Robust Rule-Based Method for Human Activity Recognition

Masafumi Sugimoto, Thi Thi Zin, Takashi Toriu and Shigeyoshi Nakajima

Graduate School of Engineering, Osaka City University, Sugimoto 3-3-138, Sumiyoshi-ku, Osaka, Japan

Summary
Human activity recognition is an active research field in computer vision and image processing. In this paper we propose a robust rule-based method for the task of recognition of human activities and scenarios in video image sequences. The methodology uses a context-free grammar based representation scheme to represent human actions. The proposed system consists of three major steps. Initially by using a single camera, in a variety of angles, the movement of the object is detected, and object silhouette is generated in each frame. Then, the proposed method for generating a silhouette is presented. The silhouette is used to determine human activities such as running and walking. In the last stage, a rule-based classifier is used to classify the action. The experimental results show that the system can recognize seven types of primitive actions with high accuracy.

Key words:
Human activity, rule-based, motion parameter, appearance parameter.

1. Introduction

Recent advances in computer vision and pattern recognition have fuelled numerous initiatives that aim to intelligently recognize human activities. High-level understanding of human activity is essential for various applications, including surveillance systems and human computer interactions. In particular, a human activity recognition system may enable the detection of abnormal activities as opposed to the normal activity of persons using public places like airports and subway stations. Automated human activity recognition may be useful for real-time monitoring of the elderly people, patients, or babies. Specifically, human action recognition aims at automatically telling the activity of a person, i.e. to identify if someone is walking, dancing, or performing other types of activities. It is a crucial prerequisite for a number of applications, including surveillance, content-based image retrieval, or human robot interaction. The task is challenging due to changes in the appearance of persons, articulation in poses, changing backgrounds, and camera movements.

Several methods have been used to process the features for scenario recognition. In particular, Bayesian approaches and Hidden Markov Models (HMM) have been extensively used to detect simple and complex events that occur in the scenarios. This paper shows that an alternative and simpler approach, based on control rule based scheme is equally effective in detecting activities that occur in a scene. The action recognition system also exploits objects' silhouettes obtained from video sequences to classify actions. It mainly consists of two major steps: manual creation of silhouette and action templates offline and automatic recognition of actions in real-time. In classifying actions of humans into predetermined classes like walking, boxing and kicking; temporal signatures of different actions in terms of silhouette poses are used.

In this paper, we concentrate on rule based activity recognition. We infer action classes based on a single recognized atomic action or based on a sequence of recognized poses. The action classes considered are often referred to as atomic or primitive actions whereas more complex activities can be understood as a sequencing of these primitive actions.

Several researchers have worked on human activity recognition at various levels [1-5]. Some researches focus on simple tracking of persons, and others focus on estimating the physical state of persons in the scene [1, 6-8]. Further, various analyses on human actions have been conducted [9-11]. Most of the previous researches focused mainly on the recognition of single (i.e. atomic) actions of humans, and complex composition of multiple movements or actions [11, 12].

Currently, there are two major approaches towards moving object classification which are shape-based and motion-based methods [1]. Shape-based methods [13] make use of the objects' 2D spatial information like bounding rectangle, area, silhouette and gradient of detected object regions; whereas motion-based methods use temporally tracked features of objects for the classification solution. In this paper, we present the most significant works classified individual activity type.

In individual activity analysis most of works focus on activities such as walking, jumping, and standing up, sitting and getting up. The human body's pose frequently gives an indication of the action that takes place. Human activity recognition system based on angular poses and velocities of the main human body parts (hands, legs, and torso) has been analyzed in [12]. Some researchers described human activity as a temporal sequence of pose vectors that represent sampled poses of these body parts [14]. An example of a human activity is a sequence of actions in which the subject enters in a room, sits down,

Manuscript received April 5, 2011
Manuscript revised April 20, 2011

stands up and walks away. Each component of the human activity, such as walking, sitting down, is a discrete action primitive. There are no distinct breaks between the executions of different actions. The angles subtended by three major components of the body (the torso, the upper component of the leg and the lower component of the leg) with the vertical axis are used as a feature vector for classifying the human actions. Traditionally, human activity recognition has been achieved mainly by the statistical pattern recognition techniques such as the Nearest Neighbor Rule (NNR), and the state-space methods, e.g. HMM [15-16].

