, image width $W = 640$ pixels and image height $H = 480$ pixels.)

A virtual grid overlays the *super foreground map* Γ^t (Fig. 4) for decimation as follows. Each of the grid cells with cell size of $S \times S$ pixels ($S = 40$ pixels) counts the number of foreground pixels (in Fig. 4) within its cell boundary and divides the number with the cell area as follows.

$$\delta_{i,j}^t = \frac{\sum \sharp\, foreground\ pixels}{S^2} \tag{1}$$

Therefore, $\delta_{i,j}^t$ forms the value of a *super pixel* representing the ratio of foreground occupancy in the cell, ranging [0, 1]:

$$\delta_{i,j}^t \in [0, 1] \in \mathbb{R} \tag{2}$$

In our current implementation, the original $2W \times 2H (= 1280 \times 960)$ super foreground map Γ^t is converted to $\frac{2W}{S} \times \frac{2H}{S} (= 32 \times 24)$ *occupancy map* $\mathcal{O}^t$ which contains super pixel values $\delta_{i,j}^t$, $i \in \{1, \ldots, 32\}, j \in \{1, \ldots, 24\}$. We concatenate all the super pixels $\delta_{i,j}^t$ across $\mathcal{O}^t$ into a vector format, Δ^t:

$$\Delta^t = [\delta_{1,1}^t, \ldots, \delta_{32,24}^t]^T \tag{3}$$

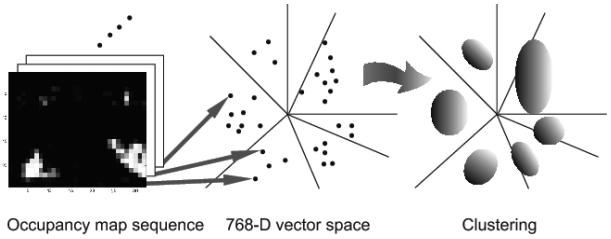

Figure 5: Transformation from the sequence of *occupancy map* $\mathcal{O}$ (color coded for visualization) to the high-dimensional vector space to the cluster labels.

where the dimensionality of Δ^t is 768 ($= 32 \times 24$) in our experiments.

Therefore, the entire foreground silhouettes from the four simultaneous camera views at time t (e.g., Fig. 4) are represented as a *single* point in the 768-dimensional vector space. Consecutive frames of a given video sequence are mapped as sample points in this vector space as depicted in Fig. 5.

Our conjecture is that two sample points, Δ^{t_1} and Δ^{t_2} from nearby frames (*i.e.*, $|t_2 - t_1| \ll \tau$) involved in the same activity will be grouped together, whereas the sample points from distant portions of the video sequence (*i.e.*, $|t_2 - t_1| \gg \tau$) involved in different activities will get separated in this vector space. The proper threshold value τ can be learned from training data.

Clouds of adjacent sample points can be grouped into clusters by standard clustering algorithms such as K-means clustering or expectation-maximization (EM) based Gaussian mixture modeling (Duda, Hart, and Stork 2001). Note that each element of vector Δ^t was normalized into the range of [0, 1]. We adopt the Linde-Buzo-Gray Vector Quantization algorithm (Sayood 1996), an efficient variant of K-means clustering, to cluster the high-dimensional vector space representation of the video data into $\mathcal{M}$ clusters. (We will show experimental results with different values of $\mathcal{M}$ from 32 to 128.)

In this approach, the long-term video sequence that involves multiple sequential activities is therefore succinctly represented as the transition of cluster labels across the $\mathcal{M}$ clusters as depicted in Fig. 5. The evolution patterns of cluster label transitions in different activities in daily living will be learned by using hidden Markov models in an unsupervised manner with expectation-maximization (EM) learning.

RFID for Training Activity Segments

Proper parsing of the temporal sequence of feature streams for activity recognition is still an open research question. Traditional approaches for generating ground-truth annotation of training data are based on cumbersome manual segmentation of feature stream by user discretion. Such approaches are not effective or error-prone for natural activities that may vary in duration.

