Journal of Information Processing Vol.24 No.6 853–866 (Nov. 2016)

[DOI: 10.2197/ipsjjip.24.853]

Regular Paper

Recognizing and Understanding Nursing Activities for a Whole Day with a Big Dataset

Sozo Inoue[1,a)] Naonori Ueda[2,b)] Yasunobu Nohara[3,c)] Naoki Nakashima[3,d)]

Received: December 20, 2015, Accepted: July 5, 2016

Abstract: In this paper, we 1) provide a real nursing data set for mobile activity recognition that can be used for supervised machine learning, 2) provide big data combined with the patient medical records and sensors attempted for 2 years, and also 3) propose a method for recognizing activities for a whole day utilizing prior knowledge about the activity segments in a day. Furthermore, we demonstrate data mining by applying our method to the bigger data with additional hospital data. In the proposed method, we 1) convert a set of segment timestamps into a prior probability of the activity segment by exploiting the concept of importance sampling, 2) obtain the likelihood of traditional recognition methods for each local time window within the segment range, and, 3) apply Bayesian estimation by marginalizing the conditional probability of estimating the activities for the segment samples. By evaluating with the dataset, the proposed method outperformed the traditional method without using the prior knowledge by 25.81% at maximum by a balanced classification rate, and outperformed by 6.5% the F-measure with accepting 1 hour of margin. Moreover, the proposed method significantly reduces duration errors of activity segments from 324.2 seconds of the traditional method to 74.6 seconds at maximum. We also demonstrate the data mining by applying our method to bigger data in a hospital.

Keywords: mobile activity recognition, nursing activity, bigdata mining

1. Introduction

In the field of healthcare, the standardization of care processes, termed *Clinical* or *Critical pathways*, has been attempted [14], [26], [34], [44], [47], [49]. In meeting such an objective, the recognition and data mining of nursing activities can lead to a better understanding and improvements in medical care, and they can help prevent unnecessary activities and excessive work. At the same time, these approaches are beneficial to patients because the overall care process is optimized, thus resulting in shorter hospitalization times and lower costs.

Recently, researchers have explored the possibility of human activity recognition with mobile sensors; for example, accelerometers, gyroscopes, and low-frequency audio have been explored [2], [5], [10], [29], [31], [33], [35], [36], [46], [52], [58]. In addition, several researchers have applied such technology to domain-specific applications in nursing activities [40], [43], [53]. However, in the available methods, several issues still remain:

The nature of the real activities is not clear

In the application of nursing activity recognition, the *activity classes* — the types of activities — are defined in a domain-specific manner (as listed in Table 1). Here, the activities are not always easy to recognize because the table includes feature value varieties even for single classes, such as blood pressure measurements starting by attaching the corresponding equipment to a patient, followed by pushing air pumps periodically, and finishing with detaching the equipment. Moreover, such activities have imbalance varieties, such as the number of occurrences among classes, starting times in a day, and duration. For example, complex activities, such as capturing X-ray, require dozens of minutes, whereas other activities are completed more quickly. Because the traditional approach normally assumes that activity classes have similar probabilities of being performed, similar probabilities any time in a day, and similar durations, the way in which the accuracy changes when we consider such imbalances is not known.

The goal is not clear

In the application of nursing activity analysis, we can set up clear goals, such as improving nursing activities effectively for timing, duration, and patient satisfaction, or optimizing the costs of the nursing process. For such goals, the technical objective is not only improving the recognition accuracy each time, derived from the traditional recognition from the current time window or those in the vicinity (called *local time windows*), but also estimating the *segment* — the range where the activity is performed continuously — attached with correct timestamps and durations. Thus, by clarifying the goals of the application, we could choose the recognition aspects to which to assign importance, but this is not the case with the existing work.

No dataset with clear goals

To overcome the aforementioned challenges, we require real

[1] Kyushu Institute of Technology, Kitakyushu 804–8550, Japan
[2] NTT Communication Science Laboratories, Sagara-gun, Kyoto 619–0237, Japan
[3] Kyushu University Hospital, Fukuoka 812–8582, Japan
[a)] sozo@mns.kyutech.ac.jp
[b)] ueda.naonori@lab.ntt.co.jp
[c)] y-nohara@info.med.kyushu-u.ac.jp
[d)] nnaoki@info.med.kyushu-u.ac.jp

data to evaluate or input into a machine learning algorithm. However, there is extreme shortage of such open datasets obtained from multiple subjects, and a set of entire days with densely annotated labels. In the literature, there are several datasets, such as Refs. [4], [7], [45] that provide data with longer times, but because they are not inteded for a clear application, it is not clear what accuracy aspects to pursue.

For this paper, we collected 1) (*labeled data*) actual activities from nurses wearing accelerometers in a hospital for approximately 2 weeks and combined them with training labels, which resulted in 25 activity classes with 5,743 labels from 22 nurses, and 2) (*unlabeled data*) the open big data for 60 nurses for 442 [days × people] in the trial for almost 2-years with the duty days which could obtain agreements from the nurses and up to 100 patients, combined with patients' wearable, vital, and environmental sensor data and medical records. From the obtained labeled data, we observed that the activities have imbalances in the number of occurrences for each activity class, the starting times in a day, and the duration of each activity class, as explained in Section 2.

Then, we propose a method for recognizing whole day activities using prior knowledge on the information of a sequence of activity segments which are obtained from a whole day training dataset, such as the daily timestamps, duration, and imbalances among activity classes, as explained in Section 3, based on our papers [22], [23].

In the proposed method, we 1) convert the set of timestamps of the training data into the prior probability of the activity segment by exploiting the concept of importance sampling, 2) obtain the likelihood for the test data with a traditional recognition method for each local time window within the range of the segments, and 3) apply Bayesian estimation by marginalizing the conditional probability of estimating the activities for the segment samples.

By evaluating with the nursing dataset in Section 4, the proposed method outperformed the naive method without using prior knowledge by 25.81% at maximum through the balanced classification rate, and outperformed by 6.5% using the F-measure with accepting 1 hour of margin. Moreover, the proposed method significantly reduces the duration of errors of activity segments from 324.2 seconds of the naive method to 74.6 seconds in k-NN, from 173.5 seconds to 90.33 seconds in NaiveBayes, and from 122.2 seconds to 7.88 seconds in RandomForest.

In order to demonstrate research probabilities with ubiquitous healthcare research to the community, we introduce an analysis of the unlabeled data utilizing the machine-learning result of the labeled data, combined with nurses' profiles and medical records, and applying RandomForest algorithm to generate regression models with considering generalizing ability, and to investigate importances of each predictor variables as well as avoiding interactions between predictor variables, and visualize the effects between predictor and response variables.

The contribution of our paper is four-fold: 1) provide the real dataset [*1] of nursing activities that can be used for supervised machine learning, and also big data combined with patient medical records and sensors, 2) propose a method for utilizing prior knowledge on activity segments in a day, 3) evaluate the proposed method for improvements on the accuracy of activity recognition and the durations of activity segments, and 4) demonstrate data mining by applying our method to bigger data in a hospital merged with additional hospital data.

2. Sensor Data Collection for Nursing Activities

We collected mobile-sensor data from the nurses of a hospital's cardiovascular center [41]. The experiment was first examined and agreed by the ethical committee of the hospital, and exclusive to those nurses who agreed to usage of the sensor data, and to the duties related to patients who consented to participate in the experiment.

It includes labeled data for 2 weeks, and unlabeled data for the duty days which we could obtain agreements from up to 100 patients in 2 years. In this section, we describe the protocols for data collection and review both of the labeled and the unlabeled datasets.

2.1 Protocol

We requested the nurses to wear mobile devices (iPod touches) that record accelerations in their breast pockets in a generally fixed direction. They also attached a small accelerometer device on their right wrist, and another on the back of their waist. **Figure 1** illustrates the attachments. Each sensor measured accelerations on three axes in the range of $\pm 2G$ at 20 Hz.

2.1.1 Labeled Data Collection

The daytime duties of 22 selected nurses over the period over two weeks on Feb. 2014 were labeled with mobile tablets by other nurses who acted as observers. Before the trial, we defined 41 activity classes from the clinical path, and asked the observers to record them. The activity classes were extracted from the clinical path, and the terms in clinical path has consistency with other medical information standard such as HL7 [*2] and NANDA [20].

Naturally, the quality assessment of the labeled data with sensors is not straightforward, because even if we visualize the sen-

Fig. 1 Nurses with three accelerometers: one on their right wrist, one attached to their breast pocket, and one on the back hip.

[*1] http://nurseact.sozolab.jp

[*2] http://www.hl7.org/

sor data, we cannot discriminate one activity class from another. That means, annotating labels for real activities requires careful design. In real nursing activities, nursing the patient has the highest priority, and there are occurences of a lot of missing labels or incorrect timestamps. Therefore, another nurse acted as an observer, and operated another iPod touch device to record the activities of the subject nurse. On the software on the iPod touch, the observer selects the activity class which the subject nurse is about to start, and pushes the finish icon when the subject finishes.