The paper is organized as follows. In Section 2, we introduce a rule based action recognition method. We present experimental results in Section 3 followed by conclusions in section 4.

2. Proposed Human Action Recognition Method

In this section, we describe our approach to human action recognition system. When humans speak of recognizing an action, they may be referring to a set of visually observable transitions of the human body such as 'raise hands', or an abstract event such as 'a person entered the room'. While recognizing the former requires only visual knowledge about certain movements of the human body, the latter requires much more than purely visual knowledge: it requires that we know about rooms and the fact that they can be 'entered into' and 'exited from', along with the relationships of these abstract concepts to lower level visual actions. In this section, we shall deal with the automatic view-invariant recognition of low level visual actions based on context free grammars. An overview of the proposed algorithm is shown in Fig. 1.

The algorithm can be divided into three steps based on low-level, intermediate-level and high-level vision processing. The low-level vision step includes video data acquisition, background-foreground separation, silhouette extraction and representation. In this paper, we do not deal with the issue of background subtraction, which is a widely studied topic of research in itself. The sequences we have used were obtained using a white background, and background subtraction becomes a straightforward task. Given such a sequence, the issue at hand is how to find a representative sequence of key features to describe the action being seen. For a given video after background removal, we define appearance and motion based features. In order to achieve scale invariance, we normalize the extracted features. At the high-level vision step, we perform feature space analysis to compute the activity decisions for each frame, and smooth these decisions over time to maintain smooth activity transitions.

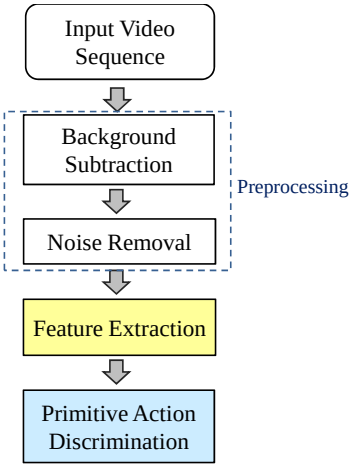


Fig. 1 Overview of proposed human action recognition.

We start by defining a comprehensive set of features that, intuitively, may distinguish the different activities. Thus here, we will tackle human behavior analysis problems focusing on intermediate-level and high-level vision processing. First, we introduce two classes of features: motion features and appearance features, which represent activities on the different spatial scales. Second, a rule-based algorithmic analysis and pattern matching are presented to model and recognize interacting activities, combining the two classes of features. The advantages of the approach are: (i) these two classes of features represent human interactions more accurately and completely; (ii) the rule-based algorithmic analysis and pattern matching reveal behavior structures well.

2.1 Intermediate Level Vision Processing

In this processing we perform:

- Appearance and motion features extraction module,
- Feature inter-relationships establishment module.

These modules will extract the following features and their relationships. Appearance features includes Gravity Center (GC), height, aspect ratio, dispersion, angle and pixel ratio. Motion features includes speed, velocity, acceleration and step of the tracked subject. The target velocity and speed are obtained through differentiation of the instantaneous position estimate. Since the target inter-frame displacements can be very small, the temporal derivative can be quite noisy. Hence, we introduce different ways of averaging the velocity and speed estimates over an interval of fixed number of frames. For details, we define the following terminologies and notations.

