Realtime Recognition of Complex Human Daily Activities Using Human Motion and Location Data

Chun Zhu and Weihua Sheng*

***Abstract*—Daily activity recognition is very useful in robot-assisted living systems. In this paper, we proposed a method to recognize complex human daily activities which consist of simultaneous body activities and hand gestures in an indoor environment. A wireless power-aware motion sensor node is developed which consists of a commercial orientation sensor, a wireless communication module, and a power management unit. To recognize complex daily activities, three motion sensor nodes are attached to the right thigh, the waist, and the right hand of a human subject, while an optical motion capture system is used to obtain his/her location information. A three-level dynamic Bayesian network (DBN) is implemented to model the intratemporal and intertemporal constraints among the location, body activity, and hand gesture. The body activity and hand gesture are estimated using a Bayesian filter and a short-time Viterbi algorithm, which reduces the computational complexity and memory usage. We conducted experiments in a mock apartment environment and the obtained results showed the effectiveness and accuracy of our method.**

***Index Terms*—Activity recognition, body sensor network, dynamic Bayesian network (DBN), wearable computing.**

I. Introduction

A. Motivation

With the growth of the elderly population, more seniors live alone as sole occupants of a private dwelling than any other population groups. Helping them to live a better life is very important and has great societal benefits. Many researchers are working on new technologies, such as assistive robots, to help elderly people [1], [2]. Human–robot interaction (HRI) is an important design problem in assistive robots. It is desirable for the robots to be able to understand human gestures and their daily activities so that they can better serve humans. Automated recognition of human gestures and activities can also be used in studying behavior related diseases, detecting abnormal behaviors, activity logging, etc.

Manuscript received November 13, 2011; revised January 16, 2012; accepted February 8, 2012. Date of publication March 12, 2012; date of current version August 16, 2012. This work was supported in part by the National Science Foundation under Grant CISE/CNS 0916864 and Grant CISE/CNS MRI 0923238. *Asterisk indicates corresponding author*.

C. Zhu is with Microsoft Corporation, San Francisco, CA 94107 USA (e-mail: chuzhu@microsoft.com).

*W. Sheng is with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: weihua.sheng@okstate.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2012.2190602

Traditional activity recognition approaches use cameras to obtain the data of full human body movement [3]. However, there are some challenging issues in vision-based approaches, such as data association for multiple human subjects, computational complexity in image processing, and data consistency under different environmental conditions. These issues sometimes make the practical use of vision-based recognition very difficult. An alternative source for human activity recognition is motion data collected from wearable motion sensors. Motion sensors worn on human body or integrated into human clothing can collect motion data with much less volume compared to those from vision-based systems. It is crucial to build a minimum wearable sensor system because too many wearable sensors on the human body may be obtrusive to the human subject. However, the limited information from the motion data usually causes ambiguity. Meanwhile, the accuracy of activity recognition from wearable sensors relies on the number of motion sensors and where they are located on the human body. Therefore, it is an important and challenging task to reduce the number of wearable sensors while improving the activity recognition accuracy.

From our daily experiences, we learned that human body activities and locations are highly correlated. For example, the "*sitting*" activity most likely occurs at a chair or sofa, while the "*lying*" activity most likely occurs in a bed. Therefore, it is possible to integrate the limited amount of motion data and the location information to enable accurate activity recognition. There are various ways to provide human location information in an indoor environment, for example, through radio-frequency identification (RFID) tags, camera tracking, or other techniques. In our previous work [4], we fused motion data from a single wearable sensor and location information to recognize eight basic body activities. However, a single motion sensor is not sufficient to recognize complex daily activities, such as using a computer, cooking, or reading a book, which involve simultaneous body and hand movement, as well as the associated environments. Therefore, we need attach motion sensors to the human body and hand to recognize body activities and hand gestures at the same time. In this paper, we propose an approach that combines motion data and location information to recognize complex daily activities in realtime. This paper has the following contributions.

1) A new wireless motion sensor node is developed, which is compact, power-aware, and has reconfigurable data output.
2) Adaptive gesture spotting is developed to detect gestures conditioned on environmental context and body activities.
3) A dynamic Bayesian network (DBN) [5] is used to model both the sequential constraints and the causal dependence

between the locations and daily activities, while a short-time Viterbi algorithm [6] is applied to recover activities with reduced computational complexity and memory use.

This paper is organized as follows. The rest of Section I introduces the related work. Section II describes the hardware platform for complex daily activity recognition. Section III presents the framework for body activity and hand gesture recognition. Section IV presents the detailed implementation of activity recognition based on a DBN. The experimental results are provided in Section V. Conclusions and future work are given in Section VI.

B. Related Work

This paper focuses on activity recognition using motion sensors; therefore, the review of related work consists of two parts: development of wearable motion sensors and complex activity recognition through wearable motion sensors.

