Extraction of Human Activities as Action Sequences using pLSA and PrefixSpan

Takuya TONARU[†] Tetsuya TAKIGUCHI[††] Yasuo ARIKI[††]
Graduate School of Engineering, Kobe University[†]
Organization of Advanced Science and Technology, Kobe University[††]
tonaru@me.cs.scitec.kobe-u.ac.jp takigu@kobe-u.ac.jp ariki@kobe-u.ac.jp

Abstract

In this paper, we propose a framework for recognizing human activities in our daily life. Since a human activity is represented as a sequence of actions, the actions are recognized from videos and then the frequently-occurring human activities can be extracted from them. We show the experimental results applied to the data taken in a deskwork environment to demonstrate the performance of the proposed framework. The experimental results were as follows: 86.0% averaged recall rate and 78.3% averaged precision rate were obtained in extracting human activities.

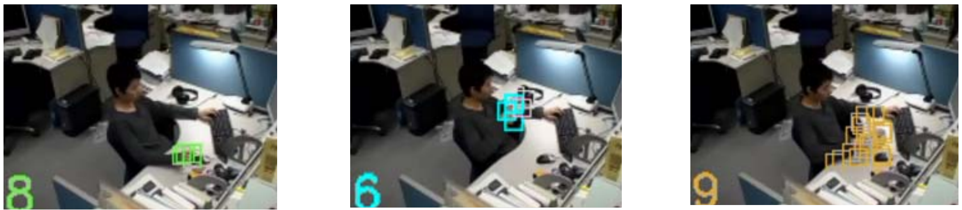
1. Introduction

Today, it is easy to record individual daily activities in video sequences. To analyze human activities in video sequences is valuable for tasks that can give helpful information to users or support their lives. For example, at a desk in an office, workers mainly use computers, sometimes drink coffee, or wear headphones to listen to music. If someone drinks coffee too much, a life-support system analyzes his activities and will issue a warning about his health. Hence our goal is to automatically detect, categorize and recognize human daily activities.

There has been much research carried out on recognition of simple actions [1] [2], such as running, walking, hand waving, boxing, etc. Niebles showed interesting results for unsupervised learning and recognition of multiple actions using pLSA models [2]. However in an actual environment, a person acts by combining various simply actions. Hence, recognition of daily human activity cannot be achieved by merely extending the previous framework.

Previous research has represented human activity as a symbolic sequence of actions in hierarchy. One popular approach applied Stochastic Context-Free Grammar (SCFG) to the symbolic sequence of actions to analyze their structure [3] [4]. However, grammar was given manually. Hamid has analyzed the human activity in the kitchen environment using a SuffixTree from a sequence of interactions with key-objects [5].

In this paper, we propose a method to analyze human activities using video, by detecting and categorizing actions based on an unsupervised learning approach and to recognize the human activities from these actions based on sequential data mining. The learning cost in obtaining a symbolic sequence of actions can be reduced by adopting the unsupervised approach. Under the assumption that daily human activities appear frequently, sequential data mining shows strong potential for obtaining frequently-appearing activities from symbolic sequences of actions in a video.



(a) “reaching for a cup” (b) “taking a cup” (c) “putting a cup down”

Figure 1. A sequence of actions forming an activity of “Drinking Coffee”. The number in the lower left indicates each action.

2. Activity representation

We define the human activity in this section. Human activity consists of various actions, and it is represented as a symbolic sequence of actions. For example, an activity S, in which a person is drinking coffee is represented as a sequence of actions as follows:

$$S \rightarrow 8 \quad 6 \quad 9$$

The numbers 8, 6, and 9 indicate the actions of “reaching for a cup”, “taking the cup”, and “putting the cup down,” respectively, as shown in Fig. 1. An activity of drinking coffee is usually represented as a flow of actions such as taking the cup, lifting the cup to the mouth, and putting the cup down. A temporal flow of such actions constitutes a human activity.

3. Approach

Our method consists of two phases. In the first phase, a histogram sequence of actions is obtained using human action categorizing method [2]. In the second phase, the obtained action histogram sequence is converted into discretized symbolic sequence of actions, and human activities are extracted using PrefixSpan based on the frequency.

