2011 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 25-30, 2011. San Francisco, CA, USA

Realtime Recognition of Complex Daily Activities Using Dynamic Bayesian Network

Chun Zhu and Weihua Sheng
School of Electrical and Computer Engineering
Oklahoma State University
Stillwater, OK, 74078
email: chunz, weihua.sheng@okstate.edu

***Abstract*— In this paper, we proposed a method to recognize complex human daily activities including body activities and hand gestures simultaneously in an indoor environment. Three wearable motion sensors are attached to the right thigh, the waist, and the right hand of a person, while an optical motion capture system is used to obtain his/her location information. A three-level dynamic Bayesian network is implemented to model the intra-temporal and inter-temporal constraints among the location, body activity and hand gesture. The body activity and hand gesture are estimated using a Bayesian filter and the short-time Viterbi algorithm, which reduces the storage memory and the computational complexity. We conducted experiments in a mock apartment environment and the obtained results showed the effectiveness and accuracy of our algorithms.**

***Index Terms*— Activity recognition, wearable computing.**

I. INTRODUCTION

In recent years, human daily activity recognition is gaining more attentions in many areas, such as health and elderly care, personal fitness, gaming accessories, human-robot interaction, etc. Traditional activity recognition approaches use visual data as input to observe full human body movements. However, there are some challenging issues in vision-based approaches which include data association for multiple human subjects, computational complexity in image processing, and data consistency under different environmental conditions. An alternative source for human activity recognition is motion data collected from wearable motion sensors, which can monitor motions with less data compared to vision-based systems. Since there is ambiguity in activity recognition due to the limited information from the motion data, the capability of activity recognition from wearable motion sensors is greatly dependent on the number and the placement of sensors. On the other hand, it is crucial to build a minimum wearable sensor system because too many wearable sensors on the human body will be obtrusive.

In most current research, inertial sensors are usually used to capture human motion data. Many applications for human activity recognition using inertial sensors can be found in [1], [2], [3]. With the advancement of MEMS, VLSI and wireless communication technologies, wearable inertial sensors have been become compact and wireless. Xsens [4] has developed the MTw module which is a wireless, accurate, small and lightweight 3D motion tracker. Multiple MTw modules can form a wireless body area network to capture human body pose. However, the cost of this device is relatively high. Philips Corporation has developed the NWS wireless activity monitor [5] which tracks the body motion every time when the user moves up, down, forwards, backwards and sideways. By measuring the acceleration of these movements, it calculates how much energy the human subject used to make them.

There are several existing approaches to human daily activity recognition. Most researchers use discriminative approaches for daily activity recognition (e.g. window based feature clustering) [6], [7]. Others apply generative approaches to utilize sequential constraints, such as hidden Markov models (HMM) [8], [9]. Location information is also used to find the context information of human daily activities [10]. However, the spatial and temporal constraints between the location and human daily activities are rarely utilized.

In our previous work, we found that human body activities and locations are highly correlated [11], and we fused data from a single motion sensor and human location information to recognize eight activities. In order to recognize more complex activities such as using a computer, cooking, and reading a book, we attach motion sensors to different parts of the human body to recognize body activities and hand gestures while maintaining the least obtrusiveness to the human subject. In this paper, we proposed an approach that combines motion data and vision-based location information to recognize complex daily activities in realtime. Adaptive gesture spotting is proposed to segment gestures for different environments and body activities. The adaptive gesture spotting method can adjust the parameters for gesture detection in different scenarios. A dynamic Bayesian network is developed to model both the sequential constraints and the causal dependency between the locations and daily activities in order to recognize the body activities and hand gestures simultaneously. The short-time Viterbi algorithm [12] is applied to recover activities with reduced computational complexity and small memory size.