1) Motion Sensors: With the advancement of microelectromechanical systems, very large scale integration, and wireless communication technologies, wearable motion sensors have become compact and wireless. The motion sensor MDP-A3U7 is a sensor unit which combines a ceramic gyro, acceleration sensor, and terrestrial magnetism sensor. It can detect the 3-D posture in realtime [7]. But the output data of this sensor are delivered only via wired universal serial bus (USB) interface. The inertia link [8] provided by MicroStrain, Inc., combines a triaxial accelerometer, triaxial gyro, temperature sensors, and an on-board processor running a sophisticated sensor fusion algorithm. The communication interface can be wireless, USB, and RS232. The supply voltage ranges from 4.5 to 16 V and the current is about 90 mA. Xsens Technologies offers several kinds of orientation trackers [9] which have power consumption of about 540 mW. MEMSense, Inc., provides a wireless inertial measurement unit [10] which has a power consumption of about 900 mW and a 2.5-h battery life. Several 3-D motion sensor nodes have also been developed in the research community. Bandala and Joyce developed a wireless inertial sensor for tumor motion tracking [11]. A real-time algorithm determined the six degree-of-freedom sensor posture, consisting of three components of position and three components of rotational orientation. Acht *et al*. in Philips research center developed a miniature wireless inertial sensor for measuring human motion [12]. The sensor measures 3-D acceleration, 3-D earth magnetic field, and 3-D angular speed. The angular accuracy of the calibrated sensor is below 3°. The aforementioned two sensor nodes realize 3-D motion data collection and transmission, but do not fully address the power saving issue. In [12], some attention has been paid to power management, but the potential to prolong the lifetime of the battery is very limited.

2) Complex Activity Recognition From Wearable Sensors: There are mainly three methods for complex daily activity recognition using wearable sensors: discriminative methods, generative methods, and hierarchical methods. Most researchers use discriminative methods for complex daily activity recognition (e.g., window-based feature clustering). For example, Yang *et al*. [13] used one sensor node attached to the front of the testee' right leg (near the ankle) to detect standing, walking, running, climbing up stairs, and climbing down stairs at certain locations using three multiclass classifiers: decision tree algorithm [14], K-nearest neighbor algorithm [15] and weighted support vector machines algorithm [16]. Others apply generative methods to utilize sequential constraints, such as hidden Markov models (HMMs). Huynh *et al*. [17] used three sensors on the thigh, the waist, and the wrist to recognize daily activities. They combined clustering-based methods and HMM with an accuracy between 11% and 90% on different activities. Raj *et al*. [18] collected GPS data in the outdoor environment and fused it with the measurement from a wearable sensor board. They considered the location information as another parallel data channel in the Bayesian network of activity recognition. However, they could not get the detailed indoor location information and did not consider hand gesture-related daily activities. Very few researchers model a hierarchy of activities that considers complex activities as high-level semantics when compared to simpler low-level body activities from sensor measurements. The constraints in complex activity sequences can be modeled by a hierarchical HMM, which is similar to the grammars in speech recognition. Some researchers call it a hierarchy of sensory grammars [19].

In summary, complex daily activity recognition using wearable sensors is still an emerging research area while majority of the human activity recognition researches use computer vision-based approaches. In this paper, we will use the motion data from wearable motion sensors and the location information to recognize complex daily activities.

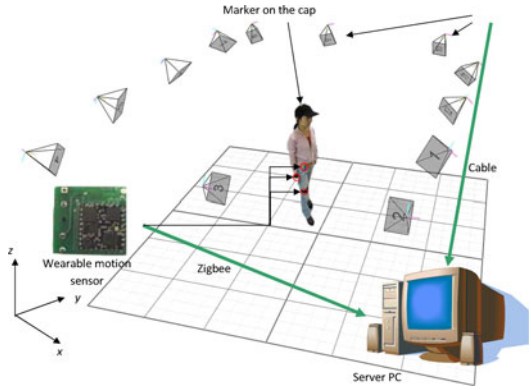

Fig. 1. Overview of the hardware platform for complex daily activity recognition.

II. Hardware Platform

Our proposed hardware system for complex daily activity recognition is shown in Fig. 1. We use three motion sensors to collect motion data and transfer them 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, the minimum setup of the wearable sensor system is combined with the motion capture system to facilitate human complex daily activity recognition. The

Fig. 2. Wireless motion sensor nodes worn on the human subject.

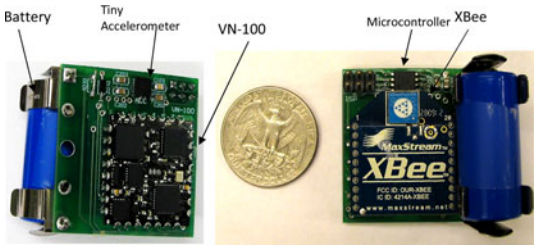

Fig. 3. Wearable motion sensor node.

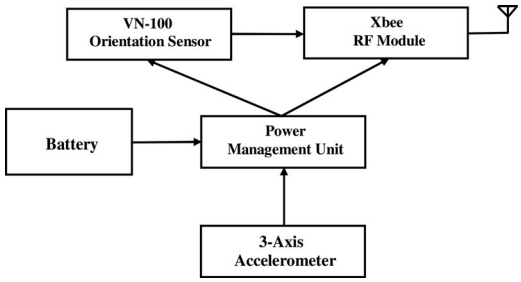

Fig. 4. Block diagram of the wearable motion sensor node.