3.1. Human action categorizing method

This method extracts spatial-temporal features and learns the action models using a pLSA model. Here, a brief review of this method is described.

3.1.1. Feature representation: Assuming a stationary camera or a process that can account for camera motion, separable linear filters are applied to the video to obtain the response function as follows

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where I is a gray-scale pixel on the image, g(x,y;σ) is a 2D Gaussian smoothing kernel, applied only along the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)\exp(-t^2/\tau^2)$ and

$h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)\exp(-t^2/\tau^2)$. The two parameters σ and τ correspond to the spatial and temporal scales of the filters, respectively. To give the response function effectively, we use $\omega = 4/\tau$.

This function detects any regions where complex motion is caused spatially. In fact, a region with complex motion can induce a strong response, but a region with simple translational motion will not induce a strong response. The spatial-temporal interest points are extracted around the local maxima of the response function. At each interest point, a spatial-temporal cube is extracted that contains the output of the response function. Its size is approximately six times the spatial and temporal scales along each dimension. To obtain a motion descriptor, the brightness gradients are computed at all the pixels in the cube and are concatenated to form a vector. Then PCA is applied to reduce the dimensionality of the descriptors.

In order to obtain the cluster prototypes, a k-means algorithm is applied to the descriptors. Then each descriptor is assigned a descriptor type by mapping it to the prototype. Therefore a collection of descriptors included in a video is represented as a histogram of the descriptor types. Hereafter, we will refer to the descriptor types as words in videos.

3.1.2. Action categorization by pLSA: The pLSA (Probabilistic Latent Semantic Analysis) method is a technique used in the analysis of co-occurrence data. This method can find meaningful topics that correspond to motion categories in terms of words in videos.

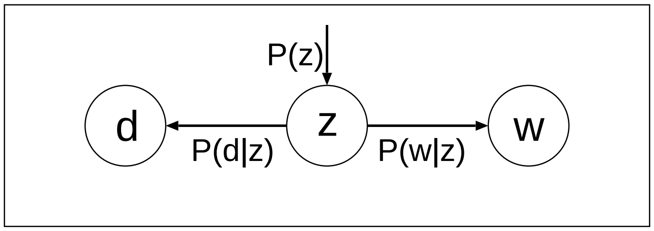


Figure 2. PLSA graphic model of symmetric parameterized version

We can create a co-occurrence table N between a word w_i in $W = \{w_1,\ldots,w_M\}$ and a video d_j in $D = \{d_1,\ldots,d_N\}$ using the feature extraction method described in 3.1.1. In addition, there is a latent topic variable z_k in $Z = \{z_1,\ldots,z_K\}$, which is not observed yet. Assuming that the observation pairs (w_i,d_j) are generated independently under the condition of the latent topic variable z_k, a joint probability model is given by

$$P(w_i,d_j) = \sum_{k=1}^{K} P(z_k)P(w_i \mid z_k)P(d_j \mid z_k) \tag{2}$$

where $P(w_i|z_k)$ is the probability of a word w_i occurring in an action category z_k, and $P(d_j|z_k)$ is the probability of video d_j occurring in an action category z_k. This model is a symmetric parameterized version of the generative model [6], and its graphic model is represented in Fig. 2.

We then determine the model parameters P(z), P(w|z) and P(d|z) by maximization of the log-likelihood function

$$L = \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) \log P(w_i, d_j) \tag{3}$$

where $n(w_i,d_j)$ denotes the word frequency, that is the number of times word w_i occurred in video d_j. Maximizing the log-likelihood function yields a model that gives high probability to the words that appear in the video. The procedure for maximization of the log-likelihood function is the Expectation Maximization (EM) algorithm.

When testing the model, each word in the testing video d_{test} is labeled topically by finding the following maximum posteriors:

$$P(z_k \mid w_i, d_{test}) = \frac{P(w_i \mid z_k) P(z_k \mid d_{test})}{\Sigma_{l=1}^{K} P(w_i \mid z_l) P(z_l \mid d_{test})} \tag{4}$$

Since $P(z|d_{test})$ is not obtained, it is required to be computed. Although this can be solved using an EM algorithm in the same way as training the model.

