Human Activity Recognition Using a Mobile Camera

Kai-Tai Song and Wei-Jyun Chen
Department of Electrical Engineering, National Chiao Tung University, Hsinchu, 300, Taiwan
(Tel : +886-3-573-1865; E-mail: ktsong@mail.nctu.edu.tw; wjc12615.ece97g@nctu.edu.tw)

Abstract **-** This paper presents a vision-based human activity recognition system using a mobile camera. This system aims to enhance human-robot interaction in a home setting for applications such as health care and companion. In the first place, the camera needs to find a human in image frames. The body pose is classified for the detected human. Then the human activity is recognized by combining information of human pose, human location and elapsed time. In order to determine the situated place of the person in a home setting, a novel space-boundary detection method is proposed in this paper. This method uses features in the environment to automatically set space boundary in the image such that human location in the environment can be obtained. In the integrated experiments, human pose recognition rate of five poses(standing, walking, sitting, squatting, lying) is 94.8%. Experiments of human activity recognition in a home setting have been conducted to verify the performance of the proposed method by using a mobile camera from different view angles and positions in a home setting. The experimental results reveal that the space boundaries are detected as expected and satisfactory results are obtained.

Keywords – vision system, human activity recognition, human pose recognition, human-robot interaction.

1. Introduction

In recent years, various domestic and service robots have been developed for many human-centered applications. Various robots have been designed to assist human with health care, surveillance and other assistive tasks. In such applications, the robot very often needs to understand its working environment as well as the person to be assisted. To work with human, it is important for a robot to identify human activities in the environment. Using the human activity information, a robot can provide timely responses and services autonomously.

It is clear that human activity can be inferred from body poses and the situated location of human in the environment. Many related works have been presented on solving human activity recognition problems and many powerful tools have been reported. Fujiyoshi and Lipton[1] used the Star-Skeleton to describe human body movement. Chen[2] proposed that human action is composed of a series of postures. In [3 [4], authors use feature matching approach to recognizing human actions. Since human actions are composed of many different postures, therefore learning mechanisms have been applied to the detection processing. Carter *et al.*[5] combined Bayesian and Markov chain as a decision-making mechanism to identify human actions. Kellokumpu *et al.*[6] achieve processing of human activity detection by using the Hidden Markov Model(HMM).

To understand human activity, in addition to analysis human body poses, one also need to give different meaning of a body pose in different situated environments. Wang *et al.*[7] proposed a human activities detection system using low-level vision and high-level vision. The low-level vision includes human detection and human tracking, and the high-level vision is responsible for behavior understanding. Zhou *et al.*[8] used a fixed monocular camera to set regions in environment, such as kitchen, doorway and bathroom. When a person is detected in one of these areas, the system can speculate the activity of this person. Shiomi *et al.*[9] reported an application of human activity detection of a robot in a shopping mall. They used laser scanner to estimate the position of customers in the environment. Their system can recognize four walking behaviors of a customer, such as fast walking, idle walking, wandering and stopping. The main purpose was to make the robot know the current activity of customers in the shopping mall, then the robot can provide appropriate response and service to customers.

Although techniques of human detection and pose recognition are essential for human activity recognition, current related studies are not sufficient to describe human activity for a camera onboard a mobile robot in practical applications. The problem of activity recognition using a mobile camera deserves urgent attention. In this paper, we will focus on how to estimate the situated environment of human in the image using a monocular camera. To cope with the problem caused by a mobile camera, we propose a method to automatically set space boundary in the image frame when using a mobile camera. After the pose and situated location of the person are detected, a finite state machine(FSM) will combine these information to find out the activity of the person.

978-1-4577-0723-0 / 11 / $26.00 ⓒ 2011 IEEE

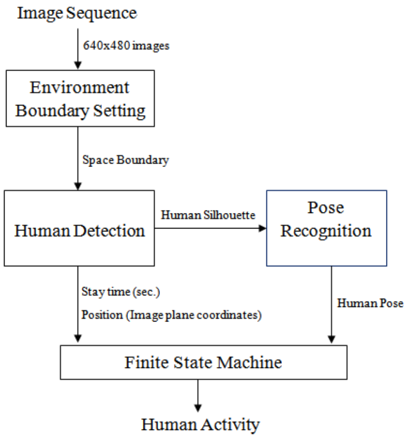

Fig. 1. Overall system architecture of activity recognition.