We are using RFID sensing for segmenting and labeling ADLs training data. Intel Research Seattle developed and supplied our lab with an RFID reader in the form of bracelet.

Table 1: Activity class descriptions.

1. Walk around (WA)	2. Sit and watch TV (ST)
Enter the scene Walk	Bring remote control Sit on couch Turn on / watch TV
3. Prepare utensil (PU)	**6. Store utensil (SU)**
Open / close cupboard Bring utensil (dish, cup, bowl) Bring flatware (spoon, knife, and fork) Open / close drawer	Open / close cupboard Return utensil Return flatware Open / close drawer
4. Prepare cereal (PC)	**5. Drink water (DW)**
Open cupboard Bring a cereal box Pour cereal in the bowl Pour milk in the bowl Eat cereal with spoon	Open refrigerator Bring water jar Pour water in a cup Drink water in the cup

It has a detection range of about 10–15 centimeters. As the person's hand approaches an RFID tagged object, the bracelet detects the tag and wirelessly transmits the time-stamped ID information to the PC-based activity recognition system. In our current configuration, the ID transmission is repeated every second until the person's hand leaves the object.

The combination of vision and RFID was pioneered by Wu *et al.* (Wu et al. 2007) to train object appearance models from high-quality color image frames. The RFID labels were used only to infer object use. A single detailed-view camera was used in their system, and no tracking of objects or human body was considered. Using color camera, however, may raise the issue of privacy in practical situations. We choose to use multi-view image silhouettes to represent human body and objects, and build a bank of HMM classifiers that model the interaction of the person and objects. RFID sensing in our system is used for learning the temporal segmentation of salient motions in video.

Activity Recognition Modeling

Activities in daily living occur in certain contexts. Such contexts may include short-range history of preceding activities, as well as a global and long-range information such as an individual's health conditions, the time of day, the time since the last occurrence of a regularly repeated activity (*e.g.*, toileting), *etc.* Activities may be observed at a coarse level, such as moving across multiple rooms during a day, as well as at a fine level, such as detailed cooking behavior in a kitchen. Our goal is to encompass both levels of analysis by developing an integrated hierarchical activity model. More specifically, our initial experiments include the six coarse-level activity classes described in Table 1.

Note that each of the six coarse-level activities is composed of a series of fine-level *unit actions*. *Wide-range* activity classes 1 and 2 occur in the wide living-room space of our smart home testbed, and are monitored by the two wide FOV cameras covering the entire space, while *narrow-range* activities 3 through 6 occur in the kitchen room space and are monitored by the two narrow FOV cameras monitoring the kitchen area.

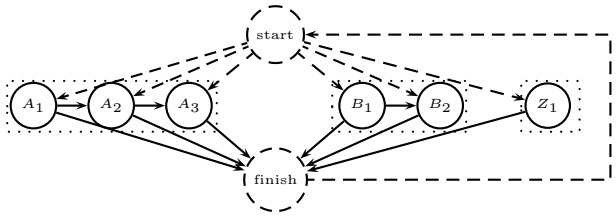

Figure 7: A simplified structure of the bank of HMMs. The dotted arrows denote null transitions.

Activity Model

Some of the activity classes (*i.e.*, wide-range activities 1 and 2) in Table 1 have very different characteristics from other activity classes (*i.e.*, narrow-range activities 3 through 6) in terms of interaction styles with different objects in the environments and available evidence from salient imaging features. Different environments such as living room *vs.* kitchen involve different objects such as couch/TV *vs.* utensil/food, respectively. In such structured environments, human behavior is constrained by the existing setup of objects. This constraint makes it easy to capture reliable visual evidence from stationary cameras. Furthermore, multi-view cameras can capture rich evidence from different perspectives.

We use the foreground map obtained by background subtraction as the *scene statistics* feature, *S*, that describe scene changes from the learned background scene. We use the motion map obtained by frame differencing as the *object statistics* feature, *J*, that represents the local motion of objects in the foreground map. The RFID signals are used as the *reference evidence, R,* of object touch, which is regarded as the baseline performance in our experiments. The *S, J* and *R* represent different sensing modalities about the scene.