Appearance parameters

	Terminology	Notations
1	Position at time t	$p(t) = (x(t), y(t))$
2	Gravity Center (CG)	$GC(t) = (x_c(t), y_c(t))$ $x_c = \sum_{i=1}^{N} \frac{x_i}{N}$ $\quad x_i = \begin{cases} x_i & \text{if } p(x_i, y_i) = 1, \\ 0 & \text{otherwise.} \end{cases}$ $y_c = \sum_{i=1}^{N} \frac{y_i}{N}$ $\quad y_i = \begin{cases} y_i & \text{if } p(x_i, y_i) = 1, \\ 0 & \text{otherwise.} \end{cases}$
3	Height (H) Width (W)	$h(t)$ $w(t)$
4	Aspect ratio	$A(t) = w(t)/h(t)$
5	Dispersion	$D = \frac{\text{Perimeter}^2}{\text{Area}}$
6	Angle	$A_L = \tan^{-1} \frac{y_c - y_L}{x_c - x_L}$, $A_R = \tan^{-1} \frac{y_c - y_R}{x_R - x_c}$
7	Pixel ratio	$R = \frac{\#(U)}{\#(L)}$, U and L are the upper and lower part of bounding box.

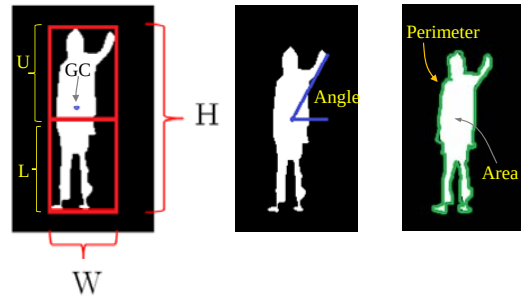
Motion parameters

	Terminology	Notations
1	velocity	$v(t) = (x(t)\text{-}x(t\text{-}1), y(t)\text{-}y(t\text{-}1))$
2	speed	$s(t) = \sqrt{(x(t) - x(t-1))^2 + (y(t) - y(t-1))^2}$
3	Acceleration	$\lambda(t) = v(t) - v(t\text{-}1)$

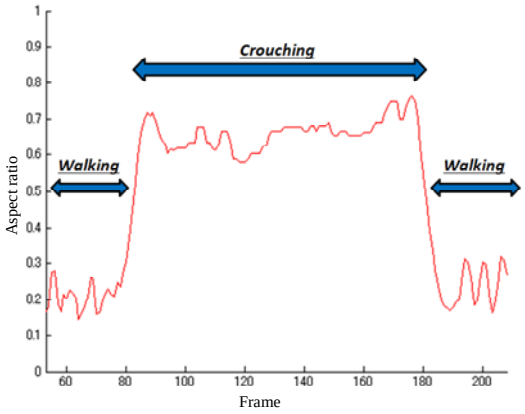
The intuition behind these features is that frames where the point of high average acceleration reaches a minimum indicate flow reversals which occur when the body reaches an extreme pose. Frames at the maxima are points where the body is exactly in between two extreme configurations, and is in the middle of a transition undergoing large overall movement. The appearance feature concepts are illustrated in Fig. 2 and motion parameter illustration is shown in Fig. 3.

2.2 High-Level Vision Processing

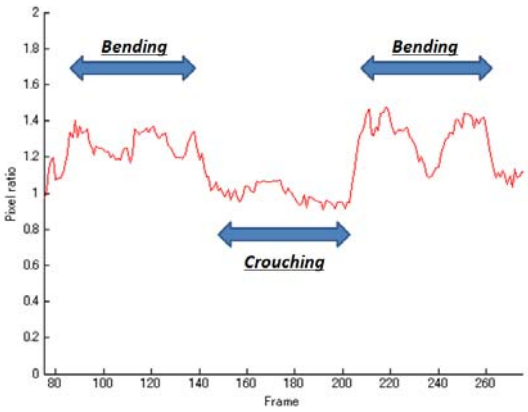
Now, we will use the extracted features for detecting human actions. In this respect, our High-Level Vision Processing will proceed as follows. Apply the moving average smoothing filter to the graphs of aspect ratios and moments ratios. In particular, the aspect ratio is used to discriminate between stand and squat or crawls, and the angle of inclination is an easy and reasonable selection to roughly distinguish between walk and run because the pose of body is different between these two movements.



(a)



(b)



(c)

Fig. 2 Some appearance parameters used in the proposed algorithm: (a) detected individual region on background subtraction image, (b) illustration of aspect ratio on walking and crouching, and (c) illustration of pixel ratio on bending and crouching.