This paper is organized as follows. Section II describes the hardware platform for the proposed complex daily activity recognition. Section III explains the activity and gesture recognition algorithm. The experimental results are provided

978-1-61284-456-5/11/$26.00 ©2011 IEEE

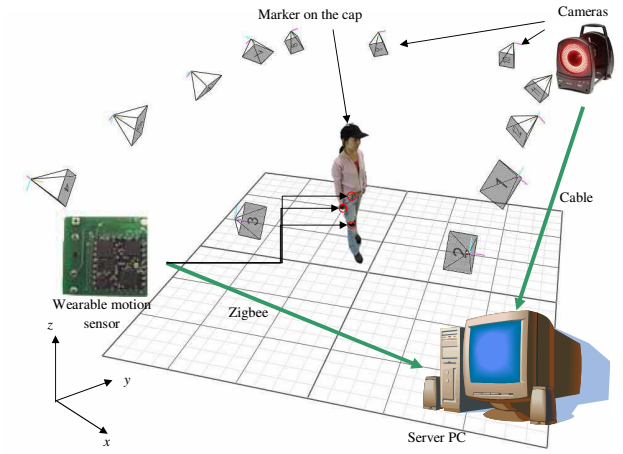

Fig. 1. The overview of the hardware platform for complex daily activity recognition.

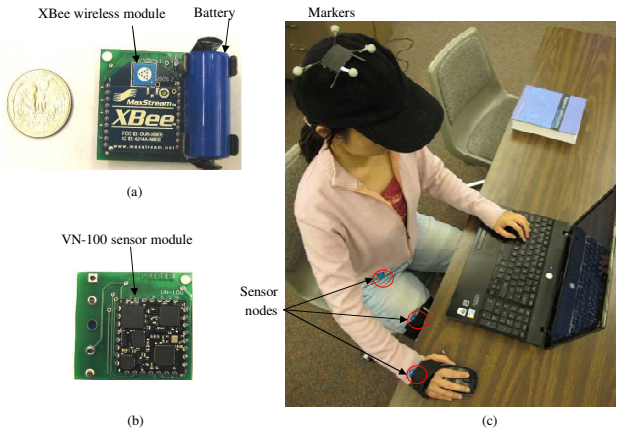

Fig. 2. The wireless motion sensor node: (a) XBee wireless module and battery (front side); (b) VN-100 sensor module (back side); (c)The sensor nodes worn on the human subject.

in Section IV. Conclusions and future work are given in Section V.

II. Hardware Platform

Our proposed hardware system for complex daily activity recognition is shown in Figure 1. We use three wireless motion sensors to collect motion data and transfer it to a server PC. The cameras in the optical motion capture system are used to provide location information of the human subject. The wearable motion sensors are synchronized with the location data from the motion capture system. Thus, this minimum setup of the wearable sensor system is combined with the motion capture system to facilitate human daily activity recognition. The three-sensor setup minimizes the obtrusiveness to the human subject. The optical system provides real-time location coordinates of the human subject rather than raw video data, which significantly reduces the computational complexity compared to traditional vision-based activity recognition algorithms.

A. Hardware Setup for Motion Data Collection

Figure 2 (a) and (b) show the prototype of the motion sensor node we developed based on a commercial VN-100 [13] orientation sensor. The sensor nodes send data (3D acceleration, 3D angular velocity, orientation, and magnetic data) through Zigbee to a receiver on the PC for processing. ZigBee is one type of short-distance wireless communication standards, which is used for wireless networking mainly in home and offices. Each motion sensor node has an ID to be distinguished from others. Therefore, multi-person activity can also be tracked in this system. Due to the limited size of our lab, we tested our approach using Zigbee. Our method can also work in larger space when other communication standards are considered, such as WiFi. A PDA is used to label the activities in the training phase. Since the position to attach the sensor is very important to activity recognition [14], we collected data using the sensors on different parts of the human body and found that the thigh and the waist are the best positions for body activity recognition using the minimum sensor setup. The third sensor is attached to the right hand to capture hand motion, as shown in Figure 2(c).

The wearable motion sensor samples the 3D acceleration and 3D angular velocity at a rate of 20Hz. In the experiments, it is observed that the angular velocity exhibits similar properties as the acceleration, so we only collect the 3D acceleration as the raw data. Features including means and variances are extracted and further clustered into discrete observation symbols for the dynamic Bayesian network.

B. Hardware Setup for Location Tracking

We use the Vicon motion capture system [15] to collect the location information. A baseball cap with four markers is used to track the human subject. The tracking software runs on the server PC to calculate the position of the markers in real-time and stream out the data. The 3D location of the markers can be resolved within millimeter accuracy. The real-time data streaming rate is 100 fps. We down-sample the location data to synchronize it with the motion sensor data. The output coordinate in the 2D (x-y) space gives us the location information of the human subject.