TABLE I
COMPARISON OF MOTION SENSORS

Motion Sensor	Voltage	Power Consumption
Xsens MTx (Bluetooth mode)	4.5 - 12 V	540mW
MEMSense Bluetooth IMU	6.0-9.0V	600mW
Inertia-Link 802.15.4	4.5-16 V	405-1440mW
Our wireless motion sensor (Active)	3.3V	396mW

three-sensor setup minimizes the obtrusiveness to the human subject. The optical system provides real-time location of the human subject. In reality, the location can be obtained through RFID or other localization methods.

Since the position to attach the sensor is very important to activity recognition [20], 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. The third sensor is attached to the right hand to capture hand motion, as shown in Fig. 2. The wireless motion sensor samples the 3-D acceleration and 3-D angular rate at a rate of 20 Hz. In the experiments, it is observed that the angular rate exhibits similar properties as the acceleration, so we only collect the 3-D acceleration as the raw data, which is represented as $D_t = [D_t^T, D_t^W, D_t^H]$ where D_t^T, D_t^W, and D_t^H indicate the 3-D acceleration from the sensor on the thigh, the waist, and the hand, respectively. Features (mean and variance) are extracted from the raw data and further clustered into discrete observation symbols.

A. Hardware Setup for Motion Data Collection

1) Hardware Design: The motion sensor node we developed consists of a VN-100 orientation sensor module [21] from VectorNav, Inc., for motion sensing, an XBee RF module [22] for wireless communication, and a power management unit. The power is provided by a 3.3 V 2/3 AA battery. The picture of the motion sensor node is shown in Fig. 3 and its functional block diagram is shown in Fig. 4. The motion data include 3-D orientation (roll, pitch, yaw), acceleration, angular rate, and magnetic field. The dimension of the sensor node is 36 mm $\times$ 35 mm $\times$ 18 mm and the weight is about 40 g. The VN-100 module calculates the orientation based on a 3-D accelerometer, a 3-D gyro, and a 3-D magnetometer. With the surface-mount package, it is possible to embed this sensor into various products. The typical operating voltage range is from 3.1 to 5.5 V, and the power supply current is 65 mA. These features make the VN-100 module ideal for incorporating accurate and reliable device orientation information in the compact embedded electronic designs. With all the ICs in normal operation mode, the sensor node can operate continuously for about 5 h. Table I shows the comparison of several motion sensors on the market in terms of voltage level and power consumption. It can be seen that our motion sensor node has an active power consumption of 396 mW, which is lower than that of most of the sensors on the market.

2) Power Management: For wearable sensors, how to reduce power consumption so as to prolong battery life is a critical issue. When the wearable sensor is used to monitor daily activity of the elderly, it is inconvenient to replace or recharge the battery frequently. Therefore, an embedded power management unit which employs a power management algorithm is proposed to reduce the power consumption of the wireless motion sensor node. The task of the power management unit is to analyze the 3-D acceleration from a tiny accelerometer and determine if the sensor node is in motion or not. If the sensor node is in stationary state, when the data variations remain below certain threshold, the VN-100 orientation sensor module and the XBee RF module can be turned into sleep mode, or disabled. Otherwise, these two modules will be woken up or enabled. With a duty cycle around 38%, we can estimate that the battery life of the motion sensor node can be prolonged from 5 to 14 h, which is sufficient for many wearable computing applications. Table II

TABLE II
COMPARISON OF THREE MODES OF THE VN-100 SENSOR

Sensor Mode	Current	Power Consumption	Power Duration
Normal	120mA	396mW	5 h
With power management unit (The duty cycle of power performance is around 38%)	--	--	14 h
Sleep	0.8mA	2.64mW	750h

shows the comparison of the power consumption between the normal and sleep mode of the VN-100, which clearly indicates that by turning the VN-100 node into sleep mode, significant power can be saved.

B. Hardware Setup for Location Tracking

We use the Vicon motion capture system [23] to collect the location information of a human subject, which in other approaches can be obtained from cameras, RFID, etc. 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 realtime and stream out the data. The 3-D location of the markers can be resolved within millimeter accuracy. The real-time data streaming rate is 100 frames/s. We downsample the location data at 20 Hz to synchronize it with the motion sensor data. The output coordinate in the 2-D space gives us the location information of the human subject.

III. FRAMEWORK FOR BODY ACTIVITY AND HAND GESTURE RECOGNITION

A. Overview

The flowchart of our recognition software is shown in Fig. 5. The PC runs the recognition program which consists of two threads. First, the data sampling thread collects data from the three motion sensors and the Vicon motion capture system. Each data packet includes the ID of the sensor, the 3-D acceleration, and the current time in milliseconds. The location data are sampled in the meanwhile. 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 personal digital assistant (PDA) to provide labels as the ground truth. The label is recorded when the user manually pushes a button on a PDA. In the real-time testing mode, we use a digital camera to record the scene for the ground truth of the locations, body activities, and hand gestures. The three motion sensors are configured to stream data at 20 Hz. Features such as the mean and variance of the 3-D acceleration are extracted from the raw data and discretized into observation symbols for body activity and hand gesture recognition in the dynamic Bayesian network described in the following.