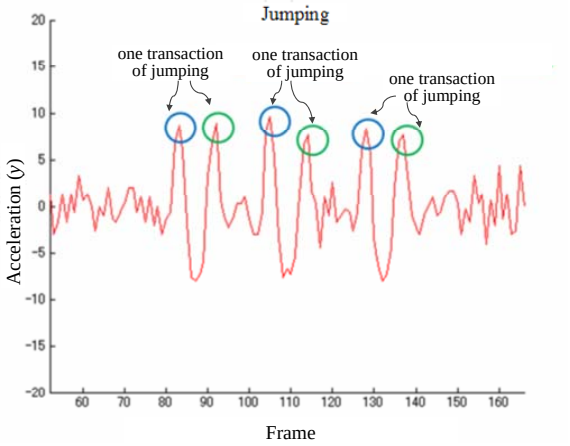


Fig. 3 Illustration of motion parameter on jumping action.

In addition, when the aspect ratio is less than threshold value (can be estimated) and the moment ratio less than different threshold then two possibilities can rise: walking or loitering. In this case, check the value of speed/velocity norm ratio. According to laws of motions in Physics, **the ratio between the average speed and the norm of the average velocity is approaching 1** if the target always moves in the same direction along a straight line and **tending to 0** if it moves irregularly in various directions, or returns to the initial position. So we can apply this rule for detection of loitering. If *y*-value is greater than threshold then hopping action is detected otherwise check the difference between two successive values of *x*-positions. This value can differentiate crawling and lying down (suspicious action).

The hierarchical expansion is a particularly useful technique, which relates to the aggregation hierarchy of sub-rules that constitute an action. These can be classified into two categories, concrete and abstract hierarchical expansion. Concrete hierarchical expansion is the relevant objects are well-defined. For example, an action of "standing" can be expanded to bending, walking, waving, jumping, and so on as shown in Fig. 4. On the other hand, an abstract hierarchical expansion is the actions are not concretely defined. Although the rules are not definite visual actions, they contain certain common characteristics. In order to perform direct extraction of high-level actions automatically, we establish associations between low-level appearance and motion features with high-level concepts to classify possible actions. An illustration of such associations is described in Fig. 5.

3. Experimental Results

In this section, we present the experimental setup used to evaluate the performance of the proposed algorithm. We also created our own activity database comprising of videos of eight activities: standing and walking, running, sitting, squatting, crawling, lying down and pointing. We used SONY HDR-CX550V/XR550V to capture indoor and outdoor activity videos with frame size of 1440x1080. The only preprocessing done on the data was to temporally down-sample the video sequences to 10fps and spatially reduced each frame to size 320x240. This step not only reduced the size of the data set (thus enabling faster computation) but also provided limited blurring, which smoothed the derived silhouette contour to an extent. The data set was developed with subjects of different physical builds; wearing indoor as well as bulky out-door clothing; moving in front of a stationary camera with static lighting conditions and a relatively static background Actions were performed such that the view angle changed frequently and limited occlusion occurred. We constructed 279 video cuts of activity database. These videos were combined to produce objects tracked frames from which the motion and appearance features are extracted. Our system correctly can detect target activities giving very promising results and indicate that there is much to be said in favor of simple methods even when the problem is complex.

We prepared two tables to show the experimental results. In Table 1, the results are organized based on number of video sequences we have tested. It can be seen that the average accuracy rate is 95%. In Table 2, we organize the results based on the number of video cuts. It shows that the accuracy rate is well above 98%. A considerable degree of correlation across the ensemble of activities is observed from the videos. For example, the Jumping and Squatting have similar kind of frames for most of their activity durations (in this case, standing still for a short duration).