III. Framework for Body Activity and Hand Gesture Recognition

A. Overview

Our recognition program consists of two threads. First, the data sampling thread collects data from three body sensors and the Vicon system. Each data packet includes the ID of the sensor, the 3D acceleration, and the current time in milliseconds. The location data is sampled at the same time. Second, the data processing thread deals with the sampled data in two steps: preprocessing and online recognition of body activities and hand gestures. This process is triggered every second and generates a vector representing the body activity and hand gesture.

In the training mode, the server PC accepts connection from a PDA to provide labels as the ground truth. The label is recorded when the user manually pushes a button on a PDA. We use a digital camera to record the scene for

the ground truth of the locations, body activities and hand gestures.

There are three steps in data preprocessing. First, data from different sensors need to be separated. Three body sensors are configured to stream data at 20Hz. However, the Zigbee receiver on the server PC receives around 60 packets of mixed data from these three sensors. Those packets need to be separated into three groups with respect to the sensor IDs. Second, features are extracted in the one-second window buffer. The mean and variance of the 3D acceleration form the feature vector. Third, feature vectors are discretized into observation symbols for body activity and hand gesture recognition in the dynamic Bayesian network.

B. Hierarchical Activity and Gesture Model

In this paper, we recognize both body activity and hand gestures at the same time. Eight body activities are to be recognized: *sitting*, *standing*, *lying*, *walking*, *sit-to-stand*, *stand-to-sit*, *lie-to-sit*, and *sit-to-lie*. The activities are categorized into two kinds: stationary and motional activities. Five specific types of hand gestures are considered: using mouse, typing on a keyboard, flipping a page while reading a book, stir-fry cooking, and dining using a spoon. Undefined gestures are categorized into the type of "other hand movements".

In indoor environments, human daily activities (body activities and hand gestures) and locations are highly correlated [11]. Given a floor plan of an apartment, we can learn the probability distribution for each specific activity on the 2D map. Such a probability distribution can be obtained through training. To simplify the activity-location correlation, the given map of the mock apartment is segmented into different areas with corresponding probabilities of body activities and hand gestures. The coordinate of the human subject given by the Vicon system is mapped into N_A semantic areas. Similarly, there are correlations between body activities and hand gestures, which can be learned from training.

In the time domain, the transition of the location of a person follows certain patterns. For example, people always walk from one area to another adjacent area and there is probability distribution according to the floor plan and personal preference. We assume the transition of locations is a discrete, first-order Markov process. Meanwhile, there are constraints between two consecutive body activities and hand gestures as well. For example, at this second the person is sitting at the computer and typing on the keyboard. It is not likely he/she will be walking in the following second without standing up. In a similar way, we assume the transition of body activity and hand gesture is also a discrete, first-order Markov process.

Since a person's location, body activity and hand gesture have both intra-temporal causal relationship and inter-temporal constraints, this pattern can be modeled using a three-level dynamic Bayesian network model shown in Figure 3. The individual nodes in this graphical model represent hidden states and shaded nodes represent observations. The solid arcs correspond to direct probabilistic

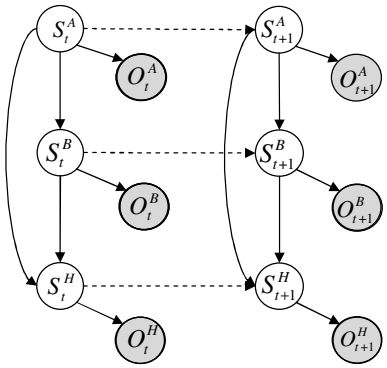

Fig. 3. Two-slice dynamic Bayesian network of the activity and gesture model, showing dependencies between the observed and hidden variables. Observed variables are shaded. Intra-temporal causal links are solid, inter-temporal links are dashed.

interactions between nodes in one time slice, while the dashed arcs correspond to temporal dependencies between two time slices t and $t+1$.

The highest level of the model represents the person's location information. S^A represents the location area. The middle level represents the person's body activity S^B and the lowest level represents his/her hand gesture S^H. In the data preprocessing step, the observed measurements from the Vicon system are clustered into the observation O^A. The data from the sensors on the right thigh and the waist are combined and clustered into the observation O^B. The right hand sensor measurements are clustered into the observation O^H.

