Environment Recognition Based on Human Actions Using Probability Networks

Hiroshi Miki Atsuhiro Kojima
Koichi Kise
Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
miki@m.cs.osakafu-u.ac.jp, ark@las.osakafu-u.ac.jp, kise@cs.osakafu-u.ac.jp

Abstract

To realize context aware applications for smart home environments, it is necessary to recognize function or usage of objects as well as categories of them. On conventional research for environment recognition in an indoor environment, most of previous methods are based on shape models. In this paper, we propose a method for recognizing objects focused on the relationship between human actions and functions of objects. Such relationship becomes obvious on human action patterns when he or she handles an object. To estimate object categories by using action patterns, we represent such relationship in Dynamic Bayesian Networks (DBNs). By learning human actions toward objects statistically, objects can be recognized. Finally, we performed experiments and confirmed that objects can be recognized from human actions without shape models.

1. Introduction

In general, to realize context aware application for smart home environments, human activities and arrangements of objects should be recognized. Although RFIDs or bar codes are effective to identify objects simply, some kinds of information can not be obtained from these sensors or labels, such as human actions related to objects.

In the field of pattern recognition, most of researches for object recognition are based on appearance models such as shapes, textures or colors of objects. On the other hand, different approaches for object recognition focused on interactions between human activities and objects are proposed. Object categories can be recognized using object functions or attributes estimated from human actions. Moore *et al*. proposed a model-based technique for recognizing objects using trajectories of hand motions [1]. In a similar work, a method for labeling and segmentation of object regions using human actions [2]. These processes are, however, performed in a 2-D image, not in a 3-D space.

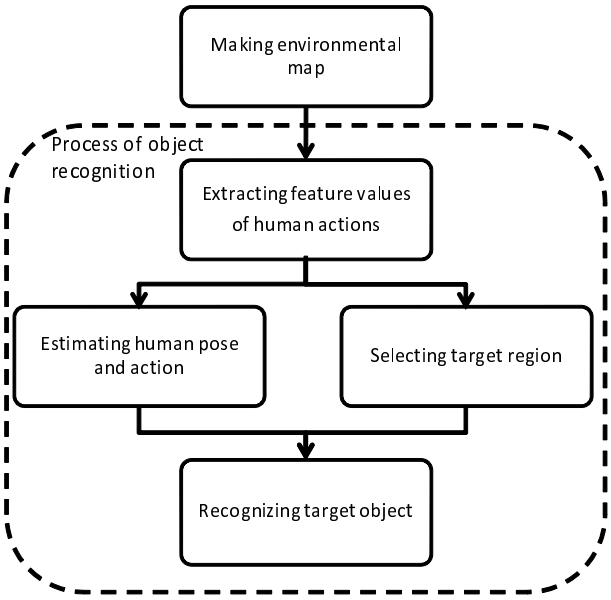


Figure 1. Process of object recognition.

Higuchi *et al*. proposed an integrated method for recognizing human actions as well as functions and usage of objects using concept hierarchies of actions [3]. Mitani *et al*. applied this method to an autonomous mobile robot [4]. In their work, objects are recognized from human motions on an environmental map via observation from multiple view points. In these methods, however, it is difficult to construct hierarchical models suitable for target objects. In another approach, a probabilistic method for describing relationships between human actions and objects is proposed in [5]. Instead of creating individual models manually, relationships between actions and objects can be learnt automatically in a probabilistic method.

In this paper, we describe a method for recognizing objects using probability networks by learning relationships between human actions and objects. Probability networks consist of Dynamic Bayesian Networks (DBNs) which are

suitable for analysing of sequential data[6].

In the following, overview of our method is presented in section 2. The process of extracting features of human action is described in section 3, and recognizing object using DBNs is described in section 4. Experimental results and discussions are presented in section 5. Finally we conclude in section 6.

2. Overview

In this method, relationships between actions and objects are represented in probability networks. These networks consist of three types of Dynamic Bayesian Networks (DBNs): *pose* DBN for estimating human poses, *basic action* DBN for estimating actions interacting with objects and *related object* DBNs for identifying objects. Each object region have an own *related object* DBN and is recognized separately.