2. System Architecture

The system architecture of the proposed human activity recognition is depicted in Fig.1. The camera onboard a mobile robot acquires 640x480 images. The first processing step of the algorithm is the space boundary setting. Using the virtual space boundary, human location in the environment can be determined. The second step is human detection. As a human is detected in the image, then the stay time and position of the human can be obtained. The situated location of the human in the environment can be determined by using the virtual space boundary. Then a silhouette of human image is generated for body pose recognition. According to a series of human motions, the current pose is recognized by using star-skeleton and HMM. Finally, these data are fed to a finite state machine to determine a suitable human activity at home.

2.1 Human Detection

Fig. 2 shows the block diagram of human detection system. It consists of two parts: human detection and SVM classifier. For robotic applications in a home setting, complicated background and illumination variation pose a problem to image processing. To cope with this problem, we adopted histogram of oriented gradient(HOG)[10-11] to extract features in human image. In the training phase of HOG, we used 1000 human images and 1000 non-human images as the training data from INRIA person dataset[12] and MIT pedestrian database[13]. Another 1000 human images and 1000 non-human images are used as test data to evaluate the SVM classifier. The test results reveal that true positive detection rate is 96.8% and false negative rate is 1.5%. For the on-line detection part, when an image sequence is acquired, image subtraction is first performed to segment foreground from the images, then a region of interest(ROI) is set by using image projection and the HOG's features extracted. Finally, features are fed to the SVM classifier to identify which is a human.

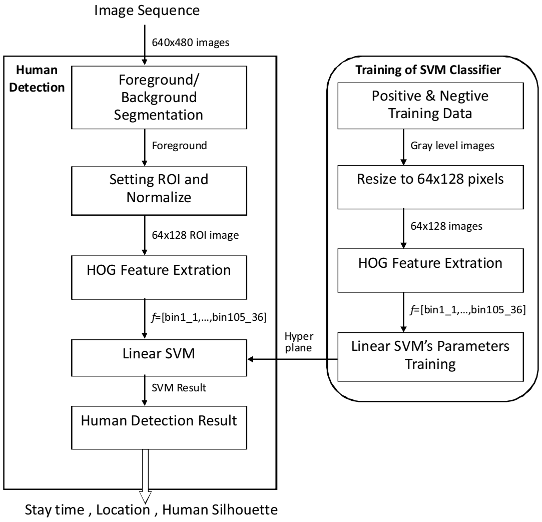

Fig. 2. Block diagram of human detection system.

2.2 Pose Recognition

After a human is detected, the stay time, location and silhouette of the human imagery can be calculated. We then use star-skeleton features [1] for body pose recognition. In this design, HMM is applied to train and detect body poses of a human. The developed pose recognition architecture is shown in Fig. 3. Five common poses can be classified, namely standing, walking, sitting, squatting and lying. Among them, lying is identified by the body geometric proportions. After a human is detected, a frame is setup for the human shape to mark the human. If the frame's width/height is greater than 1.5, it will be identified as lying posture, otherwise the system enters the feature extraction step. In feature extraction, silhouette contour extraction is performed and the star-skeleton is used to describe the human contour. These star-skeleton features will be represented as a feature vector. We use these feature vectors to build a codebook and each feature vector is represented by a number, namely the motion number, then a pose is combined of a sequence of motions. Our codebook has four poses which we want to recognize, as shown on Fig. 4. The number on the bottom is the motion number of each feature vector.

When the feature vector is extracted, it will be mapped

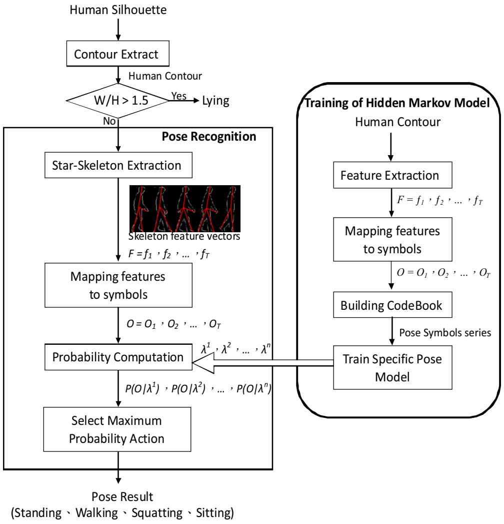

Fig. 3. Block diagram of body pose recognition.