We are developing graphical models for recognizing ADLs by incorporating a bank of multiple HMMs (Fig. 7). Observation nodes are omitted for clarity. The individual HMM models in the dotted boxes are denoted by different letters. The dotted nodes, *start* and *finish*, are virtual nodes that do not involve any observation nodes. The dashed edges do not take time step, meaning that the multiple individual HMMs are operating in parallel.

We use standard Baum-Welch and Viterbi algorithms (Rabiner 1989) to train the individual HMMs. For each HMM model, the number of hidden node states (e.g., $|A_i|$) we tested are 1 and 3 ($\mathcal{N}_{\mathcal{H}}$ = 1 or 3), and the number of observation node states ($|S_t|$) and ($|D_t|$) varied between 32, 64 and 128 ($\mathcal{M}_{\mathcal{H}}$ = 32, 64, or 128) depending on the number of cameras used as input for single- *vs.* multi-view based recognition performances (See Experiments section). The number of RFID observation node states ($|R_t|$) was fixed to 30 due to the fixed number of RFID tags used in the experiments.

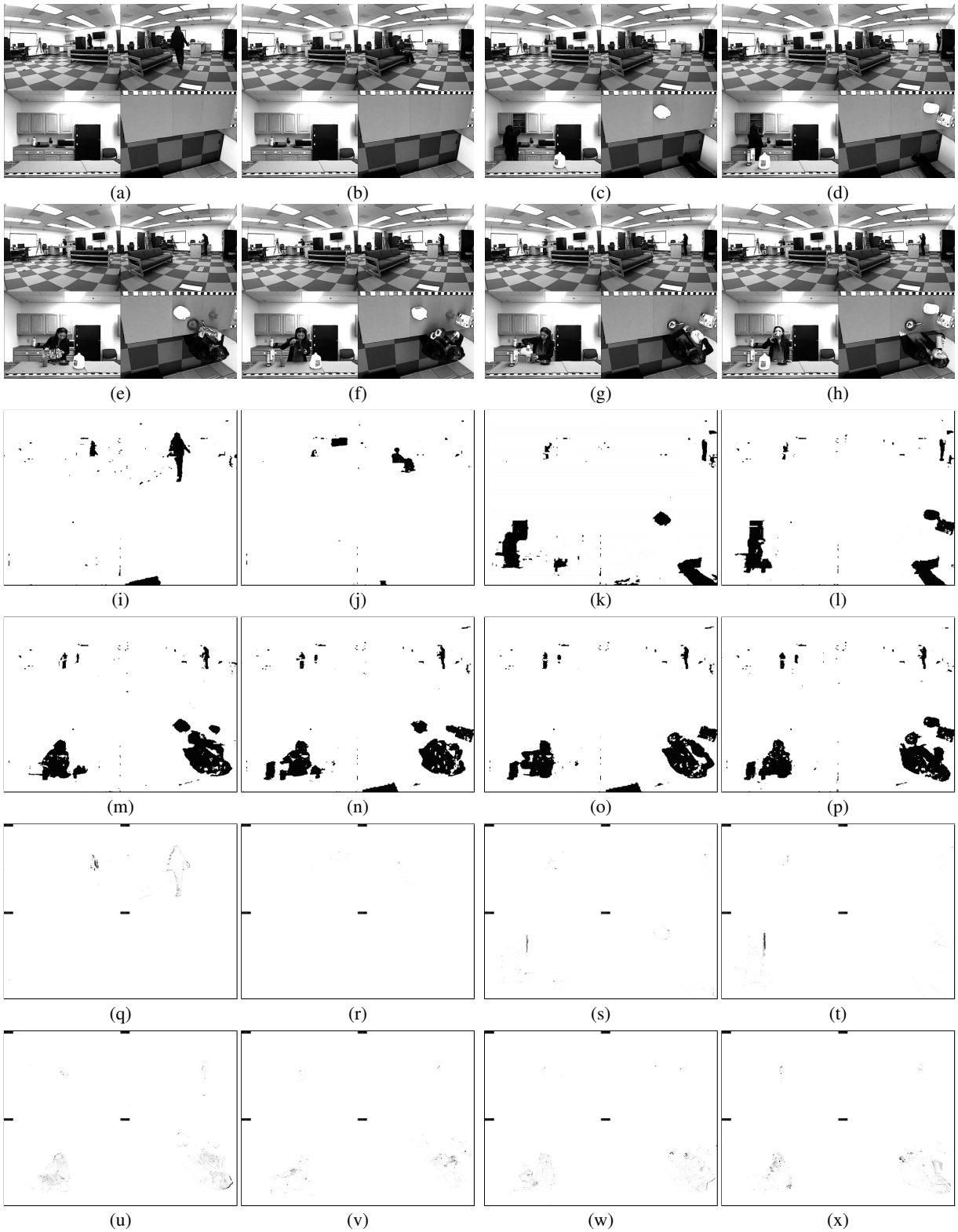

Figure 6: Example frames for each activity in Table 1. (a) Walk around, (b) sit and watch TV, (c) prepare utensil, (d) store utensil, (e)(f) prepare cereal, and (g)(h) drink water, respectively. Notice that some activities may look very similar at certain moments during the sequences. Raw frames: (a)-(h), background subtraction: (i)-(p), frame-differenced motion maps: (q)-(x).

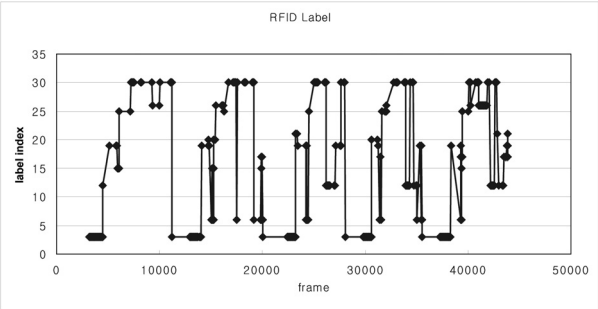

Figure 8: A participant's five epoches of RFID labels. (x-axis: associated video frame index, y-axis: RFID label indices.)

Experiments

We are currently investigating the six activity classes occurring in the smart home testbed as shown in Fig. 2. K persons ($K = 5$) participated twice in the experiments in separate sessions to conduct the activities from 1 through 6 in a sequential manner, which defines an *epoch.* In total, 5 epochs (*i.e.