In reality, if the observer waits for the subject nurse to start the activity, the start timestamp will have a latency than the correct one. Therefore, they collaborated with each other to have correct start timestamps, such that the subject nurse declares the activity to the observer before s/he will start it. Moreover, in reality, a nurse could perform several activities concurrently, and the proposed method in this paper assumes such concurrency. However, for the data collection, we gave priority to assure the accuracy of a single activity in that case, because we found that attempting to annotate concurrent activities using the iPod touch tool makes the annotation inaccurate, such that the finishing times are not completed.

2.1.2 Unlabeled Data Collection

In the same department of the hospital as above, we collected unlabeled sensor data for 2 years from the nurses who wear three accelerometers in the same way as the labeled data collection.

Since we also collected the patients' sensor and medical data associated with the nurses' mobile sensor data, — which are out of the main scope of this paper — we specifically collected the nurses' sensor data for the duty days which could obtain agreements. The data we used are collected carefully to be able to be open data, by obtaining agreements from the subject nurses and the patients.

2.1.3 Formatting the Dataset

To interoperate the data sets for labeled and unlabeled data, they were formatted uniformly as well as possible. The ID for the nurses are consistent, then an ID for a nurse is the same for both data sets.

Moreover, while each sensor on each position on the body stores their sensor data separately on the device, it is useful for data analysis to be merged into one multi-column table. Therefore, we joined the data for 3 devices' data of a duty date to a single table in an off-line manner. We first generated timestamps increasing by 20 Hz, which means 0.05 seconds, and adopted the closest sample within 0.025 seconds for each timestamp. If there are no samples within 0.025 seconds, we reused the last timestamp value.

Since each device has its own clock and they have no interaction for time synchronization with each other, there is a risk that the clock is not synchronized. To avoid this, we shook the devices together periodically — once in a day on average — as a reference timestamp, and used the relative time from the shaking time as well as possible.

2.2 Overview of the Dataset

As the result of the experiment, we collected 346.5 [hours × people] of sensor data from 22 nurses by the labeled data collection, and 1,655 [day × people] from 60 nurses by the unlabeled data collection.

To review the collected labels, we review the labels obtained by the labeled data collection in the following.

After the trial, the activity classes actually observed were 25, listed in **Table 1**. The total number of labels was 5,743. The labels for each activity class are also listed in Table 1.

Figure 2 shows the plot of the start times for each activity in a day range. **Figure 3** shows the duration of each activity class.

As shown in Fig. 2, the number of activities varies among activity classes. Moreover, we can see that the activities do not all

Table 1 Observed activity classes and numbers of labels.

No.	Activity class	# labels
1	Anamnese (patient sitting)	2
4	Measure height	45
5	Measure weight (dorsal)	8
8	Measure blood pressure (dorsal)	529
10	Sample blood (dorsal)	16
12	Start intravenous injection	61
13	Finish intravenous injection	40
15	Change drip/line	38
18	Assist doctor	19
19	Find artery	257
20	Examine edema (lie on back)	118
22	Check bedsore (sacrum/back heel)	10
23	Measure ECG	22
24	Attach ECG	54
25	Remove ECG	5
27	Attach bust bandage	29
28	Portable X-ray (prone)	5
29	Changebandage	30
31	Change posture	77
32	Clean body	27
36	Assist wheelchair	86
37	Assist walk	35
38	Move bed	19
39	Wash hands	117
41	Record work (PC)	912

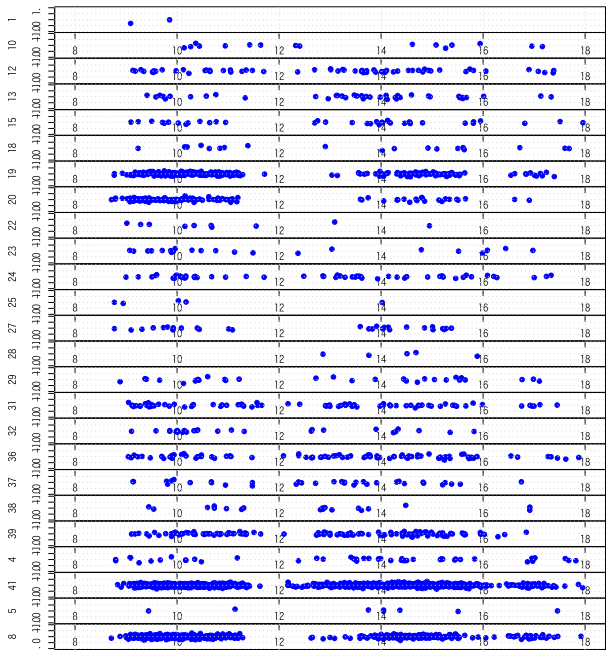


Fig. 2 Start time for each activity in a day range. Each row corresponds to an activity class (the number corresponds to the No. in Table 1), and the x-axis is the hour in a day. The dots are the recorded starting time of an activity. We can see imbalances between activity classes and times in a day.

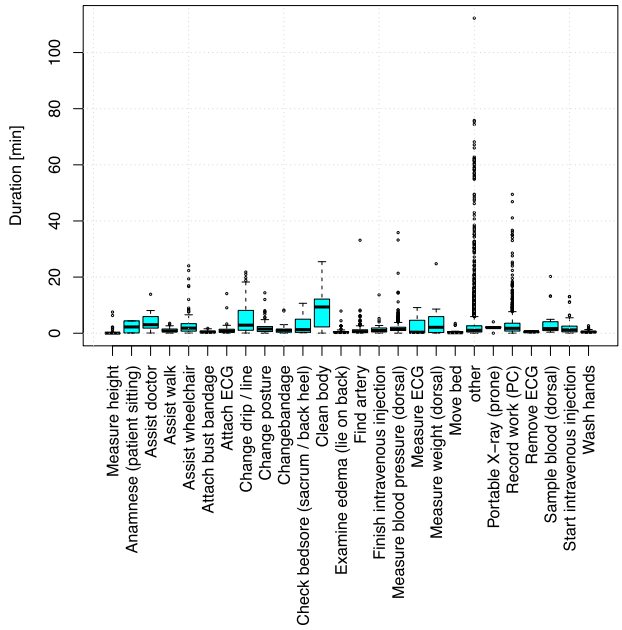


Fig. 3 Durations of each activity label in the dataset.

occur at any time uniformly. Some activities, such as No. 27, occur only during several hours in the morning or afternoon, and others occur continuously, such as No. 12. Compared with traditional experiment settings where the training data are collected in a balanced way or in a short time without considering the time of day, this may result in difficulties during activity recognition.

Moreover, as shown in Fig. 3, the activity duration also varies considerably. For example, the maximum median duration in the dataset we collected was 9.35 minutes for "clean body," whereas the minimum was 0.03 minutes for "measure height." The variances within a class are large, such that "measure weight" has a standard deviation of 8.40 minutes, and "other" has 8.09. These phenomena are considered more significant than other research fields, such as segmentation in voice recognition [8], [37], [38], [57], and chunking in natural language processing [6], [13].

In summary, the real activity dataset attempted for several entire days has imbalances in several aspects, such as class-wise, times of day, and activity duration. If such information is obtained in the training phase, we can expect it to be instructive for improving the activity recognition.

3. Activity Recognition for a Whole Day

In this section, we propose a method for recognizing activities of a whole day.

3.1 Approach

As shown in Fig. 2, nursing activities have different possibilities, depending on the time of day. If we have training data with labels and timestamps, we can convert the set of timestamps into the prior probability of the activity being performed. In addition, if we use both the starting and ending times of an activity, we can obtain information on the activity's duration. As explained in Section 2, such information of when and how long nursing activities are performed is important for analysis. In our approach, in addition to the traditional method for estimating activities from the sensor input of neighborhood time windows, we exploit the timestamp information in order to construct a prior probabilistic distribution on the activities of an entire day, implement them based on importance sampling, and utilize them for the Bayesian estimation of activities.

Table 2 Basic expressions used in the paper.

Symbol	Summary
C	The set of activity classes to be recognized.
$1:T := (1, 2, \cdots, T)$	The time sequence in a day.
x_t	The feature vector at time t ($t \in 1:T$).
a_t^c	Whether the activity at time t is c or not ($c \in C$).
L^c	The number of segments for activity $c \in C$.
$s_l^c := (b(l), e(l))$	The l'th segment ($l \in 1:L^c$).
$b(l) \in 1:T$	The start time of the l'th segment.
$e(l) \in 1:T$	The end time of the l'th segment.

