Hierarchical Recognition of Activities of Daily Living using Multi-Scale, Multi-Perspective Vision and RFID

Sangho Park and Henry Kautz

Computer Science, University of Rochester
Rochester, NY 14627, USA
{spark | kautz}@cs.rochester.edu

Keywords: activity recognition, computer vision, RFID

Abstract

Research on computer-based recognition of ordinary household activities of daily living (ADLs) has been spurred by the need for technology to support care of the elderly in the home environment. We address the issue of recognizing ADLs at multiple levels of detail by combining multi-view computer vision and radio-frequency identification (RFID)-based direct sensors. Multiple places in our smart home testbed are covered by distributed synchronized cameras with different imaging resolutions. Learning object appearance models without costly manual labeling is achieved by applying the RFID sensing. A hierarchical recognition scheme is proposed by building a dynamic Bayesian network (DBN) that encompasses both coarse-level and fine-level ADL recognition. 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 at both a coarse and fine level of detail.

1 Introduction and Research Motivation

Computer-based recognition of human activities in daily living (ADLs) has gained increasing interest from computer science and medical researchers as the portion of the elder population in society grows. 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 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 RFID reader, Intel Research Seattle's "iBracelet", remotely transmits time-stamped RFID signals to the vision system's computer. Multiple RFID tags are attached to various objects including furniture, appliances, and utensils around the smart homes. Although we currently use commercial-quality cameras and a high-end frame buffer to integrate the video feeds, the decreasing cost of video cameras and the increasing power of multicore personal computers will make it feasible in about two years to deploy our proposed system with inexpensive wireless camcorders and an ordinary laptop computer.

Previous approaches to recognizing ADLs have depended upon users wearing sensors (RFID and/or accelerometers) or using a single camera vision system. Recently, [4] 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 performance. The system also did not attempt to model or learn the *motion* information involved in the ADL. We propose a multi-sensor based activity recognition system that uses multiple cameras and RFID readers in a richer way.

Understanding human activity can be approached from different levels of detail: for example, a body transition across a room at a coarse level, versus the hand motions manipulating objects at a detailed level. Our multi-camera based vision system covers various indoor areas with different viewing resolutions from different perspectives. RFID tags and reader(s) pinpoint the nearby objects which are handled by the user. 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 object appearance models without human intervention, and the recognition of human activity at both a coarse and fine level.

2 System Architecture Overview

Fig. 1 shows the overall system architecture. Light gray modules compose the basic single-view system, while the bright (yellow) modules compose the multi-view functionality. Dark gray modules can work either in single or multi-view modes, but more cameras can increase the overall accuracy. In the single-view mode, a planar homography mapping generates a virtual top-down view of the ground plane overlayed with a warped foreground image. In multi-view mode, the foreground image is created by warping and blending all the camera views. Using multiple views not only increases robustness, but also supports simple and accurate estimation of view-invariant features such as object size.

Currently, four cameras are used for synchronized views,

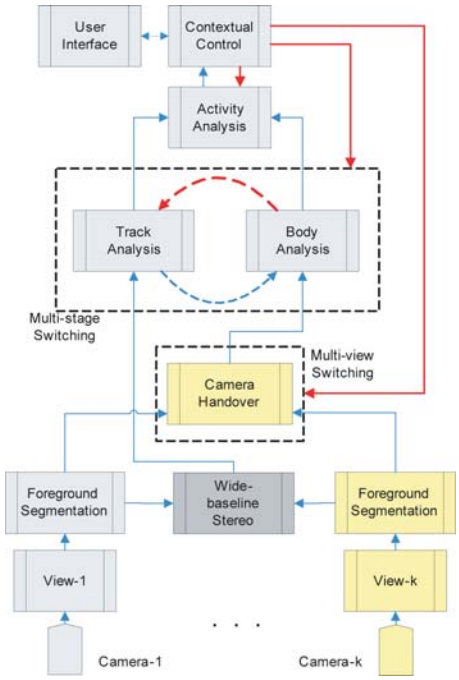

Figure 1: The overall system architecture of the Multi-scale multi-perspective vision system.

Figure 2: An example session of the ADL experiment in which a person performs the "Drink water" activity.

as shown in Fig. 2 , which are foreground-segmented and combined to form a planar-homography map for 3D localization of persons. The homography map is to generate a perspective-independent virtual top-down view for the (coarse) track-level analysis, whereas the view switching for unoccluded views of people is used for the (fine) body-level analysis. Fig. 3 shows an example of the homography mapping from the two wide-FOV cameras and an example of the foreground segmentation from the two narrow-FOV cameras. The narrow 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 on analysis task.

In Fig. 1, dynamic contextual control with optional user involvement is 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 in Fig. 1.

2.1 Appearance-Based Segmentation and Tracking

ADLs may involve multiple objects moving simultaneously, which can create challenges for a vision system — for example, changing background and object occlusion.