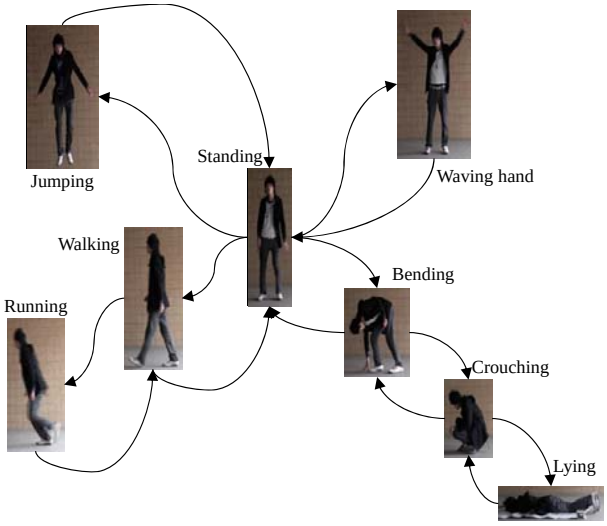
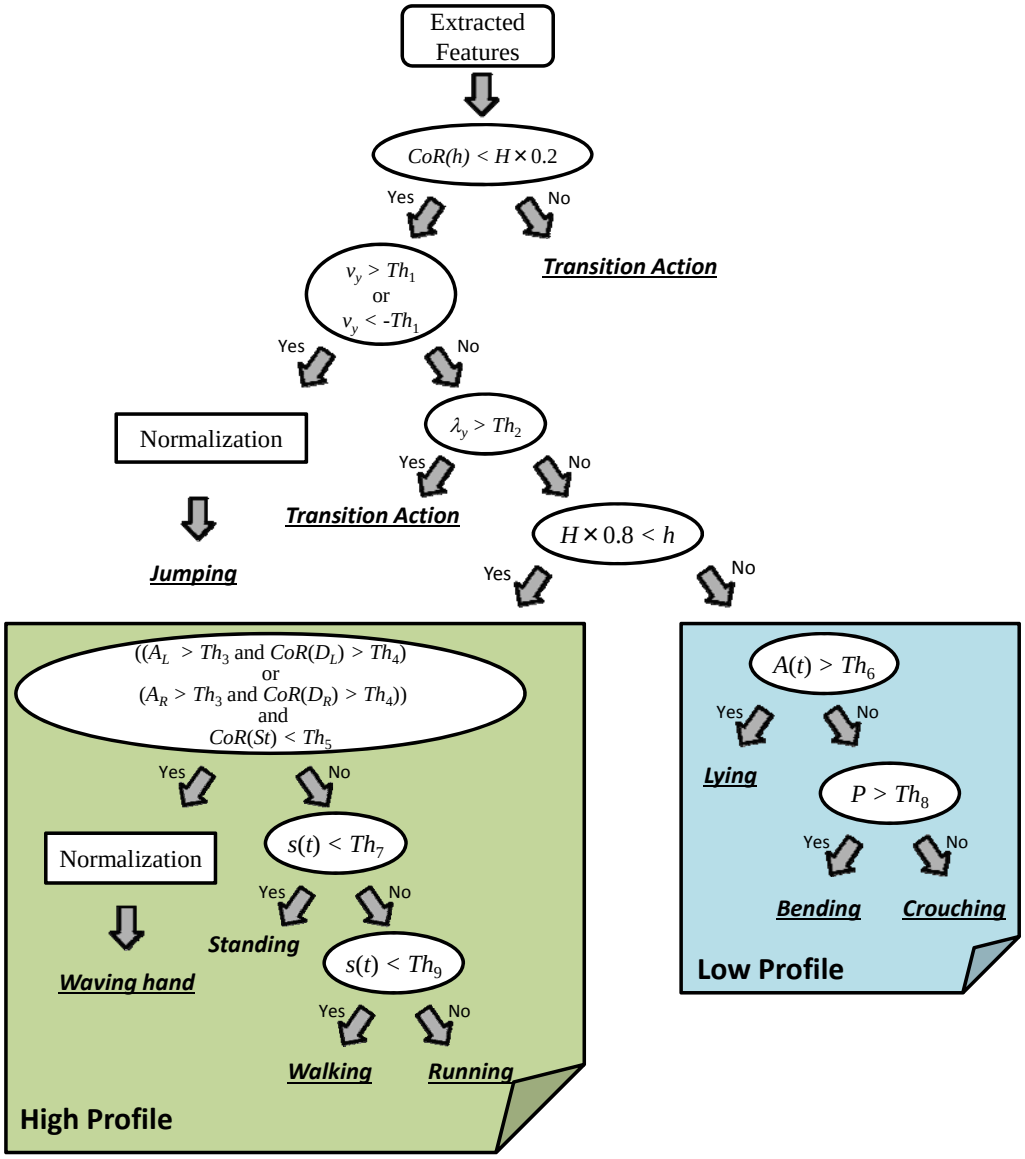


Fig. 4 Actions Transitions.