3.1. Extraction of activities

3.2.1. Action recognition by human action categorization: It is necessary to prepare video clips that include actions as learning data. However, in our method, it is not necessary to clip each action precisely from the videos because the pLSA model is a multi-topic analysis method. If two actions occur consecutively without a non-movement gap, they will be clipped as one video sequence. The pLSA model can find these action categories separately as latent topics. Accordingly, video sequences for learning are extracted easily and automatically from videos.

When learning using the pLSA model, it is necessary to decide topic K, which is the number of categorized actions. If topic K is large, although an action vocabulary becomes large, it will respond sensitively to the small difference of the feature. If topic K is small, it does well in dealing with noise, but the action vocabulary becomes small. Future research will consider how to deal with this problem automatically.

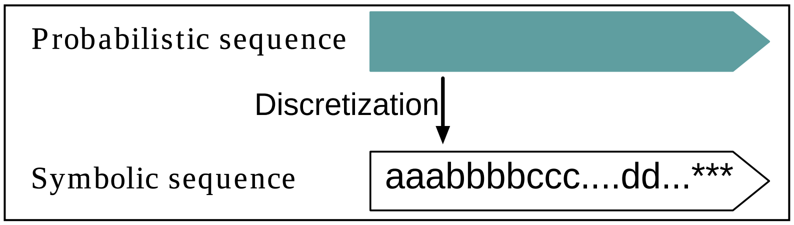


Figure 3. Conversion into a discretized symbolic sequence

3.2.2. Converting into discretized symbolic sequence: The result of action recognition for the testing video d_{test} is a histogram sequence of actions computed frame by frame. This histogram is $P(z_k|d_{test})$ as described in section 3.1.2. This histogram sequence is smoothed for denoising, and each frame is replaced by the action symbol with the maximum probability as shown in Fig. 3.

Next, the consecutive same symbols are merged into one as shown in Fig. 4. In addition, since human activity is a sequence of consecutive actions, if non-movement duration is longer than some threshold, the sequence is split into two sequences.

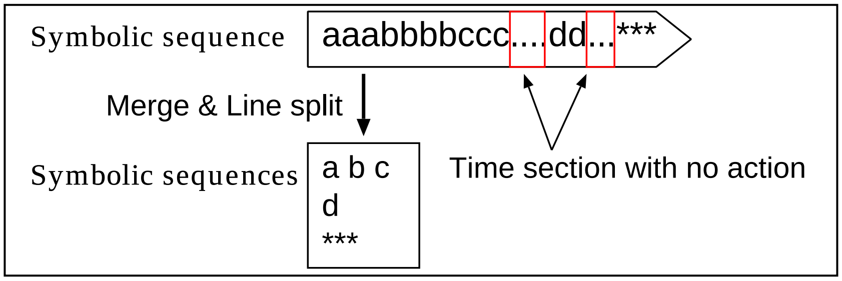


Figure 4. Conversion into symbolic sequence by merging and splitting

3.2.3. Extracting human activities: We assume that human daily activities appear frequently. To extract activities, PrefixSpan (*Prefix*-projected *S*equential *PA*tter*N* mining) [7], commonly used in sequential data mining, is employed. As shown in Fig. 5, frequent subsequences are discovered as patterns in a sequence database, where the occurrence frequency of subsequences is no less than minimum support. Its general idea is to examine only the prefix subsequences and project only their corresponding postfix subsequences into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns [7]. A mining result is a list of action sequences and they are sorted in the order of frequency. Next, the extracted sequences are manually labeled as activities if they represent the human activities.