In our model, the dependencies between the nodes in Figure 3 include both spatial and temporal domains. The observation O_t^A, O_t^B, and O_t^H depend on corresponding intra-temporal hidden state S_t^A, S_t^B, and S_t^H, respectively. The hand gesture S^H at time t depends on the previous gesture, the body activity and the location at the current time slice. The body activity S^B at time t depends on the previous activity and the location state at time t. The location state S^A only depends on its previous state.

C. Adaptive Gesture Spotting

In our system, hand gestures are first spotted from other non-gesture movements. Since hand gestures exhibit different intensity levels under different activities, the parameters for gesture spotting have to adapt to the change of environments and body activities. For example, when a person is working on the computer with the keyboard, the hand movement intensity is much less than that during cooking. Therefore, the classifiers need to be trained under different locations and body activity conditions.

The observation of body activity O^B is obtained by classifying the feature vectors from the sensors on the thigh and the waist. The detailed classification method is similar to the coarse-grained classification in [16]. The coordinates of the human subject given by the Vicon system are mapped into N_A semantic areas, which corresponds the location observation O^A in N_A distinct values.

The observation of hand gesture O^H is obtained by classifying the feature vectors from the sensors on the hand adaptive to the corresponding O^B and O^A. First, the feature vectors of the hand sensor are grouped based on O^B and O^A. Let $F^H_{(a,b,t)}$ be the feature vector at time t, when $O^A = a$ and $O^B = b$. $F^H_{(a,b)}$ stands for all the feature vectors in the training data set, when $O^A = a$ and $O^B = b$. K-means clustering is applied on $F^H_{(a,b)}$ to obtain the centroids $C_{(a,b)} = \{C_1, C_2, ..., C_i, ...C_K\}$, where C_i is the centroid for each cluster.

$$C_{(a,b)} = f_{K-means}(F^H_{(a,b)}, K) \tag{1}$$

where $f_{K-means}$ is the function for K-means classifier. K is the number of clusters in K-means clustering.

In the testing phase, the Euclidean distance between each feature vector of hand sensor $F^H_{(a,b,t)}$ and the centroids of cluster $C_1, C_2, ...C_K$ is calculated and the index of C_i, which has the minimum distance is chosen as the output of hand observation O^H_t.

$$O^H_t = \arg\min_i \left\| F^H_{(a,b,t)} - C_{(a,b)} \right\| \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm.

Since the centroid $C_{(a,b)}$ is trained on different location and body activity conditions, the feature vectors of hand sensor can be clustered adaptively to spot meaningful hand gestures.

D. Short-time Viterbi Algorithm for Online Smoothing

The standard Viterbi algorithm [17] retrieves the state sequence, which maximizes the belief value. That is, the retrieved state sequence has the maximum likelihood given the observation sequence from time 1 to t. In the standard Viterbi algorithm, finding the maximum likelihood state sequence is done by tracing back through a matrix of back-pointers starting from the end of the sequence. The belief value needs to be calculated from the beginning of the sequence. The computational complexity of the standard Viterbi algorithm is $O(T \times N^2)$, where T is the length of the sequence and N is the size of the state space. The memory storage size is $T \times N^2$. However, this approach is unsuitable for realtime implementation. The short-time Viterbi algorithm can solve this problem and enhance the efficiency [12]. The computational complexity of short-time Viterbi algorithm at each time step is $O(N^2)$, and the memory storage size is $2 \times N^2$. Therefore, the computational complexity and memory storage size are reduced compared with the standard Viterbi algorithm.

IV. Experimental Results

A. Environment Setup

We performed the experiments in a mock apartment, which has a dimension of 3 × 5 square meters as shown in Figure 4 (a). The Vicon system is installed on the wall. To represent the activity-location correlation, the given map of the mock apartment is segmented into different areas with corresponding probabilities of activity, as shown in Figure 4 (b). We use uniform distributions in each area.

The sensor setup is shown in Figure 2, regular daily complex activities were performed. We collected 5 sets of training data and 15 sets of testing data. The training datasets include motion data about 10 minutes and each of the testing datasets has a duration of about 6 minutes.