Figure 1 shows an overview of the process to identify objects from features of human action. In advance, each object region to be identified is manually enclosed in a rectangle on a 2-D plane to which a 3-D environmental map is projected [4]. The process of object recognition is explained as follows.

First, features of human actions are extracted from positions of face and hand. For each frame, posterior probabilities of human *pose* and *basic action* are computed from features of human action. In parallel with this, positions of human face and hands are projected onto environmental map. One of the object regions are selected as related object which is overlapped with face or hands positions. Finally, posterior probabilities of the *related object* are computed from probabilities of *pose* and *basic action* mentioned above. By iterating these steps, object regions are identified cumulatively.

3. Extracting Features of Human Actions

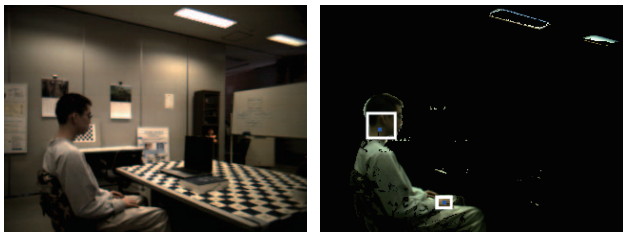
Features of human actions are extracted by tracking human face and hands positions. One reason for this is that tracking a face position can be clues to detecting human pose, such as standing, sitting, and so on. Similarly, human actions related with objects can be detected by tracking hand positions.

3.1. Tracking Face and Hands

Indoor human activities are captured from a stationary stereo camera. Simple background subtraction is applied to extraction of a human region from video sequences. From the subtracted images, skin regions are extracted and tracked using skin probabilities [7]. Skin probabilities are calculated from probabilistic distribution model of skin color using a^*, b^* of $L^*a^*b^*$ color system. Figure 2 shows an example of extracting and tracking skin regions.

Table 1. Features of Human Action.

Features	Label	Number of states
height of face	H_f	3
distance from face to hand	D_{fh}	3
direction of hand	A_{fh}	5
change of distance from face to hand	R_{fh}	3
speed of face movement	S_f	3
speed of hand movement	S_h	3
direction of face movement	M_f	6
direction of hand movement	M_h	6



(a) An input image. (b) Face and hand regions.

Figure 2. Example of extraction and tracking of skin regions.

Extracted skin regions are labeled as hand or face, assuming that only one person is present and the number of regions must be three at most. The region at highest position is labeled as face, and the others are labeled as hand.

3.2. Features of Human Action

Feature values of human actions are then evaluated from face and hands positions. Table 1 shows features used in this method, which characterize most of human motions in indoor scenes. These features are divided into two groups, H_f, D_{fh}, and A_{fh} calculated from a single frame, and D_{fh}, R_{fh}, S_f, S_h, M_f and M_h from series of T frames.

Let $F(t) = (f_x(t), f_y(t), f_z(t))$ be a coordinate of face position, and $H(t) = (h_x(t), h_y(t), h_z(t))$ be a coordinate of hand position in the 3-D camera coordinate system. For each frame, features of human actions are calculated as following equations:

$$H_f(t) = f_y(t) \tag{1a}$$

$$D_{fh}(t) = |F(t) - H(t)| \tag{1b}$$

$$A_{fh}(t) = \arccos \frac{f_y(t) - h_y(t)}{D_{fh}(t)} \quad (1c)$$

$$R_{fh}(t) = D_{fh} - \frac{1}{T} \sum_{k=t-T}^{t-1} D_{fh}(k) \quad (1d)$$

$$S_f(t) = |F(t) - F^{'}(t)| \quad (1e)$$

$$S_h(t) = |H(t) - H^{'}(t)| \quad (1f)$$

$$M_f(t) = \arccos \frac{f_y^{'}(t) - f_y(t)}{S_f(t)} \quad (1g)$$

$$M_h(t) = \arccos \frac{h_y^{'}(t) - h_y(t)}{S_h(t)}, \quad (1h)$$

where

$$F^{'}(t) = \frac{1}{T} \sum_{k=t-T}^{t-1} F(k), \quad H^{'}(t) = \frac{1}{T} \sum_{k=t-T}^{t-1} H(k).$$

Since DBNs require discrete values as input, these feature values are quantized into some states as follows.