to a codebook and get the representative motion number, then output a sequence of motion number. We use the motion numbers of each pose to build four Hidden Markov Models of each pose (standing, walking, sitting, squat). In the recognition step, the star-skeleton feature of motions is obtained and map to the codebook and output a sequence of motion numbers; then feeding these motion number to pose-HMMs to find the maximum probability pose..

3. Proposed Activity Recognition Scheme

In this paper, we want to locate a human in the environment and to integrate pose and location information to determine the human's activity. For example, for a human who is watching TV in the living room, the living room is the location of the human. But in the same imagery, there might exist several rooms of the house such as living room, dining room as shown in Fig. 5(a). A space boundary is needed to distinguish between the dining room and living room in the image plane. Furthermore, this system aims to be used by a mobile robot. When the robot moves, the boundary in the image plane will also move accordingly, as shown on Fig. 5(b). In the following, we propose a method to automatically set space boundary in the image frame.

In this method, an object is assigned first to work as a landmark as shown on Fig. 5 (red frame of the recognized object). After recognizing the object in image frame, homogarphy transform[14] is applied to identify the object location on the image plane. Because the boundary and object is fixed in the environment, there is a relationship between the object and environmental boundary in image plane. When the robot at different view angle, we can use this characteristic to calculate the relation between object location and boundary. The environmental boundary setting consists of two parts: object recognition and relation finding between object location and boundary. The procedure for finding the relation between object location and environment boundary are summarizes as follows:

1) Placed the camera away from the nearest boundary (if near to the boundary, it will disappear in the latter steps), record the four corners of the object coordinates and the space boundary position.

2) Placed the camera away from the farthest boundary to the camera (a distance that is able to recognize the landmark object), record the four corners of the object coordinates and the space boundary position.

3) Place the camera in the middle distance between the farthest and the nearest, record the four corners of the object coordinates and the space boundary position.

4) Using step1~3 to find the geometric relationship between the object and the space boundary.

The object recognition steps are summarizes as follows:

1) Find out matching points between input image and database image using speed up robust feature(SURF)

2) Using homography to identify the four corners of the object coordinates in the image plane.

3) Set the new space boundary by the derived relationship function(see later).

The flowchart of this procedure is illustrated in Fig. 6.

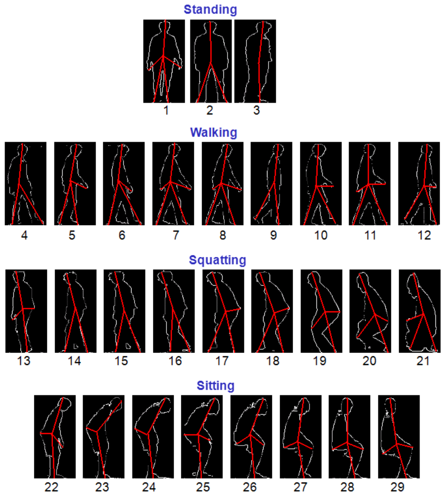

Fig. 4. The codebook of four body poses.

3.1 Object Recognition

We adopted the speed up robust features (SURF) to extract object features, because SURF feature are invariant to orientation, scale, change in illumination of the image. After getting the features of the object, we can match the features to the database then to know the object is on the image plane or not. In this design, we use nearest neighborhood algorithm to match the database and current image. The nearest neighbor is defined as the key point with minimum Euclidean distance foe descriptor vector such as (1). Then we use the homography to find the object coordinates on the image plane. (Fig. 7.)

$$d = \left(\sum_{i=1}^{128} \left(Des_c(i) - Des_d(i) \right)^2 \right)^{1/2} \qquad (1)$$

where Des_c is the current image feature descriptor vector and Des_d is the database feature descriptor vector and i denotes how many feature descriptors need to match at the same time.

3.2 Relationship Between Object and Boundary

After object recognition, we can determine the object location in the image plane, then we want to identify the environmental boundary through the object location. Therefore, we must define the relationship between the object and the boundary.