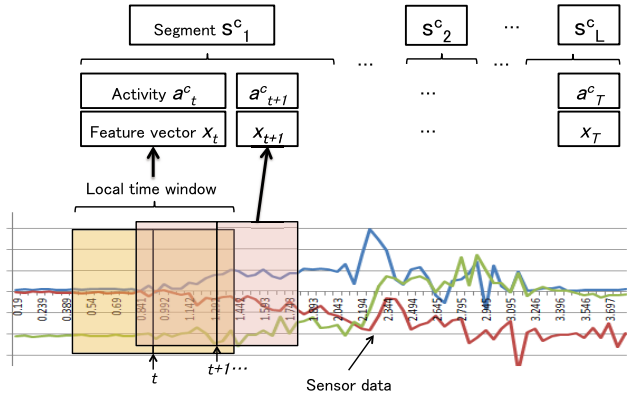


Fig. 4 Overview of one-day activities for a single activity class $c \in C$.

3.2 Preliminary

As a preliminary step, we introduce the mathematical expressions used throughout this paper. **Table 2** provides a summary of expressions, and **Fig. 4** shows an overview of the expression for a single activity class c.

For simplicity, we assume that the time of day is expressed as an integer between one and T. We abbreviate the sequence $(1, 2, \cdots, T)$ as $1:T$. For each t, we assume that a feature vector is extracted t. For each t, we assume that a *feature vector* is extracted that contains several statistic values from the time window of the sensor input around t. For example, if we adopt a time window of 5 seconds with a shift of 2.5 seconds — which we adopted throughout this paper —, we can assume that T = 24 [hour]×60 [minutes]×60 [seconds]/2.5 [seconds]−1 = 34,559. We specify the sequence of feature vectors $(x_1, x_2, \cdots, x_T)$ as $x_{1:T}$.

Moreover, C refers to the set of activity classes to be recognized. We assume that at any time t multiple activities might be included, either because the nurse is performing several activities concurrently, or because the activity-recognition algorithm conducts fuzzy estimations. Therefore, we define whether the activity at time t is $c \in C$ or not as the binary value a_t^c.

In the remainder of the section, we focus on the recognition of a single activity $c \in C$. In reality, we could apply the proposed method for each activity $c \in C$, and adopt either the most probable class $\arg_c \max \mathbf{P}(a_t^c)$, or adopt all classes estimated for a time t. In Section 4, we evaluated the accuracy using the latter strategy.

We use the term *segment* as the continuous time range where

the activity c is performed, and represent it as a pair of start and end times. When we assume that L^c segments are repeated for activity c in a day, the lth segment from time $b(l)$ to $e(l)$ is defined as:

$$s_l^c := (b(l), e(l)), \quad \text{where } 1 \le b(l) \le e(l) \le T.$$

The traditional activity recognition such as that from Bao et al. [2], can be modeled as the problem of obtaining the maximum argument $c \in C$ of

$$\mathbf{P}(a_t^c | x_t) \tag{1}$$

for the local time window t only. Note that obtaining $\mathbf{P}(x_t|a_t^c)$ is easy following Bayes' theorem. For the rest of this paper, we call $\mathbf{P}(x_t|a_t^c)$ a *local time likelihood*.

In contrast, our goal can be represented as the problem of obtaining the probability of an entire day's activities

$$\mathbf{P}(a_{1:T}^c | x_{1:T}). \tag{2}$$

For the remainder of this section, we describe the method used to conduct this.

3.3 Proposed Method

We assume the Bayesian network as shown in **Fig. 5**.

Figure 5 represents the conditional probabilities for one segment s_l^c. We assume that the probabilities between any two segments s_l^c and $s_{l'}^c$ $(l \neq l')$ are independent.

The marginal probability of the figure is written as

$$\begin{aligned}&\mathbf{P}(x_{b(l):e(l)}, a_{b(l):e(l)}^c, s_l^c)\\ &= \mathbf{P}(s_l^c) \prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a_t^c)\mathbf{P}(a_t^c|s_l^c)\end{aligned}$$

when s_l^c is fixed, then a_t^c for $b(l) \le t \le e(l)$ is straightforward, and we can eliminate $\mathbf{P}(a_t^c|s_l^c)$. Accordingly,

$$= \mathbf{P}(s_l^c) \prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a_t^c)$$

To obtain the conditional probability between $a_{b(l):e(l)}^c$ and $x_{b(l):e(l)}$, we marginalize s_t^c, then

$$\mathbf{P}(a_{b(l):e(l)}^c, x_{b(l):e(l)}) = \sum_{s_l^c} \mathbf{P}(s_l^c) \prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a_t^c). \tag{3}$$

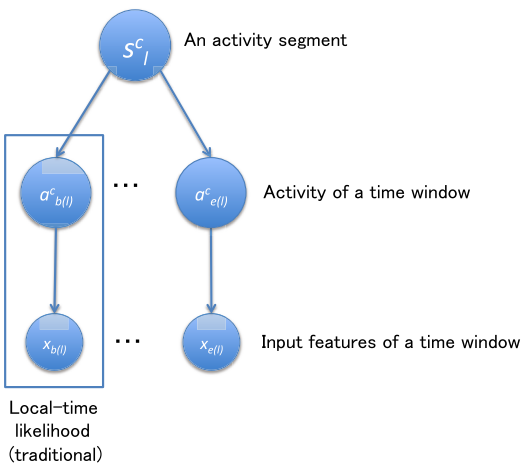


Fig. 5 Overview of the proposed method.

Next, we divide the time sequence $1:T$ to the segments

$$\{b(1):e(1)\},\ \{b(2):e(2)\},\ \cdots,\ \{b(L^c):e(L^c))\}$$

and consider the marginal probability for all the times $1:T$ as

$$\begin{aligned}&\mathbf{P}(a_{1:T}^c, x_{1:T})\\ &= \mathbf{P}\Big(a_{b(1):e(1)}^c, x_{b(1):e(1)},\\ &\qquad a_{b(2):e(2)}^c, x_{b(2):e(2)},\\ &\qquad \cdots,\\ &\qquad a_{b(L^c):e(L^c)}^c, x_{b(L^c):e(L^c)}\Big)\end{aligned}$$

Assuming any pairs of segments are independent of each other, The formula is written as the product of the segment marginal probabilities, as

$$= \prod_{l \in 1:L^c} \mathbf{P}(a_{b(l):e(l)}^c, x_{b(l):e(l)}).$$

Substituting Eq. (3),

$$= \prod_{l \in 1:L^c} \Big\{ \sum_{s_l^c} \mathbf{P}(s_l^c) \prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a_t^c) \Big\}$$

Therefore, given the input $x_{1:T}$,

$$\begin{aligned}&\mathbf{P}(a_{1:T}^c | x_{1:T})\\ &\propto \prod_{l \in 1:L^c} \Big\{ \sum_{s_l^c} \mathbf{P}(s_l^c) \prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a_t^c) \Big\}\end{aligned} \tag{4}$$

This formula utilizes not only the local time likelihood $\mathbf{P}(x_t|a_t^c)$ as the traditional approach in Eq. (1), but also the prior probability of the segments $\mathbf{P}(s_l^c)$. We use the local time likelihood $\mathbf{P}(x_t|a_t^c)$ from the result of the naive method, and also prepare and utilize the prior probability $\mathbf{P}(s_l^c)$ using the samples from the training data. Because $\mathbf{P}(s_l^c)$ can be informative when we obtain training data for an entire day, our method can lead to accuracy improvement for activity recognition of an entire day.

3.4 Implementation

In the implementation, we calculate Eq. (4) according to the following steps, where we adopt the logarithmic probability to avoid underflows, and exploit the idea of importance sampling to obtain those samples weighted by the prior knowledge of the segments.

(1) Train local time log likelihood

$$\log \mathbf{P}(x_t|a_t^c) \quad \text{for each } t \in 1:T$$

with the naive method, and store the results.

(2) Construct $\mathbf{P}(s_l^c)$ from the training data. This probability is implemented as a set of k samples from the training data. We represent the sampled segment numbers as $l[1], l[2], \cdots, l[k]$, and i'th sample as $s_{l[i]}^c$, where $1 \le i \le k$.

(3) For each $s_{l[i]}^c$ in Step 2, calculate

$$\exp\Big(\sum_{t \in b(l[i]):e(l[i])} \log \mathbf{P}(x_t|a_t^c) \Big) \tag{5}$$

using the result of Step 1).

(4) The average of Eq. (5) for $l[1], l[2], \cdots, l[k]$ can be regarded

Table 3 List of feature variables after feature selection.