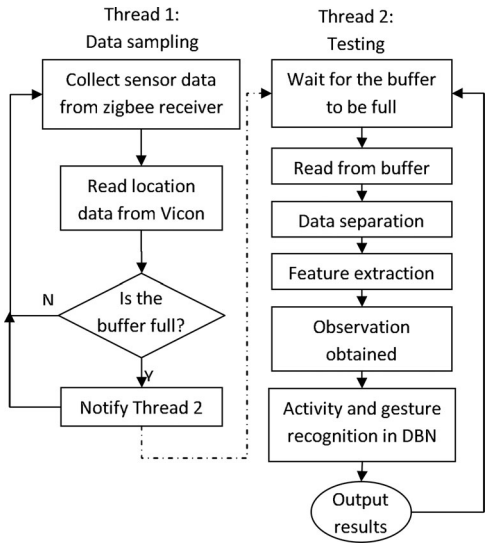

Fig. 5. Flowchart of the recognition software.

B. Hierarchical Activity and Gesture Model

In this paper, eight body activities are to be recognized: *sitting*, *standing*, *lying*, *walking*, *sit-to-stand*, *stand-to-sit*, *lie-to-sit*, and *sit-to-lie*, which are categorized into two types: 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 [4]. Given a floor plan of an apartment, we can learn the probability distribution for each specific activity on the 2-D map through training. To simplify the representation of 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. There are also correlations between body activities and hand gestures, which can also be learned from training.

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 a 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, if at one moment the person is sitting at the computer desk and typing on the keyboard, it is not likely that he/she will be walking in the following moment 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.

A person's location, body activity, and hand gesture have both intratemporal causal relationship and intertemporal constraints, which can be modeled using a three-level DBN model shown

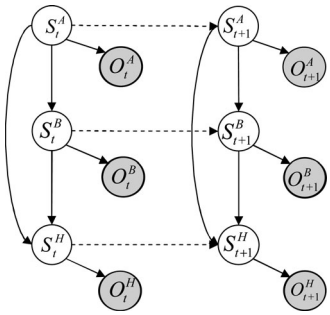

Fig. 6. Two-slice DBN of the activity and gesture model, showing dependences between the observed and hidden variables.

in Fig. 6. The individual nodes in this graphical model represent hidden states and shaded nodes represent observations. The solid arcs correspond to causal dependences between nodes in one time slice, while the dashed arcs correspond to the temporal dependences between two time slices t and $t+1$. The highest level of the model represents the person's location S^A. 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.

C. Coarse-Grained Classification for Body Observation

In the DBN, the observation of body activity O^B is obtained by classifying the feature vectors from the sensors on the thigh and the waist. Four neural networks are applied in the coarse-grained classification as shown in Fig. 7. For each network, the input is an *n*-by-1 feature vector extracted from the sensor raw data, which represents *n* features. The functions of layer 1 and 2 are log-sigmoid functions and layer 3 uses the hard limit function. The first and the second layers form a two-layer feed-forward network and the weights and biases are trained through the back-propagation method. The third layer outputs the discrete value of 0 or 1.

Features (mean and variance) are extracted from the raw data to form four input vectors for neural networks N_1, N_2, N_3, and N_4

$$\begin{aligned} I_1 &= M^T = [\text{mean}(D_x^T), \text{mean}(D_y^T), \text{mean}(D_z^T)] \\ I_2 &= V^T = [\text{var}(D_x^T), \text{var}(D_y^T), \text{var}(D_z^T)] \\ I_3 &= M^W = [\text{mean}(D_x^W), \text{mean}(D_y^W), \text{mean}(D_z^W)] \\ I_4 &= V^W = [\text{var}(D_x^W), \text{var}(D_y^W), \text{var}(D_z^W)]. \end{aligned} \quad (1)$$

Among these four neural networks, N_2 and N_4 are used to detect the motion state of the waist and the thigh with 0 for stationary and 1 for motion. N_1 and N_3 are used to detect the stationary state of the waist and the thigh with 0 for horizontal and 1 for vertical. Using the rules in Table III, the neural network outputs can be fused to generate the body observation symbol O^B, which takes value from 1 to 5. The coordinates of the human subject given by the Vicon motion capture system are mapped into one of N_A semantic areas, which corresponds to the location observation O^A in N_A distinct values.