$CoR(h)$: rate of change of silhouette's height
H : expected silhouette's height
h : actual silhouette's height
v_y : velocity in y direction
λ_y : acceleration in y direction
A_L , A_R : arm angle (left, right)
$CoR(D_L)$, $CoR(D_R)$: rate of change of dispersion on left upper box, right upper box
$CoR(St)$: rate of change of foot position
$s(t)$: speed
$A(t)$: Aspect ratio
P : pixel ratio
$Th_1 \sim Th_9$: predefined Thresholds

Fig. 5 Our proposed algorithm.

Table 1: Performance results based on video sequences

Primitive Action	***Sequence***	***Error***	Accuracy Rate (%)
Walking	9	1	89
Running	9	0	100
Jumping	9	1	89
Bending	9	0	100
Crouching	9	1	89
Lying	9	0	100
Waving hand	9	0	100
Total	63	3	95

Table 2: Performance results based on video cuts

Primitive Action	***No. of Cut***	***Error***	Accuracy Rate (%)
Walking	90	2	98
Running	27	2	92
Jumping	54	0	100
Bending	36	1	97
Crouching	27	1	96
Lying	18	0	100
Waving hand	27	0	100
Total	279	6	98

In order to retain only the visually significant information, background subtraction and normalization is performed on all the frames. Motion compensation is performed to center the subject for activities where locomotion is involved. The occasional misclassification is present between activities which share spatial coherence to a large degree, for example jumping and waving. The accuracy over the various body activities is well over 90 %.

4. Conclusions

We have presented a framework for learning to represent various kinds of human activities which can be used for recognizing them efficiently. A low-dimensional representation is learned which captures the spatial and temporal aspects of activities ideal for applications involving quick activity recognition. Although this approach does not cover all problems arising in the behavior analysis context, we suggest that simple classification rules such as those based on motion and appearance feature concepts, should be integrated with other methods such as Hidden Markov Models, possibly using a successive refinement strategy.

Acknowledgments

This work was supported in part by SCOPE: Strategic Information and Communications R&D Promotion Program (10103768). We thank Scope project members and the students of Physical Electronics and Informatics of Osaka City University, for their participations in producing tested videos.

References

[1] J. K. Aggarwal and Q. Cai, “Human motion analysis: A review”, Computer Vision and Image Understanding, pp. 428-440, 1999.

[2] K. Huang, S. Wang, T. Tan, and S. J. Maybank, “Human Behavior Analysis Based on a New Motion Descriptor”, IEEE Trans. on Circuits and Systems for Video Technology, vol.19, no. 12, Dec. 2009.

[3] Ana Paula B. Lopes, Rodrigo S. Olivera, Jussara M. de Almeida, Arnaldo de A. Araujo, “Spatio-Temporal Frames in a Bag-of-visual-features Approach for Human Actions Recognition”, Proc. of 22nd the Brazilian Symposium on Computer Graphics and Image Processing, Rio de Janeiro, Brazil, pp. 315-321, Oct. 2009.

[4] M. S. Ryoo and J. K. Aggarwal, “Recognition of Composite Human Activities through Context-Free Grammar based Representation”, Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR’06), New York, USA, pp. 1709-1718, 2006.

[5] D. Mahajan, N. Kwatra, S. Jain, P. Kalra, and S. Banerjee, “A Framework for Activity Recognition and Detection of Unusual Activities,” Proc. of Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2004), Dec. 16-18, 2004.