Feature No.	Feature	Sensor	Axis (if any)
1	Mean intensity	Chest	
2	Mean intensity	Right wrist	
3	Mean	Chest	Y
4–6	Mean	Waist	X, Y, and Z
7–9	Mean	Right wrist	X, Y, and Z
10	Variance of intensity	Right wrist	
11–13	Variance	Right wrist	X, Y, and Z
14–15	Variance	Chest	Y and Z
16	Variance	Right wrist	Z
17–18	Mean FFT-domain energy	Chest	Y and Z
19–20	Mean FFT-domain energy	Right wrist	X and Z
21	Mean sum of the absolute values of each axis	Chest	
22	Mean sum of the absolute values of each axis	Waist	
23	Number of samples out of mean intensity $\pm 0.1G$	Right wrist	
24	Number of samples out of mean intensity $\pm 0.1G$	Waist	
25	Number of crosses of the zone of the mean intensity $\pm 0.1G$	Waist	
26	Number of crosses of the zone of the mean intensity $\pm 0.1G$	Right wrist	
27	Covariance between intensities	Chest and Waist	

as importance sampling of and an approximation of Eq. (3), where the occurrence of sampled segment $s^c_{l[i]}$ follows the former term of Eq. (3): $\mathbf{P}(s^c_l)$, and Eq. (5) is the same as the latter term of Eq. (3): $\prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a^c_t)$ except for the exp-sum-log calculation to avoid underflows. Because Eq. (3) is the same function for any $l \in 1 : L^c$, we can utilize this function directly to estimate s^c_l rather than completely calculate Eq. (4). In practice, to simplify the calculation of the average for day wise, we can pick up some $l \in 1 : L^c$ with larger Eq. (3) values by a threshold such as the average of Eq. (3).

Note that $\log \mathbf{P}(x_t|a^c_t)$ can be used multiple times for different $s^c_{l[i]}$ in Step 3, and thus they are pre-calculated and stored in Step 1 to avoid redundant calculations.

4. Evaluation

In this section, we describe the dataset collected from actual nurses wearing accelerometers in a hospital for approximately two weeks, and we evaluate our proposed method by applying it to this collected data.

4.1 Objective

The goal of the evaluation is to answer the following questions:

(1) Can the proposed method improve the recognition accuracy?
(2) Can the proposed method estimate better segments?
(3) Can we obtain knowledge about nursing activities or clinical pathways from the real data?

For Question (1), we evaluate accuracy compared with the naive method indicated in Section 4.5.1 and Section 4.5.2. For Question (2), we evaluate the activity durations indicated in Section 4.5.3. Moreover, for Question (3), we discover knowledge about the nursing activities by applying our method to the two years of data collected, and explore correlations with the medical data.

4.2 Preprocessing

From the labeled data, we extracted feature vectors from the three axes using the accelerometer data. For the sensor data, time windows of 5 seconds were extracted, shifting every 2.5 seconds, as in Bao et al. [2]. For each time window, we calculated 47 feature values, following Refs. [60], [61].

We reduced the 47 feature variables to 27 by applying stepwise-feature selection [15] to 1,000 randomly sampled vectors over ten iterations. The feature variables that were selected are listed in **Table 3**.

4.3 Applying the Method

In order to evaluate our proposed method, we compared the *proposed* method with the prior knowledge about $\mathbf{P}(s^c_l)$, and the *naive* method without the prior knowledge. As underlying machine learning algorithms for $\mathbf{P}(x_t|a^c_t)$, which is the same as the naive method after applying the Bayes' theorem, we adopted k-Nearest Neighborhood (*k-NN*), naive Bayes (*NaiveBayes*) and RandomForest, and evaluated each of them. We adopted a Gaussian distribution for the naive Bayes method, which is a parametric model of probabilities. Because it assumes a specific probability function, it may lead to an incorrect modelling of the probability. Therefore, we also adopt k-NN, which can non-parametrically approximate the probability by using the powered inverse of distances with the k'th samples, as addressed in many literatures. Random forest does not have such a proven approximation, as far as we know, but we can apply Bayes' rule to the majority rate obtained from each weak-decision tree. Random forest is popular and achieves a better accuracy in many papers, then we adopted this to demonstrate the use of a high-performance baseline.

The detail of the methods are described in the following: In order to evaluate the accuracy of real usage where the training and usage data are different, we applied 1-*duty-day*-left-out cross validation, which means testing each nurse's working day with the model trained with the data that have either different days or different nurses.

4.4 Evaluation method

To evaluate the proposed method, precisions, recalls, and F-measures for each time window are not necessary for the following reasons:

(1) The targeted real data are imbalanced, as discussed previously. Standard measures, such as precision and F-measure, are affected by these imbalances, because they use the ratio of positive samples to negative samples. For example, for

a rare activity type, the negative samples increase, the false negatives tends to increase, and thereby the precision varies depending on the imbalances. Thus, it is preferable to use imbalance-independent measures.

(2) The accuracy for time windows does not consider the closeness between true and estimated segments. No matter whether the mis-estimated segment is 1 second away or 1 hour away, the traditional accuracy measure concludes the same value. Since our method tries to estimate activities considering a probable segment with start and end times, we need to develop the measures to reflect such a closeness.

(3) The activity duration is also important. The traditional measures do not consider the fragmentation of estimated activities. If the fragmentation remains in the estimated activity sequences, this can result in many segments of shorter durations. In order to analyze nursing activities, duration is one of the critical values.

To overcome these problems, we adopted the evaluation methods introduced in this section.

With regard to point 1, we adopted BCR, a measure used by Refs. [11], [12] and defined as follows:

$$\text{BCR} = \frac{\text{TP-rate} + \text{TN-rate}}{2}$$

where TP-rate is defined as $TP/(FN+TP)$, and TN-rate is defined as $TN/(TN+FP)$, where TP (FP, TN, FN) is the number of true positives (false positives, true negatives, or false negatives, respectively). In contrast with other measures such as precision and F-measure, these values are not affected by imbalanced positive and negative samples at the ground truth level, and accordingly, BCR — the mean of them — is also imbalance independent.

With regard to point 2, in order to measure the accuracy that considers the closeness between true and estimated segments, we introduce the idea of adding time margin δ as the parameter.

When we represent the true label as

$$\hat{s}^c_l = (\hat{b}_l, \hat{e}_l) \ \text{ for } 1 \le l \le \hat{L}^c,$$

and the estimated label as

$$\tilde{s}^c{}_l = (\tilde{b}_l, \tilde{e}_l) \ \text{ for } 1 \le l \le \tilde{L}^c,$$

the normal precision/recall/F-measure is calculated between

$$(\hat{b}_l, \hat{e}_l) \ \text{ and } (\tilde{b}_l, \tilde{e}_l).$$

Here,

- *δ-precision* is defined as the precision between

  $(\hat{b}_l - \delta, \hat{e}_l + \delta)$ and $(\tilde{b}_l, \tilde{e}_l)$,

- *δ-recall* is defined as the recall between

  $(\hat{b}_l, \hat{e}_l)$ and $(\tilde{b}_l - \delta, \tilde{e}_l + \delta)$, and,

- *δ-F-measure* is the harmonic mean between δ-precision and δ-recall.

That is, these measures relax the numerator by increasing TP.

To avoid double counting, these calculations have to be done keyed by each time t, which equals to be that the overlapped segments are merged. By this, we can include the sample that resides within distance δ, with the counterpart as the correct sample.

Note that the previous definitions are the same as the traditional definitions of precision, recall, or F-measure when $\delta = 0$.

With regard to point 3, we evaluate the difference between the mean durations of the estimated and true labels for each activity. If the value is smaller, the estimated segments have closer durations to the true segments.

4.5 Results

Following the evaluation approach discussed above, we explain the results shown in **Figs. 6** and **8**. From here, to easily visualize the result, we omit the result of the activity classes for no more than 5 labels (activity class No. 25 and 28) and "Other" class. Note that these samples were used in the evaluation for reality, but just removed when showing the result.

4.5.1 Accuracy by the Balanced Classification Rate

Figure 6 shows the results for k-NN, NaiveBayes, and RandomForest as the underlying machine learning algorithm.