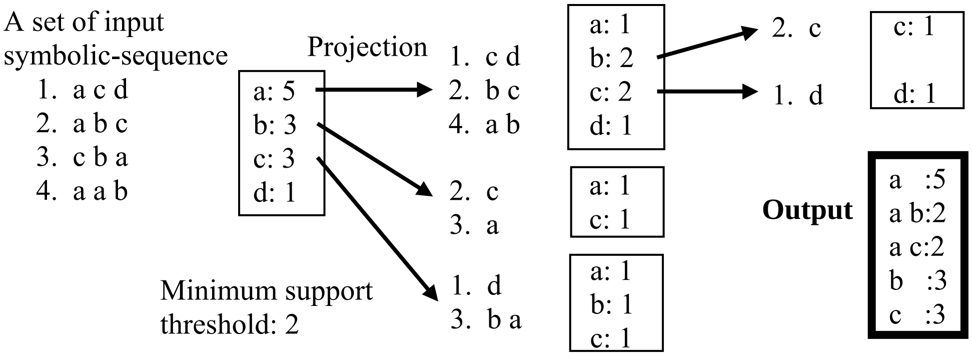


Figure 5. Frequent subsequences extraction by PrefixSpan

4. Experimental results

4.1. Experimental conditions

We verified the validity of our algorithm using a 70-minute-long video in which a person is working at a desk in the laboratory. In video, the person uses a computer and sometimes drinks coffee, wears or removes headphones, picks up or throws away tissues, and scratches his head. No one else appears in the video, and the person does not leave the desk. The resolution of the video image is 160×120.

The spatio-temporal features were extracted as described in section 3.1.1. with the two parameters $\sigma = 11$ and $\tau = 19$. A codebook containing 400 codewords was created from the training set descriptors. The latent topic K was set to 13, and the minimum support value of PrefixSpan was set to 3. A symbolic sequence was split into two if the non-movement duration is longer than 120 frames.

4.2. Experimental results

The number of human activities extracted by PrefixSpan was 43, and six activities were extracted in the order of frequency. Table 1 shows the extracted human activities. Fig. 6 shows examples of extracted human activities as images.

Table 1. Human activities extracted by the proposed method

Activity	Frequence	Sequence	Recall	Precision
Drink coffee	16	6 9	1.00	0.91
	7	6 11 9		
Remove headphones	7	4 10 3	0.86	0.86
Pick up tissues	5	8 12	0.80	0.80
Scratch the head	4	4 13	0.50	0.67
Wear headphones	3	4 7	1.00	0.86
	3	4 10 7		
Throw away tissues	3	12 10 9	1.00	0.60

In Table 1, two different sequences appear in the same activity. For example, “Drink coffee” has two different sequences: “6 9” and “6 11 9”. This is caused by slow speed of action. In Fig. 6(a) and 6(b), the action in the middle was inserted when the speed of the arm motion was very slow.

In Table 1, the averaged recall and precision are 86.0% and 78.3%, respectively. The definition of the recall and precision is as follows:

Recall = (True positive) / (True positive + False negative) (×100[%])

Precision = (True positive) / (True positive + False positive) (×100[%])

The definition of true positive, false positive and false negative is given as follows:

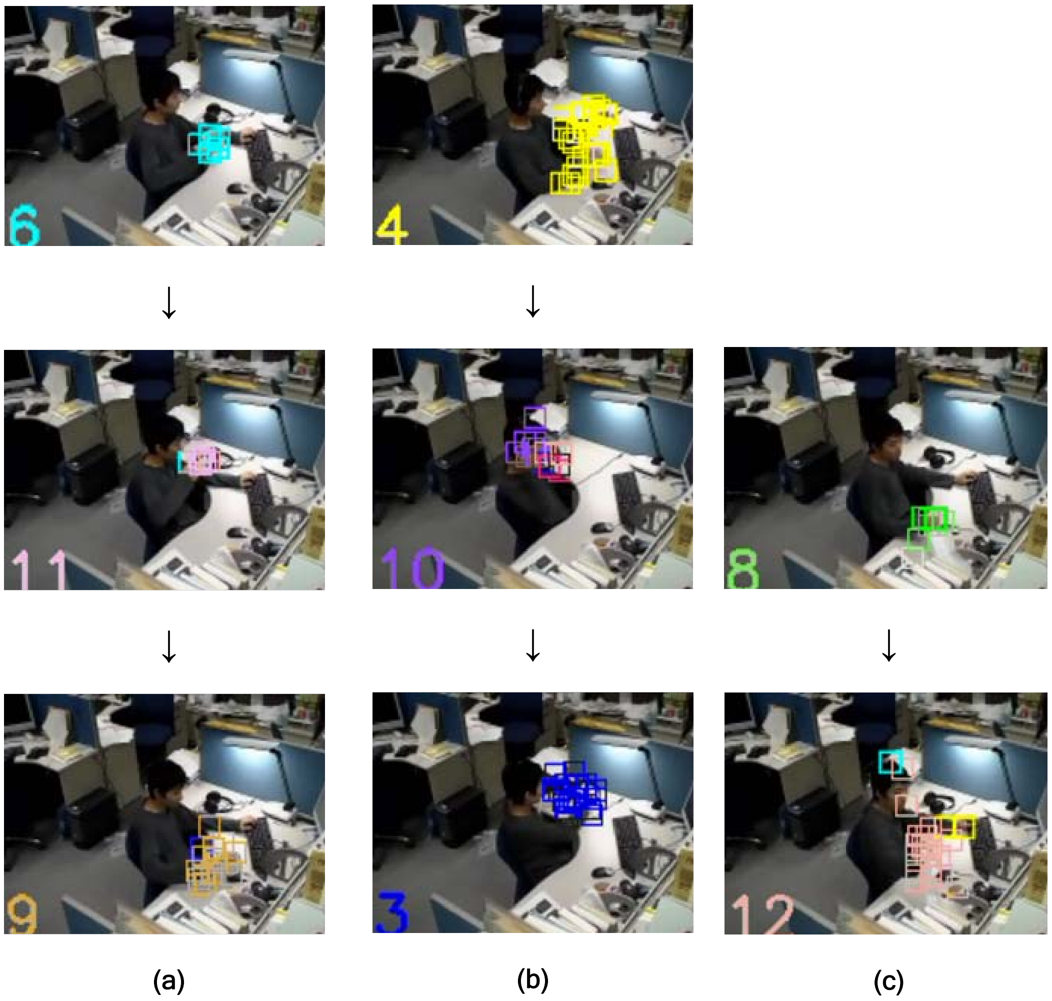
True positive :
the number of correctly extracted activities
False positive :
the number of falsely extracted activities

False negative :
the number of true activities not extracted

5. Conclusion

We proposed a framework for recognizing human activities by analyzing videos. The goal of our work is to automatically convert a video sequence into a symbolic sequence of actions and to extract frequently-occurring human activities from the symbolic sequences.

In the future, we are planning to directly extract human activities from an action histogram sequence by taking into consideration the duration of actions and by permitting multiple activity candidates.



(a) (b) (c)

Figure 6. Human activities extracted by the proposed method.
Each image shows
(a) “drinking coffee”, (b) “removing headphones”, (c) “picking up tissues”

6. References

[1] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach,” ICPR, pp. 32-36, 2004.

[2] J.C. Niebles, H. Wang, and Li. Fei-Fei, “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words,” British Machine Vision Conference, pp. 1249-1258, 2006.

[3] Y. Ivanov and A. Bobick, “Recognition of Visual Activities and Interactions by Stochastic Parsing,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 852-872, 2000.

[4] D. Minnen, I. Essa, and T. Starner, “Expectation Grammars: Leveraging High-Level Expectations for Activity Recognition,” CVPR, pp. 626-632, 2003.

[5] R. Hamid, S. Maddi, A. Bobick, and I. Essa, “Unsupervised Analysis of Activity Sequences Using Event-Motifs,” VSSN, pp. 71-78, 2006.

[6] T. Hofmann, “Probabilistic Latent Semantic Indexing”, SIGIR, pp. 50-57, 1999.

[7] J. Pei, J. Han, M. Behzad, and H. Pinto, “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth,” ICDE, pp. 215-224, 2001.

Authors

Takuya Tonaru is the graduate student at Kobe University.

Tetsuya Takiguchi received the Dr. Eng. degree in information science from Nara Institute of Science and Technology, Nara, Japan, in 1999. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Japan. He is currently a Lecturer with Kobe University. From May 2008 to September 2008 he was a visiting scholar at University of Washington. His research interests include speech and image processing. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of the IEEE, the Information Processing Society of Japan, and the Acoustical Society of Japan.

Yasuo Ariki received his B.E., M.E. and Ph.D. in information science from Kyoto University in 1974, 1976 and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IPSJ, JSAI, ITE and IIEEJ.