- Height of face (H_f)
  H_f is quantized into three states, such as *standing position* (high), *seated position* (middle), and *squatting position* (low).
- Speed of face or hand (S_f, S_h)
  S_f and S_h are quantized into three states, such as *stationary*, *moving fast*, *moving slow*.
- Direction of movement (M_f, M_h, A_{fh})
  M_f, M_h, and A_{fh} are quantized into five states of angle. These thresholds are $\frac{\pi}{8}$, $\frac{3\pi}{8}$, $\frac{5\pi}{8}$ and $\frac{7\pi}{8}$.
- Distance from face to hand(D_{fh})
  D_{fh} is quantized into three states, such as *near*, *middle*, and *far*.
- Change of distance from face to hand (R_{fh})
  R_{fh} is quantized into three states, such as *stationary*, *become near*, and *become far*.

4. Recognizing Object Using DBNs

Object regions are recognized as posterior probabilities of networks including three DBNs. In the following, a typical DBN model is presented first, and then application of DBNs to our methods described.

4.1. A Typical DBN

DBNs are suitable for handling time-series of data. A typical form of DBN is shown in Figure 3 representing state transition from time 0 to *t*, where S_t is a hidden node and

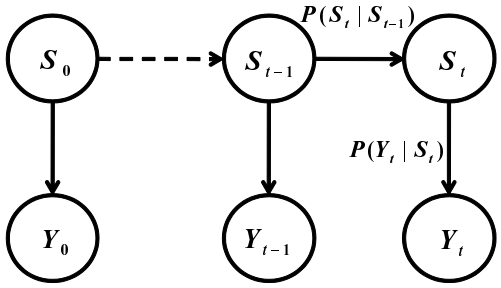


Figure 3. Typical model of DBN.

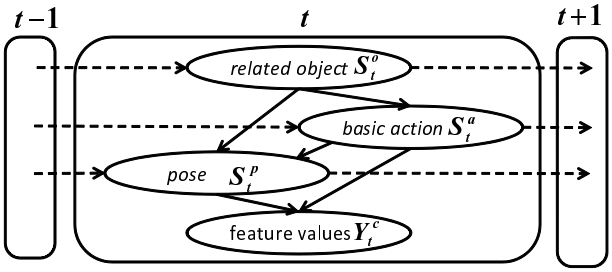


Figure 4. Probability Networks.

Y_t is an observed node at time t. $P(Y_t|S_t)$ and $P(S_t|S_{t-1})$ represent observed probability at time t and state transition probability from time $t-1$ to t, respectively. Let $Y_{1:t}$ be a series of observed values until time t, posterior probabilities $P(S_t|Y_{1:t})$ is calculated as:

$$P(S_t|Y_{1:t}) = \alpha P(Y_t|S_t) \sum_{S_{t-1}} P(S_t|S_{t-1}) P(S_{t-1}|Y_{1:t-1}) \quad (2)$$

where α is a normalization constant such that $\sum_i s^i = 1$ for $S_t = \{s^0, s^1, ...\}$.

If and only if multiple observed values $Y_t^1, Y_t^2, \ldots, Y_t^n$ are independent, observed probability $P(Y_t|S_t)$ is written as

$$P(Y_t|S_t) = \prod_{i=1}^{n} P(Y_t^i|S_t). \quad (3)$$

4.2. DBNs of Human Actions and Objects

Figure 4 shows a probability network for recognizing human actions and objects from feature values explained in section 3. Human actions are divided into two categories: *pose* S_t^p and *basic action* S_t^a. This is mainly because human actions have some hierarchies as mentioned in [8]. In what follows, we explain DBNs corresponding to each node S_t.

- *Pose* DBN
  S_t^p consists of three states shown in left column of Ta-

Table 2. States of *human pose* and *basic action*.