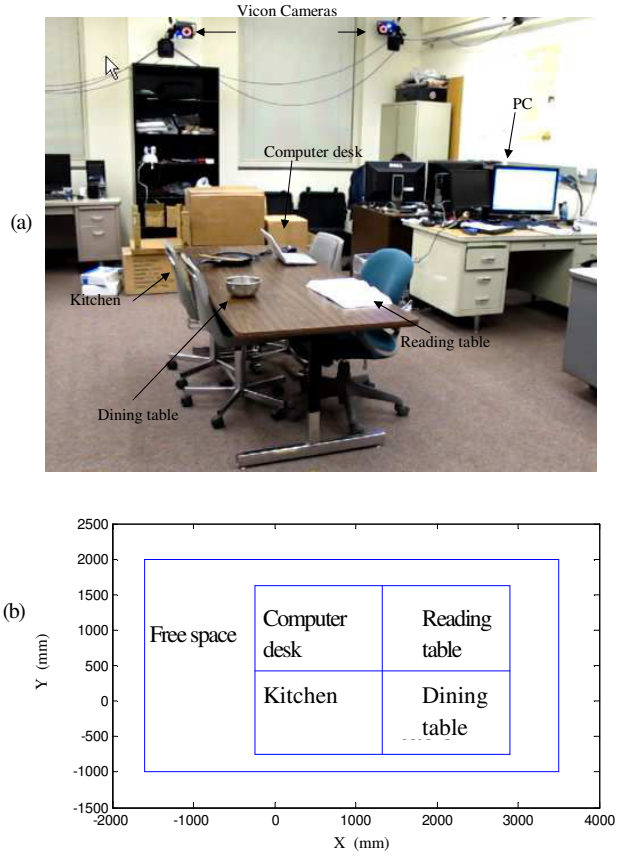

Fig. 4. (a) the setup of the mock apartment; (b) the layout of the mock apartment.

B. Recognition Result

In the experiment the online activity recognition results are compared with the ground truth recorded from a regular video recorder. Some significant frames are shown in Figure 5. In each subfigure, the plots in the top rows represent the observation symbol output of location O^A, body activity O^B, and hand gesture O^H. The plots in the bottom row show the results of body activity S_B and hand gesture S_H from the short-time Viterbi algorithm. The map and the movement trajectory of the human subject are shown in the middle plot in each subfigure. In Figure 5(a), the human subject goes to the computer desk, sits down and starts to type on the keyboard. The body activity indicates walking, and sitting. In Figure 5(b), she walks to the reading table and pulls out the chair. The body activity shows sit-to-stand and walking. The hand gesture shows *other gestures*. In Figure 5(c), she sits beside the reading table and flips pages several times. The body activity shows walking, and sitting. The hand gesture