We adopt a dynamic background model using K-means clustering [1]. Background model is updated with a certain memory decay factor to incorporate the changes in the background, and foreground-background segmentation is achieved at each pixel.

The smart home environment may include multiple persons, each of whom may disappear and reappear across non-overlapped camera views. It is important to robustly locate individual persons and re-identify each person across views. We employ a probabilistic appearance model (PAM) that represents people's color appearance in terms of Gaussian mixture models [2]. The parameters of the mixture model are learned using expectation-maximization (EM) when an individual first appears in a video frame. The tracking system can re-identify people who leave and later reappear using a color-histogram comparison.

2.2 Feature-based Body Parts Detection

Using the segmented multiple foreground regions of the input video frames, we detect human body parts such as face, shoulder, and hands. Haar-wavelet based human feature detectors [3] are trained for specific textural patterns associated with frontal face, profile face, and upper body silhouette. Skin tone is effective to segment non-textural or highly deformable body parts such as hands, but skin tone depends on illumination conditions. In order to robustly detect proper skin tone, we use the detected face area as a bootstrapped prior. That is, the trained human feature detectors are jointly used to bootstrap the proper region of face and for skin color sampling from the face.

2.3 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 provide person locations 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).

3 RFID for Learning Temporal Segmentation of Salient Motions

Properly parsing the temporal sequence of feature streams for activity recognition is still an open research question. Traditional approaches are based on manual segmentation or on moving window of fixed duration. Such approaches are not very effective for natural activities that may vary in duration.

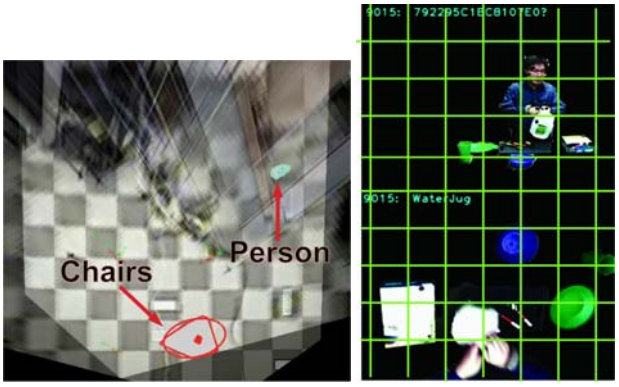

Figure 3: Homography projection of the two wide-FOV scenes in Fig. 2, and foreground segmentation of narrow-FOV with a RFID label from a different sequence (grid overlaid for scene statistics).

We are using RFID sensing for segmenting and labeling ADL training data. Intel Research Seattle developed and supplied our lab with an RFID reader in the form of bracelet. It has detection range of about 10–15 centimeters. As the person's hand approaches to an RFID tagged object, the iBracelet detects the tag and transmits the time-stamped ID information by a wireless link 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.* [4] to train object appearance models without laborious manual labeling efforts. 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. Our work expands upon their approach by incorporating human body model and object models, and building a DBN that models the interaction of the person and objects. RFID sensing in our system serves for learning temporal segmentation of salient motions as well as object appearance learning.

4 Activity Recognition Modeling

Activities in daily living occur in certain contexts. Such contexts may include a 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 instance of a regularly repeated activity occurred (*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 2.

Note that each of the six coarse-level activities is composed of a series of fine-level *unit actions*. Activity classes 1 and 2 are monitored by the two wide FOV cameras covering the entire space, while activities 3 through 6 are monitored by the two narrow FOV cameras monitoring the kitchen area.

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

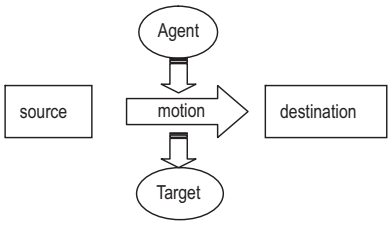

Figure 4: Unit Action Model.

4.1 The Coarse Activity Model

Some of the activity classes (*i.e.*, coarse-level activities 1 and 2) in Table 2 have very different characteristics from other activity classes (activities 3 through 6) in term of available salient features, mainly due to the different perspectives provided by the distributed cameras with different FOVs. The camera handover between the different cameras in Fig. 1 is achieved by incorporating a data-driven bottom-up process and a knowledge-driven top-down process. The bottom-up process is achieved by low-level visual procedures such as background subtraction and foreground objects tracking, while the top-down process is achieved by activity recognition using the graphical model. More specifically, *multi-view switching* and *multi-stage switching* as shown in Fig. 1 are controlled by top-down feedback.