Human pose S_t^p	Basic action S_t^a
standing	nothing
sitting	sit down
walking	stand up
	take object
	set papers to printer
	write into whiteboard
	put a object to shelf
	put in a wastebasket

ble 2. Posterior probabilities of node S_t^p are calculated from feature values Y_t^c.

- *Basic action* DBN
  Right column of Table 2 shows states of node S_t^a representing momentary actions related with objects. Posterior probabilities of node S_t^a are calculated from both Y_t^c and posterior probabilities of *pose*.

- *Related object* DBN
  Posterior probabilities of node S_t^o are calculated from posterior probabilities of *pose* and *basic action*. The number of states at node S_t^o corresponds to that of categories of objects.

4.3. Observed and State Transition Probability Matrices

Observed probability matrices and state transition probability matrices are statistically calculated from video sequences for learning. For each frame, states of *pose*, *basic action* and *related object* are labeled manually.

First, these matrices of *pose* DBNs are calculated from statistics of labeled state $Y_{1:t}$ and $S_{1:t}^p$. Next, for each frame, posterior probabilities of *pose* are calculated by applying *pose* DBNs to learning sequences. Matrices of *basic action* are, then, calculated from labeled state $Y_{1:t}$ and $S_{1:t}^a$, and posterior probabilities of *pose*. In a similar way, matrices of *related object* are calculated from labeled state S_t^o and posterior probabilities of *pose* and *basic action*.

4.4. DBNs of Object Regions

In general, indoor environment contains multiple regions of objects. Each object region have a *related object* DBN individually as described in section 2. Supposing that a person handles only one object at the same time, the related region should be selected by calculating relative positions of human and objects. When a face position comes in or goes out from an object region at time t, frames from $t-N$ to $t+N$ is extracted as a section corresponding to an action.

Figure 5. A person is putting the trash in the wastebasket. Rectangles represent regions of face and hand.

Table 3. Observed probabilities of H_f.

		H_f		
		low	middle	high
S_t^p	standing	0.043	0.056	0.891
	sitting	0.092	0.903	0.005
	walking	0.156	0.008	0.836

Frames at which hand position is in a object region are also extracted.

Finally, object probabilities of the object region o are given by:

$$P^o = \beta \sum_w P_w^o \tag{4}$$

where $P_w^o = (p_0, p_1, \dots)$ is a vector of posterior probabilities of object o at last frame of section w, and β is a normalization constant such that $\sum_i P_i = 1$.

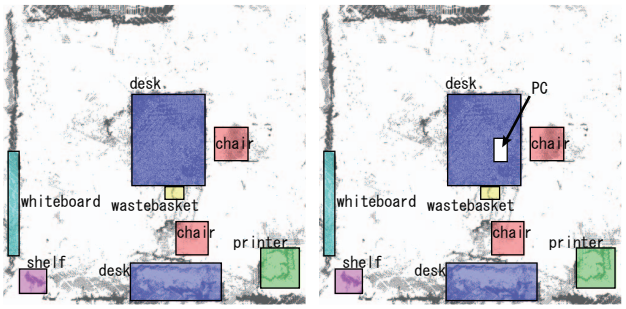
5. Experiments

To test the effectiveness of the proposed method, two experiments were performed as shown in Figure 5. In these experiments, PC (Intel Core 2 Duo T7700 2.0GHz with 4GB memory), and stereo camera (SRI International Inc.) were used. Indoor environment was captured at resolution of 320×240 pixels and 15 frames per second, T was set to be 5, and N was set to be 15.

Video sequences were captured for learning in advance, and then each probability matrix was calculated as described in section 4. Table 3 and 4 show examples of observed probability matrix and state transition probability matrix of *pose* DBN, respectively.

Table 4. State transition probabilities of S_t^p.

		S_t^p standing	sitting	walking
S_{t-1}^p	standing	0.981	0.006	0.012
	sitting	0.004	0.995	0.000
	walking	0.017	0.003	0.980



(a) Environmental map without overlapped region.

(b) Environmental map with laptop PC.

Figure 6. Environmental Map

5.1. Experiments without Overlapped Regions

Experiments were performed where objects were not overlapped with each other as shown in Figure 6(a). In these experiments, target objects for recognition were desk, chair, printer, wastebasket, shelf and whiteboard. Ten video sequences were captured in different view points, and seven of them were used, in which action features and sections related with objects were well-extracted.