In this study, there exist some preset conditions for the object: 1) The object area should be large enough; 2) It should be a planar object; 3) It needs to be parallel to the

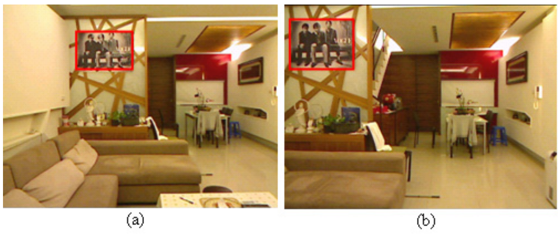

Fig. 5. Images from different camera view angles.

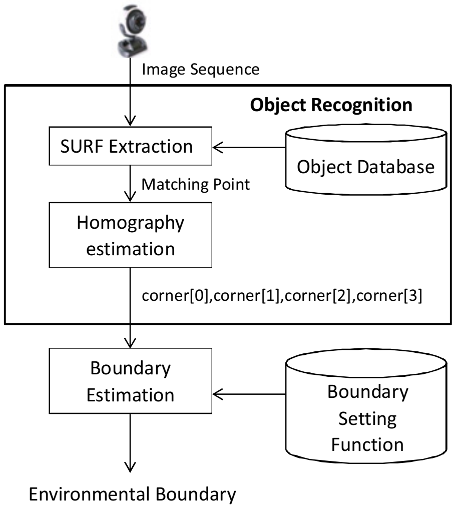

Fig. 6. System architecture of automatic space boundary setting.

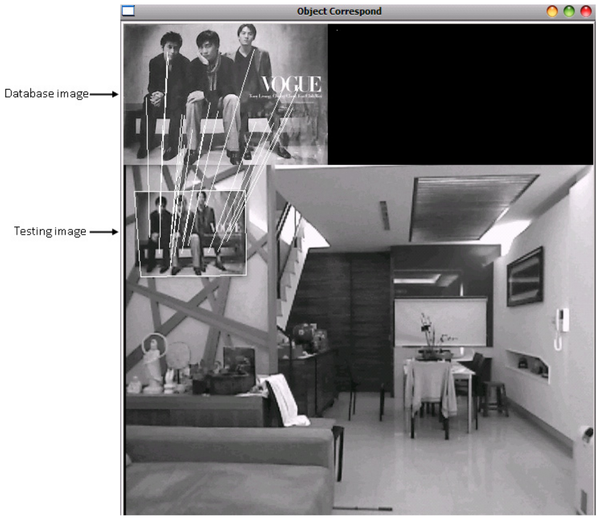

Fig. 7. Object coordinates in the image plane.

ground. We first define parameters of object and environmental boundary as shown in Fig. 8. In the figure,

$$d_1 = y_3 - y_0, \tag{2}$$

where y_3 is y coordinate of the lower left corner point on the object, and y_0 is y coordinate of the top left corner point on the object. The slope of the bottom line on the object is such that:

$$m_{23} = \frac{y_2 - y_3}{x_2 - x_3}, \tag{3}$$

where y_2 is y coordinate of the lower right corner of the object, x_2 is x coordinate of the lower right corner of the object and x_3 is x coordinate of the lower left corner point on the object. As shown in Fig. 8, the environmental boundary is obtained by combination of the reference point (b_x,b_y) and it's slope m_b . Due to the object in the real world is parallel to the ground, we will set the same slope for the object and the virtual boundary in image plane as (4). We define the x coordinate of reference point and top left corner point on the object the same as (5) , and the y coordinate of reference is defined as (6).