As we can see from the figure, most of the activity classes improve with our method. Averaging all activity classes, when we adopt k-NN as the underlying algorithm, BCR for naive method is 56.10% (σ = 9.6), and for the proposed method, it is 73.18% (σ = 14.2). When we adopt Naive Bayes as the underlying algorithm, BCR for the naive method is 55.15% (σ = 15.8), and for the proposed method, it is 80.96% (σ = 14.5). Moreover, when we adopt RandomForest as the underlying algorithm, BCR

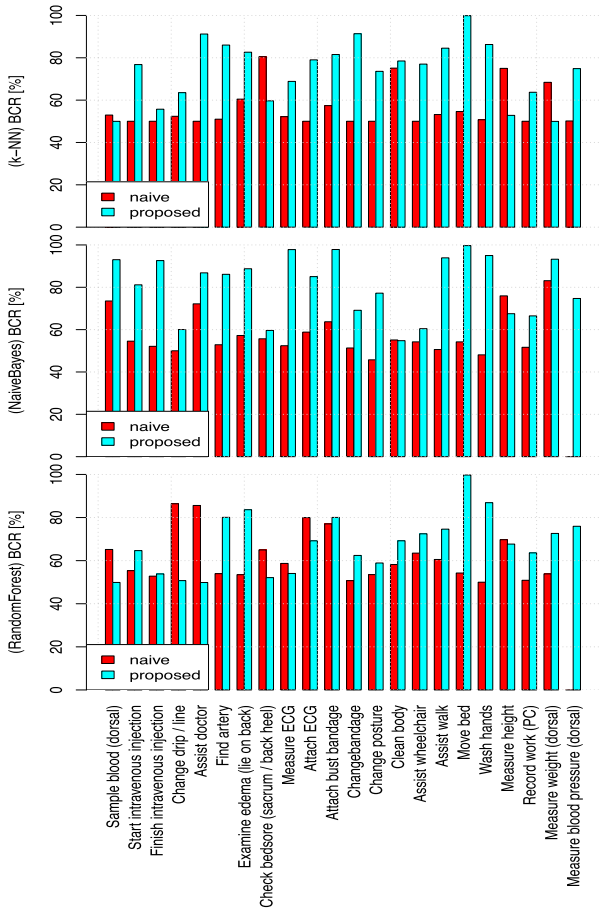
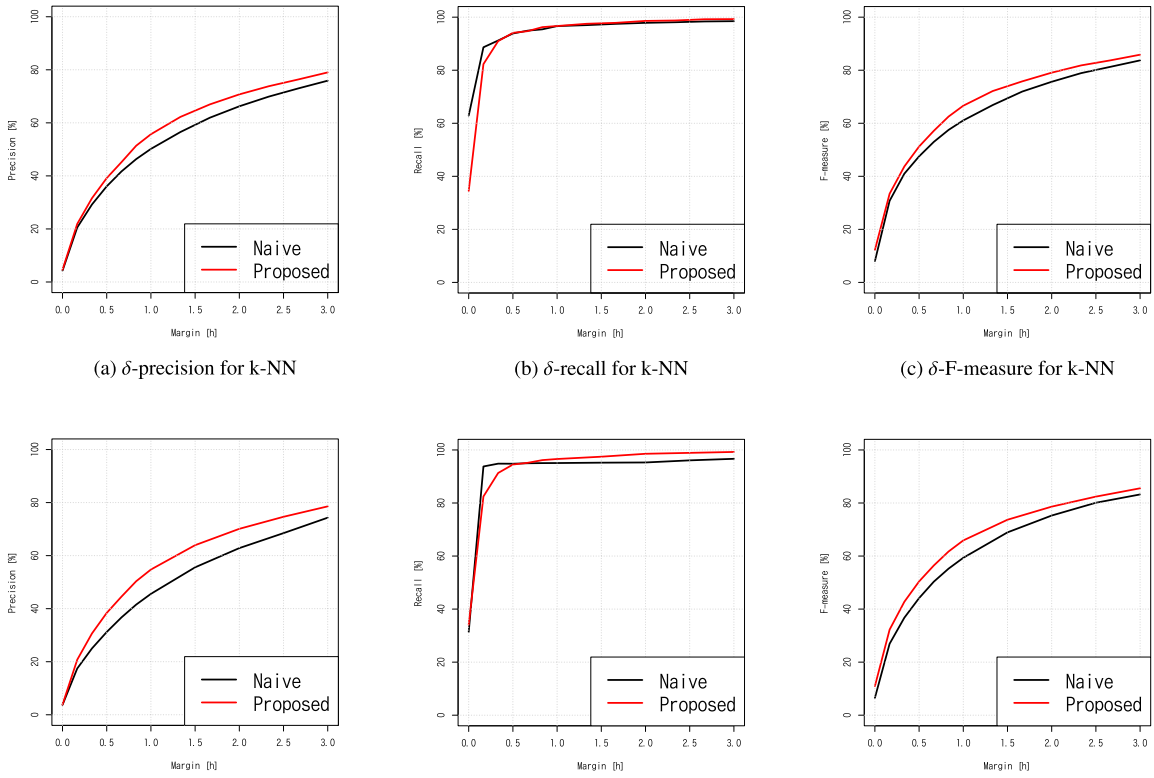


Fig. 6 BCR for naive/proposed methods for each activity with k-NN (NaiveBayes, RandomForest) (Average: 56.10 (55.15, 59.03)% for the naive method and 73.18 (80.96, 67.83, respectively)% for the proposed method).



(a) δ-precision for k-NN (b) δ-recall for k-NN (c) δ-F-measure for k-NN

(d) δ-precision for RandomForest (e) δ-recall for RandomForest (f) δ-F-measure for RandomForest

Fig. 7 δ-precision, δ-recall, and δ-F-measure for varying margin δ.

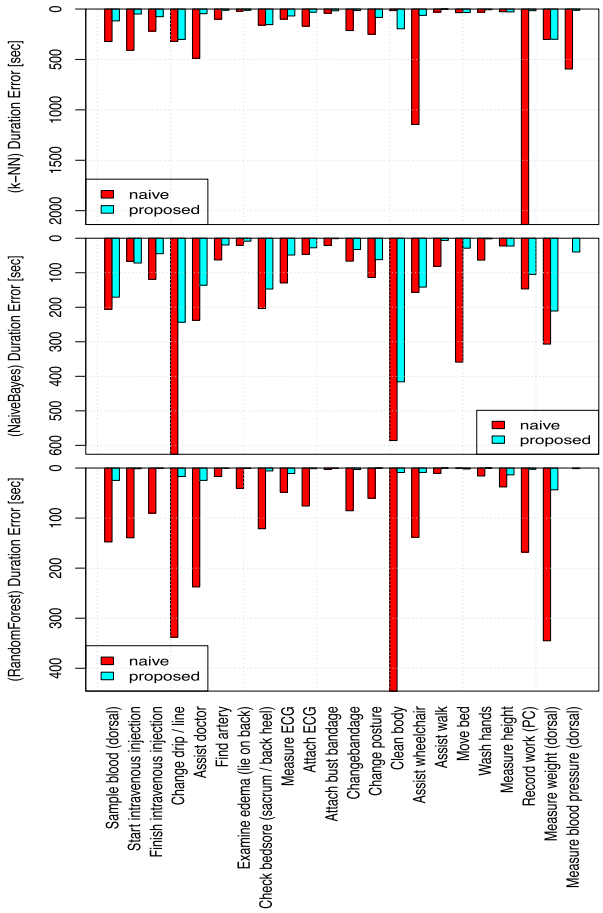


Fig. 8 Errors for activity durations for k-NN (NaiveBayes, RandomForest) (Average: 324.2 (173.5, 122.2) seconds for the naive method and 74.6 (90.33, 7.88, respectively) seconds for the proposed method).

for the naive method is 59.03% (σ = 17.3), and for the proposed method, it is 67.83% (σ = 13.4).

4.5.2 Accuracy Considering Margins

Figure 7 shows δ-precision, δ-recall, and δ-F-measure for varieties of δ from 0 to 3 hours. To know the margin effects, we omit dates with neither true nor estimated activities. From the figures, we can observe that all values increase as the margin δ increases. The precisions are relatively lower than the recalls, but the proposed method outperforms at maximum approximately 5.7% between 1 and 1.5 hours. On the other hand, the δ-recall of the proposed method underperforms at $\delta = 0$, but increases rapidly until $\delta = 0.5$ hours, and slightly outperforms the naive method. This implies that many true activities fail to be estimated in the proposed method, but often reside within 30 minutes away. Also, the underperformance of the δ-recall at $\delta = 0$ implies that the proposed method too passively estimates positive considering the segment effect. Considering the harmonic accuracies with δ-F-measure, our method outperforms the proposed method. The maximum improvement is 5.5% at 2.5 hours in k-NN, and 6.5% at 1hour in RandomForest.

4.5.3 Accuracy of Activity Durations

Figure 8 shows the error for the mean activity durations for the naive and proposed methods for each activity. Because the y-axis is the error, the smaller the y-axis, the better is the accuracy. From the figure, in any activity class, the proposed method greatly outperforms the naive method. The mean errors are 324.2 seconds for the naive method and 74.6 seconds for the proposed method, with k-NN. When using NaiveBayes, they are 173.5 seconds for the naive method and 90.33 seconds for the proposed method.

Moreover, when we use RandomForest, they are 122.2 seconds for the naive method and 7.88 seconds for the proposed method.

4.6 Discussion

As a result of the evaluations with BCR, our proposed method outperformed the naive method by 17.08 (25.81, 8.8)% with k-NN (NaiveBayes, Random Forest, respectively). Although the best absolute accuracy of the naive methods is by RandomForest, the best improvement and absolute accuracy of the proposed method is by NaiveBayes, and the second is k-NN. The reason why RandomForest was not improved so much would be because the probability approximation is not perfect, as addressed in Section 4.3, compared with other probabilistic methods.

On the other hand, the accuracy of activity durations was the best in RandomForest even in the absolute error value. Compared with the prior knowledge about the timestamps of segments, knowledge about the activity durations seems to be effective in any underlying algorithms.