*, repetitions) per each of the 6 activity classes per each of the 10 trials (*i.e.*, 5 persons $\times$ 2 sessions) were collected. Participants were free to choose different sub-sequences of the fine-level actions in each of the 6 coarse-level activity classes. The persons wore the RFID reader on their right wrists, which detected the nearby objects' RFID labels in a sequential manner as shown in Fig. 8. The five chunks of RFID patterns corresponding to the epochs were used to segment video sequences.

Table 2, Tables 3- 4, and Table 5 show the confusion matrices of activity recognition using *RFID*, *scene statistics*, and *object statistics*, respectively. The rows and columns of the tables denote the true and the recognized classes, respectively. The *leave-one-out* cross validation scheme uses four epochs for training and the remaining one epoch for testing in a round-robin manner. In each cell, the numbers with larger font denote the average accuracy, and the superscript numbers with smaller font denote the standard deviation of the five average accuracy measures. The lower right cell of each sub-table shows the average of the recall rates in the corresponding sub-table.

Comparing Tables 2 and 4 shows that RFID-based recognition accuracy is higher for kitchen activities (activity classes 3 - 6 in Table 1), while vision-based recognition accuracy is higher for living room activities (activity classes 1 and 2.) The cells of low accuracy with large standard deviation (shown in light gray) are complementary between the two sensing modalities as follows: living-room activities (e.g., *Walk around*) are better recognized with evidence from scene statistics, while cooking activities (e.g., *Prepare cereal* and *Drink water*) are better recognized with the evidence from RFID sequences. This result is consistent with the environmental configuration of the testbed; the kitchen area contains many occluding objects with RFID tag and the living area is an open space with fewer RFID tags, which are sparsely distributed.