4.2 The Unit Action Model

We model human activity as a composition of intentional *unit actions*, each of which is represented by a tuple composed of {agent, motion, target, source, destination} as shown in Fig. 4. The unit action model asserts that a meaningful atomic chuck of action may be identified by delineating the interaction between an agent and a target object associated with a certain source and destination. For example, a hand (an *agent*) may carry (a *motion*) a bowl (a *target*) from the cupboard (a *source*) to the table (a *destination*) in

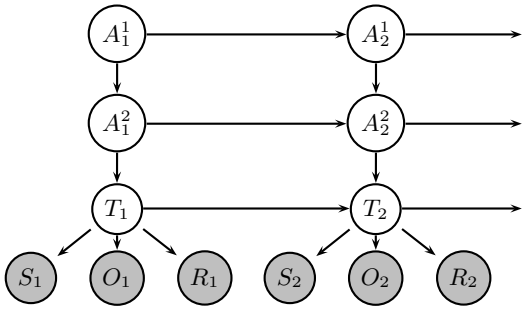

Figure 5: Hierarchical composition of dynamic Bayesian networks for recognizing ADLs. (Subscript denotes time.)

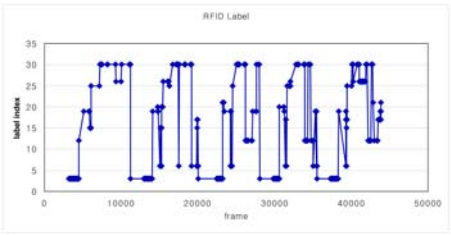

Figure 6: A participant's five epoches of RFID labels

order to "get a bowl". Moving a bowl from cupboard to table changes the states of the cupboard and the table. Such state changes can be detected by the scene statistics such as the ratio of detected foreground regions at specific locations to the entire field of view. Each of the oval and square nodes in Fig. 4 is parameter that needs to be filled with an inferred value. The parameters in the various unit actions that make up a coarse activity may be tied together. For example, "get a bowl" and "spoon cereal from bowl to mouth" may both be a unit actions in the "make breakfast" activity, where the object parameter of the first unit action is tied to the source parameter for the second unit action.

We are developing graphical models for recognizing ADL by incorporating such interdependency of the multiple nodes in Fig. 4. We incorporate the observations from scene statistics (S_i), object statistics (O_i), and RFID labels (R_i) to build a hierarchical classifier for ADL as shown in Fig. 5. The coarse level activity (A_j^1) is composed of a sequence of detailed unit activities (A_j^2) that handle objects (T_i).

5 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*. E epochs (*i.e.*, repetitions) ($E = 5$ for now) total per activity class per participant in each session are collected. Participants are free to choose different sequences of the fine-level actions in each of the 6 coarse-level activity classes. Fig. 2 shows an example session of the ADL experiment in which a person performs a kitchen activity (*i.e.*, "Drink water".) The person wears the RFID reader on his right wrist, which detects the nearby objects' RFID labels in a sequential manner as shown in Fig. 6.

Table 2: Activity recognition using only RFID (*RFID*) or scene statistics (*SS*), respectively. (Superscript denotes standard deviation from *leave-one-out* cross-validation.)

RFID	WA	ST	PU	PC	DW	SU
WA	$.70^{.22}$	$.30^{.22}$				
ST		$.86^{.09}$	$.14^{.09}$			
PU			$.88^{.08}$	$.02^{.04}$		$.10^{.07}$
PC	$.02^{.04}$			$.88^{.11}$	$.04^{.05}$	$.06^{.05}$
DW	$.02^{.04}$			$.12^{.16}$	$.78^{.16}$	$.08^{.11}$
SU				$.02^{.04}$		$.90^{.04}$
mean						.83

SS	WA	ST	PU	PC	DW	SU
WA	$.90^{.07}$		$.08^{.04}$			$.02^{.04}$
ST	$.06^{.09}$	$.78^{.13}$	$.16^{.13}$			
PU	$.02^{.04}$		$.90^{.07}$		$.08^{.07}$	
PC			$.06^{.05}$	$.52^{.08}$	$.12^{.08}$	$.30^{.07}$
DW			$.02^{.04}$	$.38^{.24}$	$.48^{.25}$	$.12^{.08}$
SU			$.24^{.18}$			$.76^{.18}$
mean						.72

Table 2 shows the confusion matrix of activity recognition using RFID and scene statistics sequences, respectively. The cells of low accuracy with large standard deviation are complementary between the two confusion matrices as follows; certain activities (e.g., *walk around*) are better recognized with evidence from scene statistics, while other activities (e.g., *prepare cereal* and *drink water*) are better recognized with the evidence from RFID sequences.

6 Conclusion

We have presented our ongoing research on hierarchical 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 evidences which include scene statistics, object statistics, and RFID labels. We showed that different types of evidences better indicate different activities such as *walking around* vs. *preparing cereal*. We are currently developing more robust algorithms for multi-object tracking to obtain better object statistics, and intend to investigate diverse and more efficient graphical models.

References

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

[2] Sangho Park and Mohan M. Trivedi. Understanding human interactions with track and body synergies (TBS) captured from multiple views. *Computer Vision and Image Understanding*, 2008.

[3] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. 2001.

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