The transition of desk's probabilities is shown in Figure 7. In this graph, human actions at the desk region was captured in frame 240-260, and 320-340. In Figure 8, similarly, all frames in a frame section 1800-1950 correspond to human actions at the whiteboard region. The result of recognition was shown in Table 5. Hatched elements of this denote correct results. In these experiment, correct probability for each region presented the highest value at 87.5% in desk region. Some objects were, however, not clearly discriminated such as whiteboard and shelf.

5.2. Experiments with Overlapped Region

In the following experiments, laptop PC was on the desk as shown in Figure 6(b), then regions were overlapped each other. In addition to this, the action of using PC was added to *basic action* of Table 2.

Table 5. Object probabilities for each object region.

Region	Object Probability (%) desk	chair	printer	waste-basket	shelf	white-board
desk	87.5	1.8	1.0	7.9	1.7	0.1
chair	11.0	26.0	13.7	25.0	18.5	5.7
printer	3.6	16.1	40.0	9.6	25.7	5.0
wastebasket	0.9	5.8	12.6	39.5	38.9	2.3
shelf	0.2	0.9	9.1	20.1	44.9	24.9
whiteboard	0.1	0.2	1.8	2.4	47.6	47.9

Table 6. Object Probabilities with PC

Region	Object Probability(%) desk	char	printer	waste-basket	shelf	white-board	PC
desk	28.3	2.5	0.8	7.7	1.7	0.1	59.0
chair	6.8	43.1	8.2	25.2	14.1	1.2	1.3
printer	1.0	16.0	36.8	9.9	25.9	5.0	5.4
can	1.1	5.0	10.9	40.8	39.5	2.0	0.6
shelf	0.5	0.8	8.7	20.0	45.4	24.7	0.0
whiteboard	0.2	0.2	1.7	2.2	47.5	48.2	0.0
PC	7.4	0.3	0.0	0.2	0.0	0.0	92.1

Table 6 shows the result of the experiment. This shows that most of the results were almost the same as the experiment 5.1, however, different results were found in desk and PC. The correct probability of PC region was high at 92.1%. The PC probability of desk region, that is false, was also high at 59.0%

5.3. Discussions

We made discussions on the experiments as follows.

- Selection of action-related object
  In the experiments, we found that wrong region was selected as related object in some results. For example, a desk region was selected as a target of person's action at frame 240-260 in Figure 7, while correct target was chair. In this frame section, in spite of the fact that human actions was related with chair region, desk region was detected. This is because the corresponds between human actions and object regions are often mistaken due to inaccuracy on 2-D environmental map and placement of object regions. This problem can be solved by obtaining accurate relative positions of objects to human body in a 3-D environmental map.

- Discrimination between similar actions
  Objects related with similar actions were difficult to discriminate, such as whiteboard and shelf. For exam-

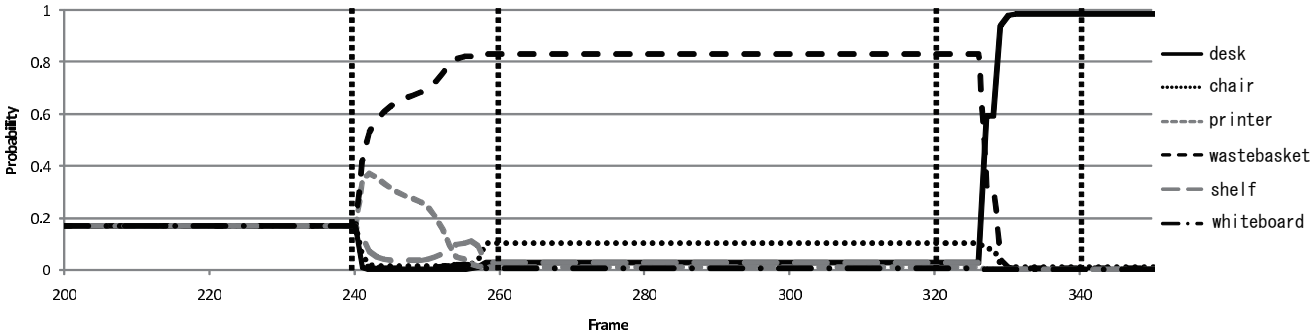


Figure 7. Transition of desk's probabilities.