Comparison of Table 3 and the first Sub-table in Table 4 shows the advantage of using multiple cameras for input observation over a single camera. The multi-view approach enhances the overall accuracy from 74 % to 80%, and significantly reduces the variance of the recognition accuracy. The robustness is enhanced as indicated by the reduced number of nonzero off-diagonal cells.

Comparing the two Sub-tables in Table 4 shows that the more complex model increases the recognition accuracy from 80% to 83%. The small difference may imply that the two models have already converged. Further improving the recognition accuracy may require the fusion of multiple observation streams as input, which is part of our ongoing investigation.

The most difficult class of activity in the recognition task is *Drink water* (DW), as indicated by the low recall rates in Tables 3 and 4. This activity often gets confused with *Prepare cereal* (PC), because the two activities are not well-distinguishable by their silhouettes.

Table 5 shows the significant increase of the recognition accuracy of the *Drink water* (DW) class from 46 % to 72 % by using the *frame-differenced* motion maps (e.g., in Fig. 6 (q)-(x)) as input observation. Unlike the background subtraction, the motion maps obtained by frame differencing effectively captures the moving parts within the silhouette.

In summary, the experimental evaluation shows the following findings:

- Multi-view vision enhances recognition accuracy and robustness of recognition of ADLs.
- Different input modalities capture different aspect of activity.
- Complex model improves the recognition accuracy slightly.
- Data fusion by an integrated model is desirable.

The physical world including the human body is basically a 3-dimensional space. The sensitivity analysis shown in Tables 2 - 5 in terms of the standard deviations of recognition accuracy supports the following argument: the quasi-orthogonal configuration of the multiple cameras better represents the scene structure than a single camera does, by providing redundancy and removing ambiguity due to occlusion between objects. With this redundancy, our multi-view vision system achieves reasonable performance rates (compared to the RFID-based system) with very coarse visual representation in terms of silhouette mosaic or motion blur as input evidence.

Conclusion

We have presented our ongoing research on privacy-preserving recognition of activities in daily living. Our approach uses a distributed multi-view vision system and RFID reader/tags for view independence and robustness in obtaining evidence for scene statistics, object statistics, and RFID labels. RFID-based evidence better indicates cooking activities involving multiple utensils, while multi-view vision-based evidence better indicates living-room activities

Table 2: Recognition of ADLs using RFID labels. ($\mathcal{N}_{\mathcal{H}} = 1$, $\mathcal{M}_{\mathcal{H}} = 30$) (Legends in Table 1). The numbers of larger font in each cell denote the average accuracy, and the superscript numbers denote the standard deviation (*i.e.*, not exponent.)

RFID	WA	ST	PU	PC	DW	SU	recal
WA	$0.70^{0.22}$	$0.30^{0.22}$					0.70
ST		$0.86^{0.09}$	$0.14^{0.09}$				0.86
PU			$0.88^{0.08}$	$0.02^{0.04}$		$0.10^{0.07}$	0.88
PC	$0.02^{0.04}$			$0.88^{0.11}$	$0.04^{0.05}$	$0.06^{0.05}$	0.88
DW	$0.02^{0.04}$			$0.12^{0.16}$	$0.78^{0.16}$	$0.08^{0.11}$	0.78
SU				$0.02^{0.04}$		$0.98^{0.04}$	0.98
precision	0.95	0.74	0.86	0.85	0.95	0.80	0.85

Table 3: Single-view based recognition of ADLs. Occupancy maps $\mathcal{O}$ from the single-view foreground silhouettes (e.g., Fig. 6 (i)-(p)) were used as input observation. ($\mathcal{N}_{\mathcal{H}} = 1$, $\mathcal{M}_{\mathcal{H}} = 32$).