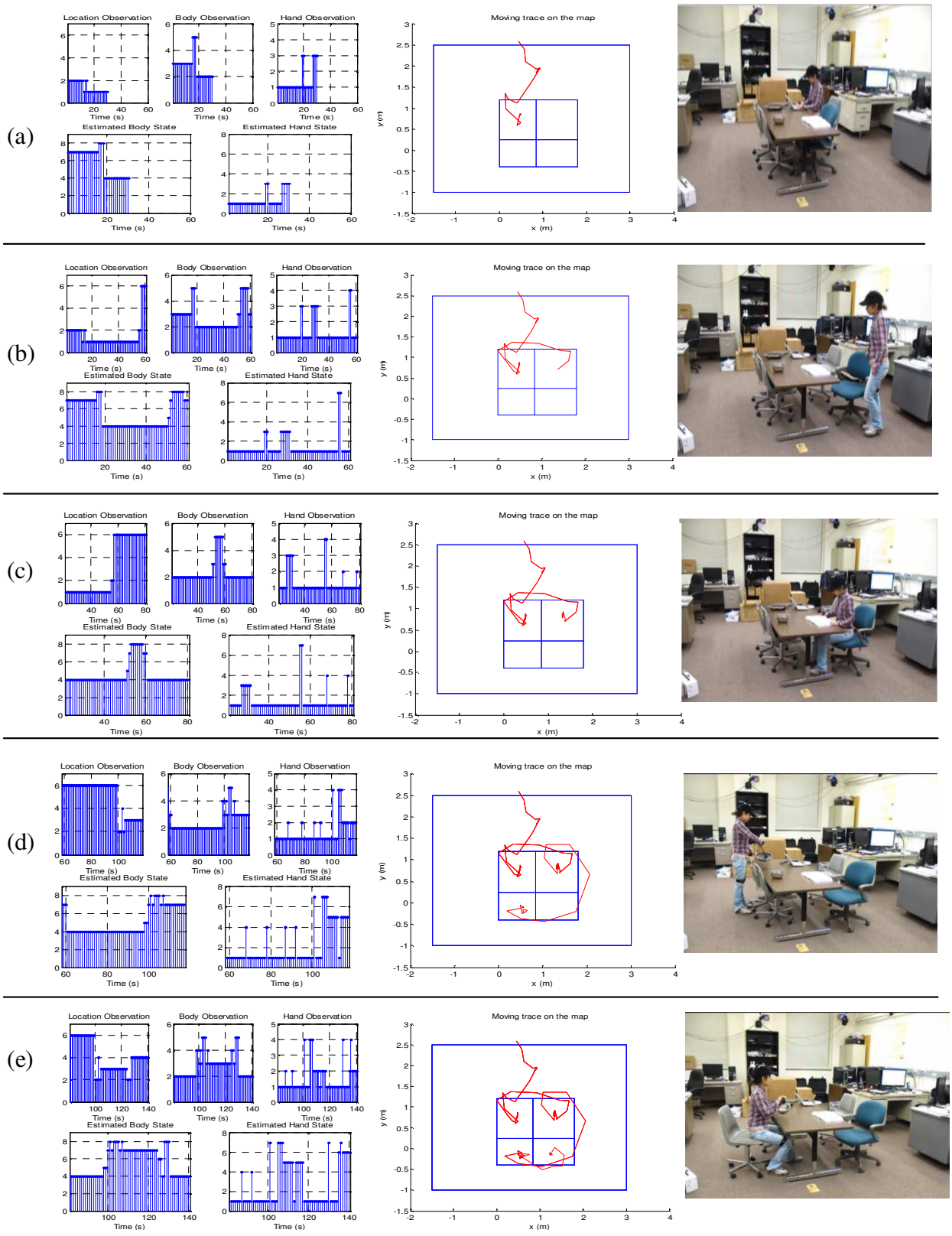

Fig. 5. Results captured from video and server PC. The top block shows the raw sensor data and the layout of the apartment on the right. Labels for activity result: 1) lying, 2) lie-to-sit, 3) sit-to-lie, 4) sitting, 5) sit-to-stand, 6) stand-to-sit, 7) standing, 8) walking. Labels for gesture result: 1) non-gesture, 2) using a mouse, 3) typing on a keyboard, 4) flipping a page, 5) stir-frying, 6) eating, 7) other hand movements.

Ground truth	Decision type											Accuracy
	Sitting	Sit-to-stand	Stand-to-sit	Standing	Walking	Typing on keyboard	Using the mouse	Flipping a page	cooking	Eating	Missed	
Sitting	**1.00**	--	--	--	--	--	--	--	--	--	--	1.00
Sit-to-stand	--	**0.92**	--	--	0.08	--	--	--	--	--	--	0.92
Stand-to-sit	--	--	**0.90**	--	0.06	--	--	--	--	--	0.04	0.90
Standing	--	--	--	**1.00**	--	--	--	--	--	--	--	1.00
Walking	--	--	0.02	--	**0.98**	--	--	--	--	--	--	0.98
Typing on keyboard	--	--	--	--	--	**0.83**	0.08	--	--	--	0.09	0.83
Using the mouse	--	--	--	--	--	0.05	**0.76**	--	--	--	0.19	0.76
Flipping a page	--	--	--	--	--	--	--	**0.85**	--	--	0.15	0.82
cooking	--	--	--	--	--	--	--	--	**0.82**	--	0.18	0.82
Eating	--	--	--	--	--	--	--	--	--	**0.80**	0.20	0.80

TABLE I

THE ACCURACY OF THE DYNAMIC BAYESIAN NETWORK FOR BODY ACTIVITY AND HAND GESTURE RECOGNITION.

shows *flipping a page*. In Figure 5(d), she stands in the kitchen and the hand gesture is *stir-frying*. In Figure 5(d), she sits at the dining table and the hand gesture is *eating*.