From the result considering margins, we can get better accuracies if we allow 0.5–2.5 hours of margins. Our method can partly replace the manual work for recording their nursing care to the electronic medical record system, which takes long time as shown in Section 5. In the current manual work, such margins often occur such that records in the morning are input at once before lunch time. Taking into account such current situation, the above-mentioned margins can be acceptable for such application.

Although we achieved improvements for BCR, further work for other types of improvements such as the precision and the F-measure, are important. The low precision shown in the result is not still sufficient. However, in the rest of the paper, we focus on the analysis of activity segments, which includes the information of activity durations, and applied big data analytics in Section 5, where we believe that we could avoid the problem of low precision as much as possible. This is inherently difficult to achieve, for example, prediction of disasters or diseases that hardly occur, but other approaches, such as feature engineering, and considering state-transition probabilities, such as Ref. [56] should be explored as future work.

Overall, from the underperformance of the δ-recall at lower δ in Fig. 7, the proposed method seems to passively estimates positive, considering the segment effect, whereas the BCRs in Fig. 6 improved relatively well, which means that the TN-rate was generally improved. On the other hand, our method allows to estimate concurrent activities, but the dataset is labeled serially, so there is a possibility that the false positives in the evaluation may be true positives if we use concurrent labels for training and test data. However, the inference that the TN-rate was improved implies that it will still be effective even when we use concurrently labeled data.

Instead of the prior knowledge about the timestamps, it is possible to use the timestamps (in our example, time-of-day) as a feature. However, the prior knowledge about the activity duration cannot be utilized. Since the activity duration is only known when the segment is defined, it is not applicable for the features in the traditional method. The activity durations are drastically improved on our method as in Fig. 8, it would be an advantage of our method.

We assume that we can obtain a multiple activity classes simultaneously. If we assume that we can restrict to a single activity class at a time, the problem is more difficult. Approaches such as optimizing multi-class ROC [16], [51] can be the candidates for solving this problem.

In this paper, we adopt k-NN, NaiveBayes, and RandomForest as the underlining algorithms. Nonetheless, our approach can be used as a post-process of any type of estimation algorithm that can output local-time likelihood.

5. Applying to Bigger Data

Using our method, we demonstrate an example of recognizing and analyzing bigger datasets, such as correlation with nurses' experience, correlation with patients' levels of nursing needs, and the relationship between delays of discharges.

For the unlabeled data, we extracted 265,002 time windows, which corresponded to 771 duty days × nurses, and applied our proposed method in order to estimate the real activities involved in nursing duties.

In this section, for each nursing activity durations in a day, we 1) first show their average durations, 2) show the correlation between nurses' profiles and them, and 3) show the correlation between them and patients' discharge delays.

5.1 Nursing Times in a Day

For 658 daytime duties, the average time for the defined care time is 277.8 minutes with $\sigma = 55.7$.

Figure 9 is the estimated average care times for each activity class in one daytime. From the figure, we can see the types of activity on which the nurses spend more time, such as "Measure blood pressure", and "Find artery". We can also see that the nurses spend significant time recording their work on a PC, which were introduced after the electronic medical record system was introduced, and hence there is an opportunity for reducing this time.

5.2 Correlation with the Nurses' Profile

If we join the results with additional data, such as nurses' profiles, we can data mine further knowledge. To demonstrate this, we joined the results with the number of experienced years, age,

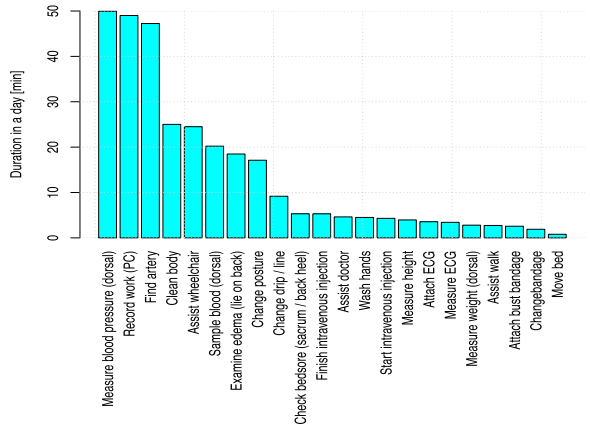
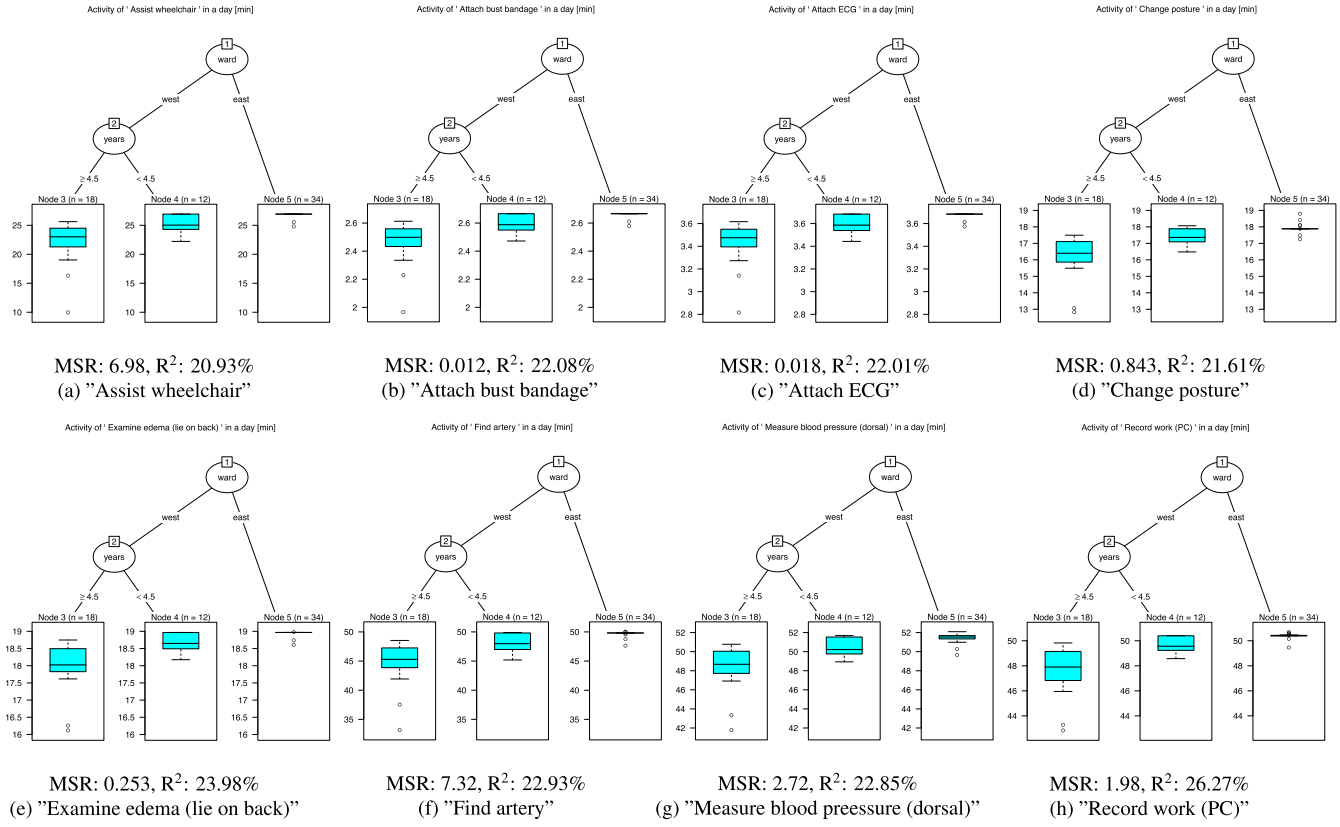


Fig. 9 The nursing times for each activity class in one daytime.



MSR: 6.98, R^2: 20.93%
(a) "Assist wheelchair"

MSR: 0.012, R^2: 22.08%
(b) "Attach bust bandage"

MSR: 0.018, R^2: 22.01%
(c) "Attach ECG"

MSR: 0.843, R^2: 21.61%
(d) "Change posture"

MSR: 0.253, R^2: 23.98%
(e) "Examine edema (lie on back)"

MSR: 7.32, R^2: 22.93%
(f) "Find artery"

MSR: 2.72, R^2: 22.85%
(g) "Measure blood preessure (dorsal)"

MSR: 1.98, R^2: 26.27%
(h) "Record work (PC)"

Fig. 10 Regression trees for each activity duration in a day by nurses' profiles with over 20% of pseudo R squared with RandomForest. Each caption shows the mean of squared residuals (MSR) and the pseudo R squared with RandomForest.

gender, title, and the ward (west/east, where the west ward is for internal medicine, and the east is for surgery) for each nurse. The mean experienced years is 7.36 years with $\sigma = 5.65$, the minimum is 1, and the maximum is 25. The joined data consists of 64 samples each of which corresponds to each nurse.