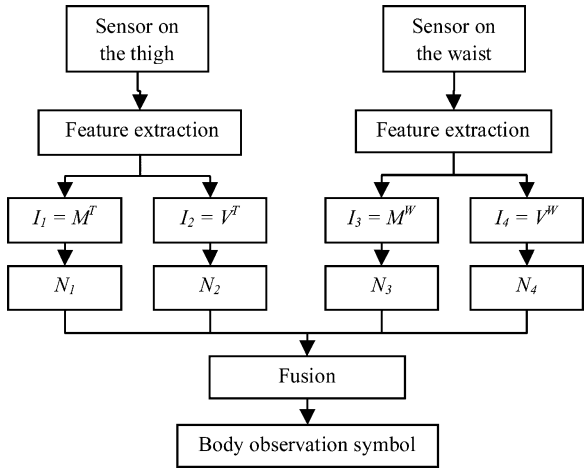

Fig. 7. Neural network-based coarse-grained classification.

TABLE III
FUSION RULES FOR NEURAL NETWORKS

Fusion rules			Sensor on the waist		
			$N_3 = 0$		$N_3 = 1$
			$N_4 = 0$	$N_4 = 1$	
Sensor on the thigh	$N_1 = 0$	$N_2 = 0$	1	2	5
		$N_2 = 1$	5	3	
	$N_1 = 1$		5		4

D. Adaptive Gesture Spotting

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

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 motion data 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 dataset, when $O^A = a$ and $O^B = b$. *K*-means clustering is applied on $F^H_{(a,b)}$ to obtain the centroid set $C_{(a,b)} = \{C_1, C_2, \ldots, C_i, \ldots, C_K\} = f_{K\text{-means}}(F^H_{(a,b)}, K)$ where $f_{K\text{-means}}$ 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 motion data $F^H_{(a,b,t)}$ and the centroids of cluster $C_1, C_2, \ldots, C_K$ are calculated and the index of C_i, which has the minimum distance, is chosen as the output of hand observation O^H_t

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

where $\|\cdot\|$ is the Euclidean norm. Since the centroid set $C_{(a,b)}$ is trained on different location and body activity conditions, the feature vectors of hand motion data can be clustered adaptively to spot meaningful hand gestures.

IV. Implementation of the DBN

A. Mathematical Representations

In the three-level DBN model, the superscript of states and observations represents the level: area (top), body (middle), and hand (bottom), while the subscript represents the time index. Each level has three basic elements.

1) State Transition Probability Distribution: The state transition probability distribution in each level reflects the intratemporal dependence in Fig. 6.

The top-level location transition probability represents the topology of the layout and the personal preference of the transition in the environment

$$a^A_{i,j} = P(S^A_{t+1} = j | S^A_t = i). \tag{3}$$

The middle-level body activity transition probability depends on the location

$$a^B_{i,j,p} = P(S^B_{t+1} = j | S^B_t = i, S^A_{t+1} = p). \tag{4}$$

The bottom-level hand gesture transition probability depends on the location and the body activity

$$a^H_{i,j,p,q} = P(S^H_{t+1} = j | S^H_t = i, S^B_{t+1} = q, S^A_{t+1} = p). \tag{5}$$

2) Observation Symbol Probability Distribution: Since the observed variables only depend on the corresponding states in the same level, the observation symbol probability distribution can be expressed as

$$b^A_{i,j} = P(O^A_t = j | S^A_t = i) \tag{6}$$

$$b^B_{i,j} = P(O^B_t = j | S^B_t = i) \tag{7}$$

$$b^H_{i,j} = P(O^H_t = j | S^H_t = i). \tag{8}$$

3) Initial State Distribution: Since the intratemporal dependence exists from the beginning of the sequence, the initial state distribution also follows the relationship of the links between levels in Fig. 6

$$\pi^A_i = P(S^A_1 = i) \tag{9}$$

$$\pi^B_{j,i} = P(S^B_1 = j | S^A_1 = i) \tag{10}$$

$$\pi^H_{k,j,i} = P(S^H_1 = k | S^B_1 = j, S^A_1 = i). \tag{11}$$

Based on the DBN model, we have the joint probability of the sequence as

$$\begin{aligned} &P(S^A_{1:t}, S^B_{1:t}, S^H_{1:t}, O^A_{1:t}, O^B_{1:t}, O^H_{1:t}) \\ &= P(S^A_1) \prod_{t=2}^{T} P(S^A_t | S^A_{t-1}) \prod_{t=1}^{T} P(O^A_t | S^A_t) P(S^B_1 | S^A_1) \\ &\quad \prod_{t=2}^{T} P(S^B_t | S^B_{t-1}, S^A_t) \prod_{t=1}^{T} P(O^B_t | S^B_t) P(S^H_1 | S^B_1, S^A_1) \\ &\quad \prod_{t=2}^{T} P(S^H_t | S^H_{t-1}, S^B_t, S^A_t) \prod_{t=1}^{T} P(O^H_t | S^H_t) \end{aligned} \tag{12}$$

where T is the length of the observation sequence.

Due to the computational complexity, this general formula cannot be used for real-time processing directly. Therefore, the Viterbi algorithm is applied to estimate the probability recursively.