Cam-1	**WA**	**ST**	PU	PC	DW	SU	recall
WA	$1.00^{0.00}$						1.00
ST	$0.06^{0.05}$	$0.94^{0.05}$					0.94
PU			$0.76^{0.05}$			$0.24^{0.05}$	0.76
PC			$0.24^{0.15}$	$0.56^{0.22}$	$0.16^{0.13}$	$0.04^{0.05}$	0.56
DW			$0.12^{0.16}$	$0.36^{0.19}$	$0.52^{0.18}$		0.52
SU			$0.24^{0.11}$		$0.04^{0.05}$	$0.72^{0.11}$	0.72
precision	0.94	1.00	0.56	0.61	0.72	0.72	0.75
Cam-2	**WA**	**ST**	PU	PC	DW	SU	recall
WA	$1.00^{0.00}$						1.00
ST	$0.04^{0.05}$	$0.96^{0.05}$					0.96
PU			$0.60^{0.07}$		$0.02^{0.04}$	$0.38^{0.08}$	0.60
PC				$0.84^{0.15}$	$0.12^{0.13}$	$0.04^{0.05}$	0.84
DW				$0.58^{0.22}$	$0.42^{0.22}$		0.42
SU			$0.22^{0.04}$		$0.02^{0.04}$	$0.76^{0.09}$	0.76
precision	0.96	1.00	0.73	0.59	0.72	0.64	0.76
Cam-3	WA	ST	**PU**	**PC**	**DW**	**SU**	recall
WA	$0.64^{0.13}$	$0.34^{0.17}$	$0.02^{0.04}$				0.64
ST	$0.06^{0.05}$	$0.94^{0.05}$					0.94
PU			$0.88^{0.08}$			$0.12^{0.08}$	0.88
PC			$0.02^{0.04}$	$0.78^{0.18}$	$0.18^{0.19}$	$0.02^{0.04}$	0.78
DW			$0.02^{0.04}$	$0.48^{0.13}$	$0.50^{0.12}$		0.50
SU			$0.04^{0.09}$			$0.96^{0.09}$	0.96
precision	0.91	0.73	0.90	0.62	0.74	0.87	0.78
Cam-4	WA	ST	**PU**	**PC**	**DW**	**SU**	recall
WA	$0.54^{0.34}$	$0.44^{0.36}$	$0.02^{0.04}$				0.54
ST	$0.32^{0.41}$	$0.68^{0.41}$					0.68
PU			$0.90^{0.10}$			$0.10^{0.10}$	0.90
PC			$0.02^{0.04}$	$0.66^{0.13}$	$0.28^{0.08}$	$0.04^{0.05}$	0.66
DW			$0.02^{0.04}$	$0.46^{0.11}$	$0.52^{0.08}$		0.52
SU			$0.12^{0.08}$			$0.88^{0.08}$	0.88
precision	0.63	0.61	0.83	0.59	0.65	0.86	0.70

in large space. Overall, the RFID-based recognition accuracy is comparable to the multi-view based recognition accuracy that uses coarse silhouette mosaic or even coarser representation by frame-differenced images. The complementary nature of the two sensing modalities provides useful integration for better recognition of activities in daily living. We have presented the basic performance evaluations for each of the sensing methods. The multi-view vision system achieves reliable performance in recognition accuracy from privacy-preserving simple profiles such as silhouette mosaic and frame differenced motion maps. To be truly privacy preserving, the process of obtaining the silhouette would need to be performed at the camera module such as thermal or near infrared cameras. Our approach is directly applicable to those cameras without significant modification. The current system assumes that the situation involves a single person, which would be the most critical situation in practice. But this bar of constraint would sometimes need to be raised. We

Table 4: Multi-view based recognition of ADLs. Occupancy maps $\mathcal{O}$ from foreground silhouettes (e.g., Fig. 6 (i)-(p)) were used as input observation. Different HMM complexities were compared between *1-state* HMMs: ($\mathcal{N}_\mathcal{H}$ = 1, $\mathcal{M}_\mathcal{H}$ = 32) *vs. 3-state* HMMs : ($\mathcal{N}_\mathcal{H}$ = 3, $\mathcal{M}_\mathcal{H}$ = 128).