For the joined data, we applied RandomForest algorithm for each activity class as an response variable, and the profiles as predictor variables.

When we use RandomForest for correlation analysis, there are several advantages compared with a traditional regression: 1) RandomForest automatically avoids overfitting, and outputs general models, 2) we can see the *importances* of variables after neutralizing interactions among variables unlike a traditional regression, 3) also with neutralizing interactions, we can see the effect of each variable to the response variably by *a partial dependent plot*, 4) if we pick up a tree from the set of obtained trees, we can easily understand the partitioning conditions compared to other algorithms such as SVM.

For the models for each activity class as response variable, we picked up the models which have the *pseudo* R^2, which is defined as 1 − (mean squared error)/(variance of the response variable), are more than 20%, and showed the scores and (no-random but naive) regression trees in **Fig. 10** to visualize example trees.

In any tree in the figure, the first partitioning is done by "ward". They have higher activity durations for the east ward, and are divided to experienced years < 4.5 [years] with middle durations and the rest with lower durations. It seems that there are differences in nursing activity durations in a day between the west internal medicine department ward and the east surgery one. For the internal medicine department, it seems that there are varieties of durations, and unexperienced nurses performed longer compared with experienced ones.

From such results, we or the nurses can estimate the differences of work load between wards or the years of experiences, and reallocate and equalize unbalanced work load, if any.

5.3 Correlation with Patients' Discharge Delays

We joined the estimated activity data with the patient record, and compared the amount of time spent by nurses for each activity with the duration of hospitalization, where 4 inpatient days are normal, and over 5 days increase in medical costs. In the experiment, we asked the nurses to attach RFID tags, which communicates with readers which were equipped at the entrances of the patients' rooms, each of which is a personal room. Therefore, the log of the RFID readings provides relationships between nurses who took care and patients who have been taken care of during the day. We first joined the RFID records with the estimated activity data, and then joined with the patients' data about the hospitalization days. The number of patients after joining with the patient record is 28 with 24 nurses for 35 days, and the number of the samples is 54.

For the joined data, we applied the Random Forest algorithm again with the numbers of hospitalization days as a response variable, and with the activity durations in a day as predictor variables.

Then, we can obtain the importances for each predictor vari-

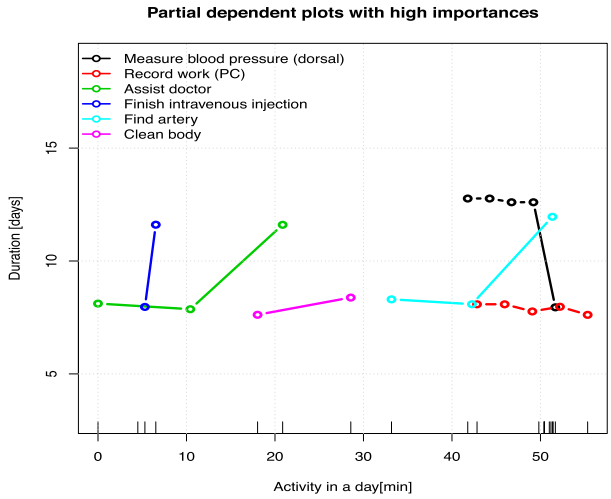


Fig. 11 Partial dependent plots to show the effect of each nursing activity time in a day (with high importance) to hospitalization durations of the cared patients.

ables, each of which are the mean increase of accuracy by the variable. For the importances, we picked up the variables with the importances of over 50, and showed the partial dependent plots in **Fig. 11**. The partial dependent plot is a plot between a predictor and the response variable, which plots the effect of the predictor to the response after marginalizing the other predictor variables. Unlike traditional regression, it can approximately eliminate the interaction effects between predictor variables.

From the figure, we can observe that "Measure blood pressure" takes shorter times for patients with longer hospitalizations, the times for "Assist doctor" and "Find artery" increase if the patients have longer hospitalizations, "Finish intravenous injections" takes slightly longer if the hospitalizations take longer, and "Clean body" takes quite longer if the hospitalizations take longer. As such, we can estimate the effect of each nursing activity time to the hospitalization duration, and we can estimate the work load for each patient. For example, if we observe a patient with longer activity durations such as "Assist doctor", "Find artery", "Finish intravenous injection", and "Clean body", then we can estimate the patient might delay for the discharge, and vise versa for "Measure blood pressure". On the other hand, if we can predict the delay of a patient from other data such as the result of the operation, we can estimate and prepare some nursing activities with longer durations.

As shown in this section, by linking our proposed method with additional data which already exist in hospitals, we can produce a valuable knowledge for reflecting and improving medical processes.

6. Related Work

In the literature, many works attempted mobile activity recognition [2], [5], [10], [29], [31], [33], [35], [36], [46], [52], [58]. Recently, in the medical field, many experiments have collected activity data from doctors, nurses, and patients; many studies make use of these collected (big) data for improving the efficiency of duties or for offering the appropriate medical services [3], [39], [42], [48], [53]. In Ref. [40], activity recognition of nurses and development of a labeling automation system using activity label information such as nurse activity, meeting information, audio and video data collected through the sensor network comprised of wearable acceleration sensors or environmental setting type sensors was conducted. Also, Ref. [1] collected labeled nursing activity sensor data from nurses coat pockets for 14 days, as in our paper, and tried activity recognition for 14 activity classes, with/without combining with nurses workflow data. However, unlike our paper, they do not define the activity classes based on the standardized clinical processes, which makes recognition accuracy higher, but less realistic. Moreover, we also contribute to make the real nursing activity data publicly open.

Because activity recognition manages sequential data, techniques for sequential data such as Hidden Markov Model (HMM) [9], [29] and Conditional Random Fields (CRF) [9], [19], [30], [32], used in speech recognition and natural language processing, are related. Some works have attempted to apply these techniques to mobile activity recognition [50], [54], [55], [59]. Here, we claim that using HMM and CRF are independent of our contribution. Basically, HMM and CRF are not segmenting wise, but they are time window wise if we use them straightforwardly. Then, we can apply our method to utilize prior knowledge independently. Applying HMM and CRF for segment wise is not straightforward since they are not determined from the first. And, HMM and CRF are complex to estimate the parameter, but our method can simply integrate and utilize other popular methods of non-sequential machine learning.

Another approach that is applicable to sequential data is Bag-of-Features (BoF), which makes histograms of feature values and utilizes their statistic features [60], [61]. However, this can only be applicable to data that is already segmented. The segmentation technique is common in speech recognition [6], [13] and natural language processing [8], [37], [38], [57].

However, among the aforementioned work, to the best of our knowledge, none addresses the challenges of real-world applications, nor tries to utilize the prior knowledge on a daily basis, like our method. Class-wise prior probability, timestamps in a day, and activity durations have large variances. These can result in difficulties in activity recognition when applying the existing work.

With regard to activity durations, Ref. [56] adopts the CRF model that can integrate the knowledge of activity duration using Semi-CRF, which learns segmentation in addition to the Markov transitions, as well as the traditional CRFs. Moreover, it improves computation costs by considering omitting "other" activities. This work generates promising results in accuracy, although the computation and parameter estimation often becomes complex in such a high-dimensional approach. Our method manages the duration and segments as a prior knowledge obtained from the training dataset, and infers the activities considering them by Bayesian network and importance sampling approaches, which is demonstrated to be tractable in real nursing big data.

In addition, the challenge lies in recognizing complex domain-specific activities such as nursing activities which, we resolve in our paper.

For machine learning from imbalanced data, problems and approaches are addressed in the literature. Classification for im-

balanced data is highly important in the area of risk management, such as medical decision domains, where a positive instance, such as a specific disease, hardly occurs. Reference [18] introduced several assessment metrics, such as ROC that is robust for imbalanced data, and reviewed several approaches, such as importance sampling, cost-sensitive methods, and active learning. It also addresses the effectiveness of *one-class learning*, a binary classification of positive or negative. Reference [27] applied the empirical evaluation of RandomForest algorithm for imbalanced data. Reference [21] proposed a sampling method for bagging, and evaluated their method using AUC. Reference [28] evaluated several boosting and bagging algorithms comprehensively for noisy and imbalanced data, and concluded that bagging generally outperforms boosting. In this paper, we incorporate the robustness of one-class learning to our method.

In the literature, several datasets for mobile activity recognition are available. Reference [17], with their large-scale activity collection, collected over 35,000 activities from more than 200 people over approximately 13 months. Reference [24] provided a dataset that consists of 28 days of sensor data from a single person with annotations added by their proposed system. Reference [25] was a unique trail to collect activity recognition datasets from the laboratories of multiple universities. In the 5 years, the total number of activities reached over 50,000 samples. Reference [7] provided a dataset with varieties of sensor displacement status for 33 fitness activities from 17 participants. References [4], [45] provided an activity dataset with sensor-rich environment where the subjects wore multiple sensors on the body, with more than 27,000 activities from 12 subjects. Among them, Refs. [4], [7], [45] provided an entire day data/multi-day data as a part of them. However, the activity classes are of common types, such as those that appear in Activity in Daily Life (ADL) records, and not similar to our dataset, which is closely coupled with the application domain and domain data, such as medical records.