B. Short-Time Viterbi Algorithm for Online Smoothing

The standard Viterbi algorithm [6] retrieves the state sequence, which maximizes the belief value. 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 backpointers q^*_T starting from the end of the sequence. The belief $\delta_t(i,j,k)$ and the index variable $\psi_t(i,j,k)$ need to be calculated from the beginning of the sequence. The computational complexity of the standard Viterbi algorithm is $O(T \times |Q|^2)$, where T is the length of the sequence and Q is the size of the state space. The memory storage size is $T \times |Q|^2$. However, this approach is unsuitable in the case of real-time input and output. The short-time Viterbi algorithm can solve this problem and enhance the efficiency [24]. The computational complexity of short-time Viterbi algorithm at each time step is $O(|Q|^2)$, and the memory storage size is $L \times |Q|^2$, where $L \geq 3$ is the length of the sequence. Therefore, the computational complexity and memory storage size are reduced compared with the standard Viterbi algorithm.

The short-time Viterbi algorithm has three steps: initialization, recursion for Bayesian filtering, and path smoothing.

1) *Initialization:*

$$\begin{aligned} \delta_1(i,j,k) = &P(S^A_1 = i) P(O^A_1 | S^A_1 = i) \\ &P(S^B_1 = j | S^A_1 = i) P(O^B_1 | S^B_1 = j) \end{aligned} \tag{13}$$

Algorithm 1 Short-time Viterbi for smoothing in DBN

Initial Viterbi sequence length $L = 3$, δ_1, and ψ_1 using Eq (13), (14);
for each new observation O_t **do**
  obtain $\delta_t(i,j,k)$ and $\psi_t(i,j,k)$ using Eq (15), (16);
  obtain current state estimate q_T^* using current $\delta_t(i,j,k)$ using Eq (17);
  backward one step and calculate the path (previous state estimate) using Eq (18);
  correct previous state output if q_{t-1}^* changes;
  save current $\delta_t(i,j,k)$ for next loop.
end for

$$P(S_1^H = k|S_1^B = j, S_1^A = i)P(O_1^H|S_1^H = k)$$
$$\psi_1(i,j,k) = [0,0,0]. \quad (14)$$

2) *Recursion:*

$$\begin{aligned}&\delta_t(i,j,k)\\ &= \max_{p,q,r}(\delta_{t-1}(p,q,r)a^A b^A_{pi}\delta_{t-1}a^B b^B_{qj}\delta_{t-1}a^H b^H_{rk})\\ &= \max_{p,q,r}[\delta^H_{t-1}(p,q,r)P(S_t^A = i|S_{t-1}^A = p)b^A_{pi}\\ &\quad P(S_t^B = j|S_{t-1}^B = q, S_t^A = i)b^B_{qj}\\ &\quad P(S_t^H = k|S_{t-1}^H = r, S_t^B = j, S_t^A = i)b^H_{rk}] \quad (15)\end{aligned}$$

$$\psi_t(i,j,k) = \arg\max_{p,q,r}\delta_t(i,j,k) \quad (16)$$

$$q_t^* = \arg\max_{i,j,k}\delta_t(i,j,k). \quad (17)$$

3) *Path Smoothing:*

$$q_{t-1}^* = \psi_t(q_t^*). \quad (18)$$

The pseudocode for short-time Viterbi algorithm is shown in Algorithm 1.

V. Experimental Results

A. Environmental Setup

We performed the experiments in a mock apartment, which has a dimension of 3 m× 5 m as shown in Fig. 8. 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. To simplify the calculation, we use uniform distributions for different activities in each area.

The sensor setup is shown in Fig. 2; regular daily activities were performed: *standing, sitting, sleeping*, and *transitional activities*. We collected eight sets of training data and 20 sets of testing data. Each testing dataset had a duration of about 6 min. We recorded video as the ground truth to evaluate the recognition results.

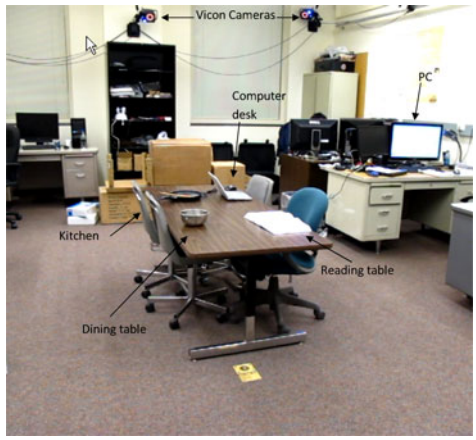

Fig. 8. Setup of the mock apartment.

B. Recognition Result

In the experiment, each output decision value represents the decision for a 1-s time window. The accuracy is calculated based on the individual decision made for each sliding window. The complex activity recognition result is compared with the ground truth recorded on the video and labeled by an observer. The recorded video of the experiment is synchronized with the output of the activity recognition. The video clips of the experiments are available at the link [25]. Some significant frames are shown in Fig. 9. In each subfigure, the plots in the top row 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 body activity S_B and hand gesture S_H from the short-time Viterbi algorithm. The map and the moving trace of the human subject are shown in the middle plot in each subfigure. In Fig. 9(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 Fig. 9(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 Fig. 9(c), she sits beside the reading table and flips pages several times. The body activity shows *walking* and *sitting*. The hand gesture shows *flipping a page*. In Fig. 9(d), she stands in the kitchen and the hand gesture is *stir-frying*. In Fig. 9(e), 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 IV. 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 [17], [26].