1-state HMMs	**WA**	**ST**	**PU**	**PC**	**DW**	**SU**	recall
WA	$0.84^{0.13}$	$0.14^{0.15}$	$0.02^{0.04}$				0.84
ST	$0.04^{0.05}$	$0.96^{0.05}$					0.96
PU			$0.94^{0.09}$			$0.06^{0.09}$	0.94
PC			$0.02^{0.04}$	$0.68^{0.04}$	$0.30^{0.00}$		0.68
DW				$0.50^{0.10}$	$0.50^{0.10}$		0.50
SU			$0.10^{0.07}$			$0.90^{0.07}$	0.90
precision	0.95	0.87	0.87	0.58	0.63	0.94	0.80

3-state HMMs	**WA**	**ST**	**PU**	**PC**	**DW**	**SU**	recall
WA	$1.00^{0.00}$						1.00
ST	$0.02^{0.04}$	$0.98^{0.04}$					0.98
PU			$0.94^{0.05}$			$0.06^{0.05}$	0.94
PC			$0.08^{0.08}$	$0.68^{0.13}$	$0.08^{0.08}$	$0.16^{0.15}$	0.68
DW			$0.02^{0.04}$	$0.42^{0.11}$	$0.42^{0.04}$	$0.14^{0.11}$	0.42
SU			$0.06^{0.09}$			$0.94^{0.09}$	0.94
precision	0.98	1.00	0.85	0.62	0.84	0.72	0.83

Table 5: Recognition of ADLs using occupancy maps from *frame-differenced* motion maps (e.g., Fig. 6 (q)-(x)) as the input observation. ($\mathcal{N}_\mathcal{H}$ = 1, $\mathcal{M}_\mathcal{H}$ = 32)

	WA	ST	PU	PC	DW	SU	recall
WA	$0.92^{0.04}$	$0.06^{0.05}$	$0.02^{0.04}$				0.92
ST	$0.08^{0.08}$	$0.92^{0.08}$					0.92
PU			$0.94^{0.05}$			$0.06^{0.05}$	0.94
PC				$0.66^{0.11}$	$0.32^{0.08}$	$0.02^{0.04}$	0.66
DW				$0.28^{0.11}$	$0.72^{0.11}$		0.72
SU			$0.06^{0.05}$			$0.94^{0.05}$	0.94
precision	0.92	0.94	0.92	0.70	0.69	0.92	0.85

are currently developing more robust algorithms for multi-object tracking to obtain better object statistics, and intend to investigate more efficient graphical models.

Acknowledgement

This work is funded by DARPA SBIR Contract W31P4Q-08-C-0170, NSF award ISS-0734843, NYSTAR award C020010, and DARPA SBIR award 07SB2-0196.

References

Abowd, G. D.; Rehg, J.; Starner, T. E.; and Arriaga, R. 2007. Using information technologies to support detection of developmental delay. In *Assisted Cognition Workshop*.

Aggarwal, J., and Park, S. 2004. Human motion: modeling and recognition of actions and interactions. *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on* 640–647.

Duda, R.; Hart, P.; and Stork, E. 2001. *Pattern Classisifcation*. New York: Wiley, 2nd edition. chapter 10.

Kim, K.; Chalidabhongse, T.; Harwood, D.; and Davis, L. 2005. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging* 11.

Mihailidis, A.; Hoey, J.; Boutilier, C.; and Boger, J. 2007. Towards the development of an intelligent home to support aging-in-place. In *Assisted Cognition Workshop*.

Oliver, N.; Horvitz, E.; and Garg, A. 2002. Layered representations for human activity recognition. In *Proc. IEEE Int'l Conference on Multimodal Interfaces*, 3–8.

Patterson, D.; Fox, D.; Kautz, H.; and Philipose, M. 2005. Fine-grained activity recognition by aggregating abstract object usage. In *Proc. IEEE International Symposium on Wearable Computers*.

Rabiner, L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257 –286.

Sayood, K. 1996. *Introduction to Data Compression*. Morgan Kaufmann Publishers.

Wu, J.; Osuntogun, A.; Choudhury, T.; Philipose, M.; and Rehg, J. M. 2007. A scalable approach to activity recognition based on object use. In *Proceedings of Int'l conference on computer vision*.