[6] Pyke Tin, Thi Thi Zin, H. Hama and T. Toriu, “Challenges and Promises in Human Behavior Understanding Research”, In the 4th Intl. Symp. on Intelligent Informatics (ISII2011), Qingdao, China, May. 1-3, 2011. (Accepted)

[7] T. B. Molesund and E. Granum, “A survey of computer vision based human motion capture”, Computer Vision and Image Understanding, pp. 231-268, 2001.

[8] L. Wang, W. Hu and T. Tan, “Recent developments in human motion analysis”, Pattern Recognition, pp. 585-01, Vol. 36, 2003.

[9] B. Fan, Z.-F. Wang, “Pose estimation of human body based on silhouette images”, Proc. of Intl. Conf. on Information Acquisition, pp. 296-300, 2004.

[10] J. Krumm, S. Harris, B. Meyers, B. Brummit, M. Hale, S. Shafer, “Multi-Camera multi-person tracking for easy living”, Proc of 3rd IEEE Intl. Workshop on Visual Surveillance, 2000.

[11] S. Dagtas, W. A. Khatib, A. Ghafoor and R. L. Kashyap, “Models for motion-based video indexing and retrieval”, IEEE Trans. on Image Processing, vol. 9(1), pp. 88-101, Jan. 2000.

[12] Jezekiel et al, “Human activity recognition using multidimensional indexing”, IEEE Trans. on Pattern

Analysis and Machine Intelligence, Vol. 24, No. 8, Aug. 2002.
[13] H-C. Mo, J-J. Leou, and C-S. Lin, "Human Behavior Analysis Using Multiple 2D Features and Multicategory Support Vector Machine", Proc. of IAPR Conference on Machine Vision Applications (MVA2009), Yokohama, Japan, May 20-22, 2009.
[14] Anjum Ali and J. K. Aggarwal, "Segmentation and recognition of continuous human activity", Proc. of the IEEE Workshop on detection and recognition of events in video, 2001.
[15] H.C.C. Tan and Liyanage C. De Silva, "Human activity recognition by head movement using Elman Network and Neuro-Markovian Hybrids", Image and Vision Computer, New Zealand, 26-28 Nov. 2003.
[16] F-S. Chen, C-M. Fu, C-L. Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models", Image and Vision Computing, 21, pp. 745-758, 2003.

Masafumi Sugimoto received the B.E. in Information Engineering from Osaka City University, Osaka, Japan in 2011. He is now in his Master first year in Information Engineering, Graduate School of Engineering, Osaka City University. His research interests include image processing and human behavior and action recognition.

Thi Thi Zin received the B.Sc. degree (with honor) in Mathematics in 1995 from Yangon University, Myanmar and the M.I.Sc degree in Computational Mathematics in 1999 from University of Computer Studies, Yangon, Myanmar. She received her Master and Ph.D. degrees in Information Engineering from Osaka City University, Osaka, Japan, in 2004 and 2007, respectively. From 2007 to 2009, she was a Postdoctoral Research Fellow of Japan Society for the Promotion of Science (JSPS). She is now a specially appointed Assistant Professor of Graduate School of Eng, Osaka City University. Her research interests include human behavior understanding, ITS, and image recognition. She is a member of IEEE and WIE.

Takashi Toriu received the B.Sc. in 1975, M.Sc. and Ph.D. degree in physics from Kyoto University, Kyoto, Japan, in 1977 and 1980, respectively. He was a researcher in Fujitsu Laboratories Ltd. from 1982 to 2002, and now he is a Professor of Osaka City University. His research interests are in the areas of image processing, computer vision, and especially in modeling of human visual attention. He is a member of IEEE, IEICE, IPSJ, ITE and IEEJ.

Shigeyoshi Nakajima received the B. E. and M. E. degree in Electric Engineering from Kyoto University, Kyoto, Japan in 1982 and 1984, respectively. He received the Ph.D. degree in Information Engineering from Osaka City University, Osaka, Japan in 1997. He is now an Associate Professor in Osaka City University, Japan. His research interests include signal processing, image processing, medical engineering and optimization algorithm and so on. He is a member of IEEE, IEICE and IPSJ.