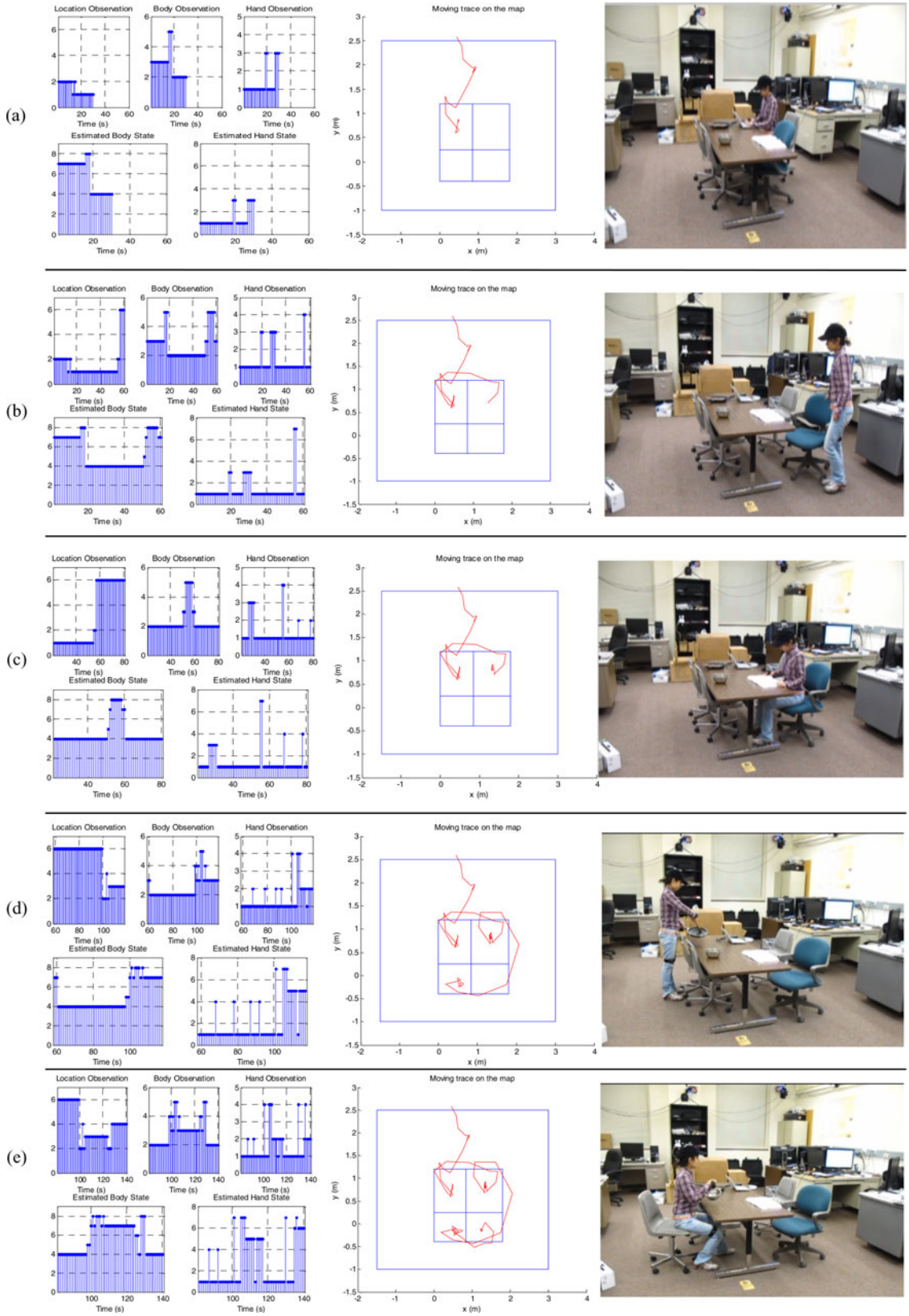

Fig. 9. (a)–(e) Results captured from video and the server PC. The top block shows the raw sensor data and the layout of the apartment on the right. The second, third, and fourth blocks show the observation symbol output of O^A, O^B, O^H and results of S^B, S^H, respectively, with the picture of that scenario on the right. Labels for body activity result: 1) lying, 2) lie-to-sit, 3) sit-to-lie, 4) sitting, 5) sit-to-stand, 6) stand-to-sit, 7) standing, and 8) walking. Labels for gesture result: 1) nongesture, 2) using a mouse, 3) typing on a keyboard, 4) flipping a page, 5) stir-frying, 6) eating, and 7) other hand movements.