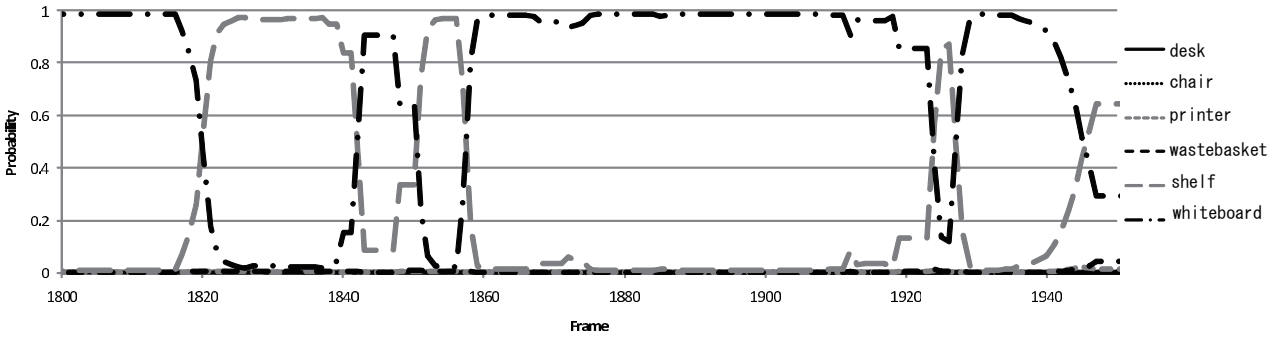


Figure 8. Transition of whiteboard's probabilities

ple, writing to a whiteboard is similar to taking some stuffs from a shelf. Such relationship is also found in PC and desk. When a person handles objects on the desk including PC, he/she takes almost the same pose. In order to discriminate these objects, estimation of human actions should be improved.

6. Conclusion

This paper presents the method for recognizing objects from human actions using probability networks representing relationships between actions and objects. The networks consist of DBNs considering human poses and actions interacting with objects. In the experiments, we demonstrated the effectiveness of the proposed method for recognizing objects. The problem of this method, however, was found that some objects related with similar actions were difficult for recognition. In the future work, we are to improve our method by incorporating human actions into dynamic segmentation of objects in a 3-D map.

References

[1] D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. In *IEEE International Conference on Computer Vision 1999*, Mar. 1999.

[2] P. Peursum, S. Venkatesh, G. A. W. West, and H. H. Bui. Object labelling from human action recognition. In *IEEE International Conference on Pervasive Computing and Communication 2003*, pages 399–406, 2003.

[3] M. Higuchi, S. Aoki, A. Kojima, and K. Fukunaga. Scene recognition based on relationship between human actions and objects. In *17th International Conference on Pattern Recognition*, volume 3, pages 73–78, Aug. 2004.

[4] M. Mitani, M. Takaya, A. Kojima, and K. Fukunaga. Environment recognition based on analysis of human actions for mobile robot. In *the 18th International Conference on Pattern Recognition*, volume 4, pages 782–786, Aug. 2006.

[5] M. Saitou, A. Kojima, T. Kitahashi, and K. Fukunaga. Dynamic recognition of human actions and objects using dual hierarchical models. In *the First International Conference on Innovative Computing*, pages 306–309, Aug. 2006.

[6] K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, 2002.

[7] A. Kojima, T. Tamura, and K. Fukunaga. Textual description of human activities by tracking head and hand motions. In *16th International Conference on Pattern Recognition*, volume 2, pages 1073–1077, Aug. 2002.

[8] M. Kitahashi, A. Kojima, M. Higuchi, and K. Fukunaga. Toward a cooperative recognition of human behaviors and related objects. In *15th European Japanese Conference on Information Modelling and Knowledge Bases*, pages 321–330, May 2005.