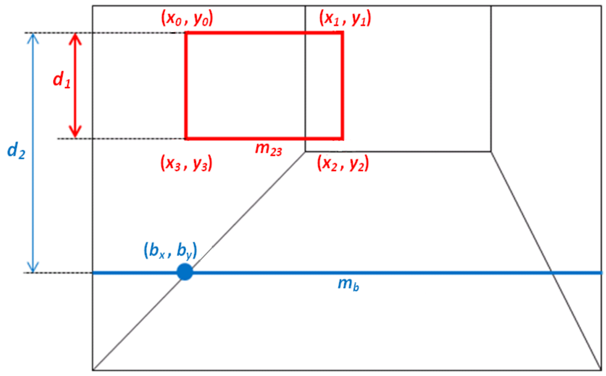

Fig. 8. The parameters of virtual space boundary and the landmark object.

$$m_b = m_{23} \quad , \tag{4}$$

$$b_x = x_0 \quad , \tag{5}$$

$$b_y = y_0 + d_2 \, , \tag{6}$$

where x_0 is x coordinate of the top left corner point on the object. We set three different distances between object and camera to derive the functional relationship between object and boundary, then we obtains d_1 and d_2 in (7) for these three different distances. In the current implementation, the farthest is d_1=71, d_2=257, the middle is d_1=93, d_2=341 and the nearest is d_1 =114, d_2=449. Assume a functional relationship of d_1 and d_2 such that

$$d_2 = \alpha * d_1^{\,2} + \beta * d_1 + \gamma \tag{7}$$

We put each d_1 and d_2 into (7), then we can get α=0.0308, β=-1.2341 and γ=229.3572, so this linear equation can be expressed as :

$$d_2 = 0.0308 * d_1^{\,2} - 1.2341 * d_1 + 229.3572 \tag{8}$$

In summary, when an image is acquired and the object successfully recognized, (8) is applied to automatically set the environmental boundary.

3.3 Finite State Machine

When human location, elapsed time and body poses are obtained, we need a method to integrate these information to infer the activity. A finite state machine(Fig.11) was designed to determine the human activity from calculated data. In Fig. 11, two locations can be identified: dining room and living room. For example, when a human crossing the boundary of dining room, the activity is “go to the dining room”, if the human stays in the dining room for over 10 frames, then the activity is ” in the dining room”, and so on. In particular, when the body pose change from standing directly into lying, the activity will be recognized as "fall down".

4. Experimental Results

In this section, we present several interesting experiments to examine the proposed human activity recognition design.

4.1 Human Detection and Pose Recognition

In human detection, we tested the recognition rate at different angles(0^{o}, 45^{o}, 90^{o}, 135^{o}, 180^{o}, 225^{o}, 270^{o}, 315^{o}). The detection rates are shown in Table. 1. In the experiment of human detection, the detection rate is from 90.66% to 99%, average is 95.33%. In the body pose recognition, we tested the recognition rate of five poses in different distances from the camera(2.5m, 3m, 3.5m, 4m, 4.5m, 5m).The detection rate is shown in Table. 2. In the experimental of pose recognition, the accuracy rate is from 86% to 100%, average rate is 94.8%.

4.2 Experimental Results of Automatic Boundary Setting

In this experiment, we first determine the object(a painting on the wall) in the environment. This object is with a size of 80cm x120cm and the edge is parallel to the ground. To verify the accuracy of the virtual boundary, we drew a black line on the ground as the ground truth boundary in the experiment. The camera was 3.5 meters from the object as shown in Fig. 9.

The system automatically set the boundary as shown in Fig. 10(a). As we turned the camera 10 degrees to the right, a new boundary was calculated as shown in Fig. 10(b). Fig.10(c) is referred to turning the camera 10 degrees to the left. We then moved the camera forward for 1 meter, such that the camera was 2.5 meters from the object. After we moved the camera, a new space boundary was calculated as shown in Fig. 10(d). We then turned the camera 5 degrees to right and left respectively, new boundaries were set as shown in Fig. 10(e) and Fig. 10(f). In summary, as we move the camera in a range of 1 meter and 10 degrees, the boundary is set successfully.

Table 1 Human detection rate at different angles.

	0^{o}	45^{o}	90^{o}	135^{o}	180^{o}	225^{o}	270^{o}	315^{o}	Total
Right	297	282	272	284	293	291	288	279	2286
Wrong	3	18	28	16	7	9	12	21	114
Detection rate	**99%**	**94%**	**90.6%**	**95.3%**	**97.6%**	**97%**	**96 %**	**93%**	**95.3 %**

Table 2 Pose recognition rate at different distances.