TABLE IV
RECOGNITION ACCURACY BY DBN

Ground Truth	Decision Type: Sitting	Sit-to-sand	stand-to-sit	Standing	Walking	Typing on keyboard	Using the mouse	Flipping a page	Cooking	Eating	Missed	Accuracy
Sitting	**1.00**	-	-	-	-	-	-	-	-	-	-	1.00
Sit-to-sand	-	**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-keyboard	-	-	-	-	-	**0.83**	0.08	-	-	-	0.09	0.83
Using mouse	-	-	-	-	-	0.05	**0.76**	-	-	-	0.19	0.76
Flipping a page	-	-	-	-	-	-	-	**0.85**	-	-	0.15	0.85
Cooking	-	-	-	-	-	-	-	-	**0.82**	-	0.18	0.82
Eating	-	-	-	-	-	-	-	-	-	**0.80**	0.20	0.80

VI. 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. Three wireless motion sensors are worn by the human subject to provide motion data, while an optical motion capture system is used to obtain his/her location. A three-level DBN is implemented to model the intratemporal and intertemporal constraints among the location information, body activities, and hand gestures. 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. It is worth pointing out that our daily activity recognition algorithm can be extended to detect abnormal behaviors such as falling down on the floor, lying on the bed for an exceedingly long time, etc. In our future work, we will also combine the location and human activities for simultaneous tracking and activity recognition (STAR) [27], which will remove the need for the Vicon motion capture system.

REFERENCES

[1] Z. Khalila and M. Merhia, "Effects of aging on neurogenic vasodilator responses evoked by transcutaneous electrical nerve stimulation," *J. Gerontol. Ser.*, vol. 55, pp. B257–B263, Jun. 2000.

[2] W. C. Mann, *Smart Technology for Aging, Disability, and Independence*. New York: Wiley, 2005.

[3] T. B. Moeslunda, A. Hiltonb, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 90–126, Nov. 2006.

[4] "Motion- and location-based online human daily activity recognition," *Pervasive Mobile Comput.*, vol. 7, no. 2, pp. 256–259, Apr. 2011.

[5] Z. Ghahraman, "Learning dynamic Bayesian networks," in *Adaptive Processing of Sequences and Data Structures*. New York: Springer-Verlag, 1998, pp. 168–197.

[6] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[7] MDP-A3U7. (2011). [Online]. Available: http://www.nec-tokinamerica.com

[8] Inertial link. (2011). [Online]. Available: http://www.microstrain.com/inertia-link.aspx

[9] Xsens Inc. (2011). [Online]. Available: http://www.xsens.com

[10] MEMSense, LLC. (2011). [Online]. Available: http://www.memsense.com

[11] M. Bandala and M. Joyce, "Wireless inertial sensor for tumor motion tracking," *J. Phys.*, vol. 76, no. 1, pp. 024006-1–024006-9, 2007.

[12] V. Acht, E. Bongers, N. Lambert, and R. Verberne, "Miniature wireless inertial sensor for measuring human motions," in *Proc. IEEE Eng. Med. Biol. Sci. Conf.*, Lyon, France, Aug. 23–26, 2007, pp. 6279–6282.

[13] J. Yang, S. Wang, N. Chen, X. Chen, and P. Shi, "Wearable accelerometer based extendable activity recognition system," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 3641–3647.

[14] T. Mitchell, "Decision tree learning," in *Machine Learning*. New York: McGraw-Hill, 1997, pp. 52–78.

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[16] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Disc.*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[17] T. Huynh, U. Blanke, and B. Schiele, "Scalable recognition of daily activities with wearable sensors," in *Proc. Location Context-Awareness*, 2007, pp. 55–67.

[18] A. Raj, A. Subramanya, D. Fox, and J. Bilmes, "Rao-blackwellized particle filters for recognizing activities and spatial context from wearable sensors," *Exp. Robot.*, pp. 211–221, 2008.

[19] A. Bamis, D. Lymberopoulos, T. Teixeira, and A. Savvides, "The behaviorscope framework for enabling ambient assisted living," *Int. J. Pers. Ubiquit. Comput.*, vol. 14, no. 6, pp. 473–487, Sep. 2010.

[20] U. Maurer, A. Smailagic, and D. P. Siewiorek, M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *Proc. Int. Workshop Wearable Implantable Body Sens. Netw.*, 2006, pp. 113–116.

[21] VectorNav Technologies, (2011). [Online]. Available: http://www.vectornav.com.

[22] Digi International Inc. (2011). [Online]. Available: http://www.digi.com.

[23] Vicon Motion Systems. (2011). [Online]. Available: http://www.vicon.com.

[24] J. Bloit and X. Rodet, "Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 2121–2124.

[25] C. Zhu. (2011). Youtube video on complex daily activity recognition [Online] Available: http://youtu.be/98l4KHpCmZE.

[26] X. Long, B. Yin, and R. M. Aarts, "Single-accelerometer-based daily physical activity classification," in *Proc. 2009 Annu. Int. Conf. IEEE Eng. Med. Bio. Soc.*, Sep., 2009, pp. 6107–6110.

[27] D. Wilson and C. Atkeson, "Simultaneous tracking & activity recognition (star) using many anonymous, binary sensors," in *Proc. 3rd Int. Conf. Pervasive Comput.*, 2005, pp. 62–79.

Authors' photographs and biographies not available at the time of publication.