The accuracy in terms of the percentage of correct decisions is listed in Table I. The values in bold are the percentages of the correct classifications corresponding to the specific types of activities. Other numbers indicate the percentages of wrong classifications. The overall accuracy of our approach is above 85%, which is higher compared to some recent existing human daily activity recognition methods [7], [18].

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a method to recognize human complex daily activities which consist of body activities and hand gestures simultaneously in an indoor apartment environment. A three-level dynamic Bayesian network is implemented to model the intra-temporal and inter-temporal constraints among the location, body activities and hand gestures. The body activity and hand gesture are estimated online using the short-time Viterbi algorithm. Our approach has the advantage of reducing the obtrusiveness and the complexity of vision processing, while maintaining high accuracy of activity recognition. We conducted experiments in a mock apartment environment and the accuracy of the real-time recognition is evaluated. In the future, we will combine the location and human activities for simultaneous tracking and activity recognition (STAR) [19], which will remove the need of the Vicon motion capture system.

ACKNOWLEDGMENTS

This project is partially supported by the NSF grant CISE/CNS 0916864 and CISE/CNS MRI 0923238.

REFERENCES

[1] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bula, and P. Robert. Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly. *IEEE Trans on Biomedical Engineering*, 50:711–723, 2003.

[2] K. Aminian, Ph. Robert, E. E. Buchser, B. Rutschmann, D. Hayoz, and M. Depairon. Physical activity monitoring based on accelerometry: validation and comparison with video observation. *Medical and Biological Engineering and Computing*, 3:304–308, 1999.

[3] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. *PERVASIVE 2004*, pages 1–17, 2004.

[4] Xsens InC. http://www.xsens.com/, 2011.

[5] Philips InC. http://www.directlife.philips.com/, 2011.

[6] Jie Yang, Shuangquan Wang, Ningjiang Chen, Xin Chen, and Pengfei Shi. Wearable accelerometer based extendable activity recognition system. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3641–3647, May 2010.

[7] T. Huynh, U. Blanke, and B. Schiele. Scalable Recognition of Daily Activities with Wearable Sensors. *Location- and Context-Awareness*, pages 55–67, 2007.

[8] L. R. Rabiner. A tutorial on hidden markov models and selected application in speech recognition. In *Proc. IEEE*, volume 77, pages 267–296, 1989.

[9] H. Junkera, O. Amft, P. Lukowicz, and G. Troster. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition*, pages 2010–2024, 2008.

[10] A. Raj, A. Subramanya, D. Fox, and J. Bilmes. Rao-blackwellized particle filters for recognizing activities and spatial context from wearable sensors. *Experimental Robotics*, pages 211–221, 2008.

[11] C. Zhu and W. Sheng. Motion- and location-based online human daily activity recognition. *Pervasive and Mobile Computing*, In Press, 2010.

[12] J. Bloit and X. Rodet. Short-time viterbi for online hmm decoding: Evaluation on a real-time phone recognition task. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2121–2124, 2008.

[13] VectorNav Technologies. http://www.vectornav.com/, 2011.

[14] U. Maurer, A. Smailagic, D.P.Siewiorek, and M. Deisher. Activity recognition and monitoring using multiple sensors on different body positions. In *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks*, pages 113–116, 2006.

[15] Vicon Motion Systems. http://www.vicon.com/, 2011.

[16] C. Zhu and W. Sheng. Human daily activity recognition in robot-assisted living using multi-sensor fusion. In *IEEE International Conference on Robotics and Automation*, pages 2154–2159, 2009.

[17] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory*, 13:260–269, 1967.

[18] Xi Long, Bin Yin, and R. M. Aarts. Single-accelerometer-based daily physical activity classification. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6107–6110. IEEE, September 2009.

[19] D. Wilson and C. Atkeson. Simultaneous tracking & activity recognition (star) using many anonymous, binary sensors. *Proceedings of PERVASIVE*, pages 62–79, 2005.