	Standing	Walking	Sitting	Squatting	Lying
2.5m	94%	96%	92%	92%	100%
3.0m	98%	96%	90%	94%	100%
3.5m	100%	94%	90%	86%	100%
4.0m	90%	94%	92%	94%	100%
4.5m	96%	96%	88%	94%	100%
5.0m	88%	98%	98%	94%	100%
Average recognition rate	**94.33%**	**95.67%**	**91.67%**	**92.33%**	**100%**

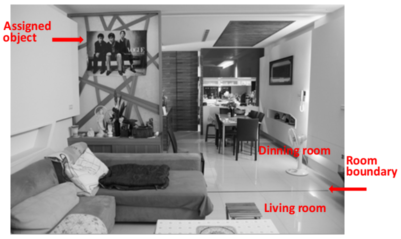

Fig. 9. Test of automatic boundary setting, a line was drew on the ground to be used as the ground truth in the experiment.

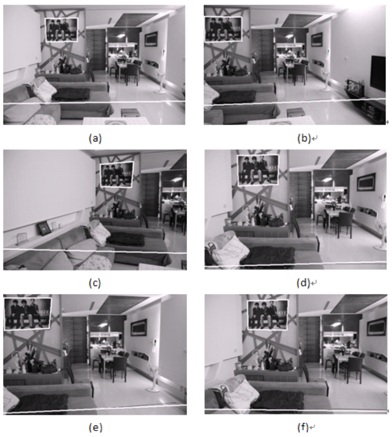

Fig.10. Experiments of automatic boundary setting, (a)3.5m and 0° from the object. (b)3.5m and 10° right from the object. (c)3.5m and 10° left from the object. (d)2.5m and 0° from the object. (e)2.5m and 5° right from the object. (f)2.5m and 5° left from the object.

4.3 Experimental Results of Activity Detection

Fig.11 presents recorded image sequence of human activity recognition in the experiment. In the beginning, the camera was fixed and the system started to find and localize the landmark object in the image plane. We use the relation function to calculate the space boundary as shown in Fig. 11(a). A human was identified and his location tracked through updated space boundary in the image sequence. The activity was recognized as "Go to the dining room" as shown in Fig. 11(b). If the person stays more than 10 frames, then the activity will become "In the dining room" as shown in Fig.11(c). When the person crossed the space boundary, he was recognized as "Go to the living room" as shown in Fig. 11(d). Similarly, if the person stays more than 10 frames, the activity will be recognized as "In the living room" as shown in Fig. 11(e). When the sit posture was recognized, the system will display "Sitting in the living room" as shown in Fig. 11(f).

When the camera moves some distance, the system will detect a new object position, then set a new space boundary in the image plane by using the relation function (8). The new space boundary is shown in Fig. 11(g). We observe that the system uses the new boundary to identify the location of human and human activities as shown in Figs. 11(h)~(l). In particular, if the current pose of human is standing and the next pose of human is lying, and

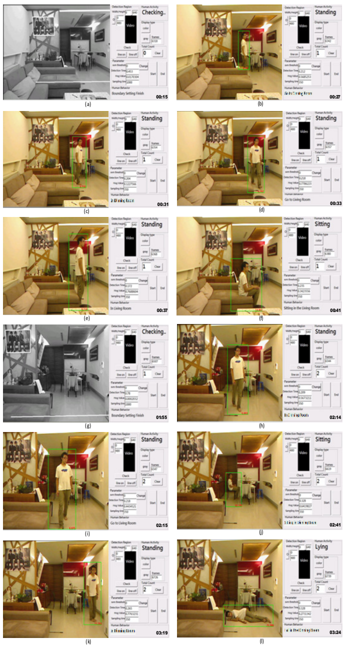

Fig. 11. Sequence of human activity recognition in a home setting, (a)automatic boundary setting. (b)successful detect human and the pose (standing).(c)human in the dining room. (d)human cross boundary line to the living room. (e)human in the living room. (f)human sitting in the living room. (g)camera moved and new boundary was set. (h)human in the dining room (i)human crossed the new boundary line to the living room. (j)human sitting in the dining room. (k)human standing in the dining room. (l)human fell in the dining room.

according to location of the human, the system detects the human "Fall in the dining room" as shown in Fig. 11(k) and Fig. 11(l). A video clip of the experiment can be found in [15].

5. CONCLUSIONS

In this paper, a design of automatic space boundary setting for identifying human activity in a home setting is proposed. In this design, human activity recognition of a mobile camera is achieved by using the relationship between a calibrated landmark object and space boundary in image plane. Experimental results reveal that for a camera motion in a range of 1m and 10°, the space boundary is set accurately for activity recognition. Combining the space boundary, human location, and human body pose recognition, the human activity in a home setting is detected satisfactorily. In the future, we will focus on simplification of human detection and pose recognition algorithms to reduce the computing time. It is interesting to extend the method to multiple calibration objects and thus extend the moving range of the mobile camera and the robot working space. We will also add face recognition to the system, such that the robot can record the activities of multiple persons and interact differently with different members.

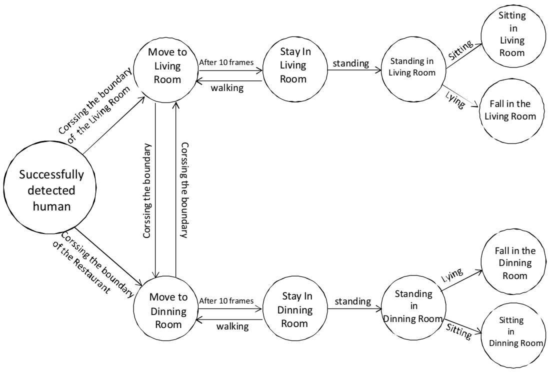

Fig. 12. Finite state machine of human activity.

Acknowledgement

This work was partly supported by grants from Industrial Technology Research Institute, ITRI-A353C40000B1-4 and National Science Council, Taiwan, NSC-99-2221-E-009-144.

REFERENCES

[1] H. Fujiyoshi and A.J. Lipton, "Real-Time Human Motion Analysis by Image Skeletonization," *Proc. IEEE Workshop Applications of Computer Vision*, Princeton, New Jersey, 1998, pp.15-21.

[2] H.S. Chen, H.T. Chen, Y.W. Chen and S.Y. Lee, "Human Action Recognition Using Star Skeleton," *Proc. of the 4th ACM international workshop on Video Surveillance and Sensor Networks,* Santa Barbara, California, USA.

[3] H. Miyamori and S.I. Iisaku, "Video Annotation for Content-Based Retrieval Using Human Behavior Analysis and Domain Knowledge," *Proc. of IEEE International Conference Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp.320-325.

[4] T. Zhao and R. Nevatia, "Tracking Multiple Humans in Complex Situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 26, No.9, 2004.

[5] N. Carter, D. Young and J. Ferryman, "A Combined Bayesian Markovian Approach for Behavior Recognition," *Proc. of IEEE International Conference on Pattern Recognition*, Hong Kong, China, 2006, pp.761-764.

[6] V. Kellokumpu, M. Pietikainen and J. Heikkila, "Human Activity Recognition Using Sequences of Postures," *Proc. IAPR Conference on Machine Vision Applications,* Tsukuba Science City, Japan, 2005, pp.570-573.

[7] L. Wang, W. Hu and T. Tan, "Recent Developments in Human Motion Analysis," *Pattern Recognition*, Vol. 36, No.3, 2003, pp.585-601.

[8] Z. Zhou, X. Chen, X. Han, J. Keller and Z. He, "Activity Analysis, Summarization and Visualization for Eldercare," *IEEE Transactions on Circuits and System for Video Technology*, Vol. 18, No.11, 2008, pp.1489-1498.

[9] M. Shiomi, T. Kanda, D.F. Glas, S.Satake, H. Ishiguro and N. Hagita, "Field Trial of Networked Social Robots in a Shopping Mall," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems,* St. Louis, USA, 2009, pp.2846-2853.

[10] N. Dalal, "Finding People in Images and Videos," *Ph.D. dissertation*, Institut National Polytechnique de Grenoble, Grenoble, France, 2006.

[11] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for Human Detection," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp.886-893.

[12] http://pascal.inrialpes.fr/data/human/

[13] http://cbcl.mit.edu/software-datasets/PedestrianData.html.

[14] S. Benhimane and E. Malis, "Homography-Based 2D Visual Tracking and Servoing," *International Journal of Robotics Research*, 2007, pp.661-676.

[15] http://isci.cn.nctu.edu.tw/video/URAI2011/activity