7. Conclusion

In this paper, we collected a real nursing dataset for mobile activity recognition that can be used for supervised machine learning, and proposed a method for recognizing activities for an entire day utilizing prior knowledge about the activity segments in a day. The results showed accuracy improvement compared with the baseline method that did not employ our method; in particular, there were significant improvements in activity durations. It implies that the dataset are valid, and that the proposed method is effective.

We also demonstrated data mining by applying our method to bigger data combined with 2 years of patient medical records, and demonstrated the value of linking with additional day utilizing RandomForest regression. The future work includes expanding the data mining in order to explore the knowledge about clinical paths, such as finding important activities that lead to earlier discharge from the hospital.

Because activity recognition in nursing domain is new and challenging, there is no statement or reference how much accuracy is required, and our method cannot be benchmarked. However, we believe the result of the paper can be a reference of how challenging it is, and moreover, we claim that we could achieve a non-negligible improvement for the durations of activities, and demonstrated the durations could be used for nursing activity analysis.

The data we used were collected carefully to be used as open data by obtaining agreements from the subject nurses and patients. The data are also related to RFID tag data in order to recognize nurses' entry into patients' rooms, vital data from hospitalized patients (e.g., cardiograms, bed sensors to measure heart rate/breathing/body movements), accelerometer, in-room sensors, and medical information recorded in the electronic clinical pathways, and indirectly, inpatient sensor data. As future work, data mining these whole data combined with the activity recognition result and extracting valuable knowledge which contributes to efficient clinical pathways and better health care will be important.

Acknowledgments The part of this research are supported by Funding Program for World-leading Innovation R&D on Science and Technology (FIRST) "Developement of the Fastest Database Engine for the Era of Very Large Database and Experiment and Evaluation of Strategic Social Services Enabled by the Database Engine (Main Researcher: Masaru Kitsuregawa)", and Grant-in-Aid for Scientific Research (B) "Sensor Context Estimation Techniques by Semantic and Physical Layer (Principal Investigator: Sozo Inoue)". The authors would like to thank their support. We also appreciate the cooperation for experiment by the staff of Saiseikai Kumamoto Hospital, Japan.

References

[1] Bahle, G., Gruenerbl, A., Lukowicz, P., Bignotti, E., Zeni, M. and Giunchiglia, F.: Recognizing hospital care activities with a coat pocket worn smartphone, *2014 6th International Conference on Mobile Computing, Applications and Services* (*MobiCASE*), pp.175–181, IEEE (2014).

[2] Bao, L. and Intille, S.S.: Activity Recognition from User-Annotated Acceleration Data, *Pervasive Computing*, pp.1–17 (online), DOI: 10.1007/b96922 (2004).

[3] Bardram, J.E. and Christensen, H.B.: Pervasive computing support for hospitals: An overview of the activity-based computing project, *IEEE Pervasive Computing*, Vol.6, No.1, pp.44–51 (online), DOI: 10.1109/MPRV.2007.19 (2007).

[4] Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Troster, G., Millan, J.D.R. and Roggen, D.: The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition, *Pattern Recognition Letters*, Vol.34, pp.2033–2042 (online), DOI: 10.1016/j.patrec.2012.12.014 (2013).

[5] Chen, L., Hoey, J., Nugent, C.D., Cook, D.J. and Yu, Z.: Sensor-based activity recognition, *IEEE Trans. Systems, Man, and Cybernetics*, Vol.42, No.6, pp.790–808 (2012).

[6] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P.: Natural Language Processing (almost) from Scratch, *Journal of Machine Learning Research*, Vol.12, pp.2493–2537 (2011), available from ⟨http://arxiv.org/abs/1103.0398⟩.

[7] Damas, M., Rojas, I., Amft, O., Toth, A.M., Baños, O., Pomares, H. and Tóth, M.A.: A benchmark dataset to evaluate sensor displacement in activity recognition, *Proc. 2012 ACM Conference on Ubiquitous Computing – UbiComp '12*, pp.1026–1035 (online), DOI: 10.1145/2370216.2370437 (2012).

[8] Demuynck, K. and Laureys, T.: A comparison of different approaches to automatic speech segmentation, *Text, Speech and Dialogue*, pp.277–284 (2002), available from ⟨http://link.springer.com/chapter/10.1007/3-540-46154-X_38⟩.

[9] Dietterich, T.: Machine learning for sequential data: A review, *Structural, syntactic, and statistical pattern recognition*, pp.1–15 (online), DOI: 10.1146/annurev.cs.04.060190.001351 (2002).

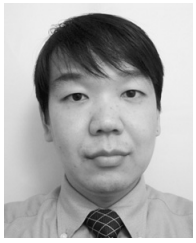
[10] Farringdon, J., Moore, A., Tilbury, N., Church, J. and Biemond, P.: Wearable sensor badge and sensor jacket for context awareness, *Digest of Papers. 3rd International Symposium on Wearable Computers*

(online), DOI: 10.1109/ISWC.1999.806681 (1999).
[11] Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers, *ReCALL*, Vol.31, pp.1–38 (online), DOI: 10.1.1.10.9777 (2004).
[12] Fawcett, T.: An introduction to ROC analysis, *Pattern Recognition Letters*, Vol.27, pp.861–874 (online), DOI: 10.1016/j.patrec.2005.10.010 (2006).
[13] Franco, A. and Destrebecqz, A.: Chunking or not chunking? How do we find words in artificial language learning?, *Advances in Cognitive Psychology*, Vol.8, pp.144–154 (online), DOI: 10.2478/v10053-008-0111-3 (2012).
[14] Gooch, P. and Roudsari, A.: Computerization of workflows, guidelines, and care pathways: A review of implementation challenges for process-oriented health information systems, *Journal of the American Medical Informatics Association: JAMIA*, Vol.18, pp.738–48 (online), DOI: 10.1136/amiajnl-2010-000033 (2011).
[15] Guyon, I. and Elisseeff, A.: An introduction to variable and feature selection, *Journal of Machine Learning Research*, Vol.3, pp.1157–1182 (online), DOI: 10.1162/153244303322753616 (2003).
[16] Hand, D. and Till, R.: A simple generalisation of the area under the ROC curve for multiple class classification problems, *Machine Learning*, pp.171–186 (2001), available from ⟨http://link.springer.com/article/10.1023/A:1010920819831⟩.
[17] Hattori, Y., Inoue, S. and Hirakawa, G.: A Large Scale Gathering System for Activity Data with Mobile Sensors, *2011 15th Annual International Symposium on Wearable Computers*, pp.97–100 (2011).
[18] He, H. and Garcia, E.A.: Learning from imbalanced data, *IEEE Trans. Knowledge and Data Engineering*, Vol.21, pp.1263–1284 (online), DOI: 10.1109/TKDE.2008.239 (2009).
[19] He, X.H.X., Zemel, R. and Carreira-Perpinan, M.: Multiscale conditional random fields for image labeling, *Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004*, Vol.2 (online), DOI: 10.1109/CVPR.2004.1315232 (2004).
[20] Herdman, T.H.E. and Kamitsuru, S.E.: NANDA International nursing diagnoses: Definitions and classification 2015–2017 (2014).
[21] Hido, S., Kashima, H. and Takahashi, Y.: Roughly balanced Bagging for Imbalanced data, *Statistical Analysis and Data Mining*, Vol.2, pp.412–426 (online), DOI: 10.1002/sam.10061 (2009).
[22] Inoue, S., Ueda, N., Nohara, Y. and Nakashima, N.: Understanding Nursing Activities with Long-term Mobile Activity Recognition with Big Dataset, *The 47th ISCIE International Symposium on Stochastic Systems Theory and Its Applications* (*SSS*), Hawaii, 10 pages (2015).
[23] Inoue, S., Ueda, N., Nohara, Y. and Nakashima, N.: Mobile Activity Recognition for a Whole Day: Recognizing Real Nursing Activities with Big Dataset, *ACM Int'l Conf. Pervasive and Ubiquitous Computing* (*Ubicomp*), Osaka, pp.1269–1280, ACM (online), DOI: 10.1145/2750858.2807533 (2015).
[24] Kasteren, T.V., Noulas, A., Englebienne, G. and Kr, B.: Accurate Activity Recognition in a Home Setting, *Proc. 10th International Conference on Ubiquitous Computing*, *UbiComp '08*, pp.1–9 (online), DOI: 10.1145/1409635.1409637 (2008).
[25] Kawaguchi, N., Ogawa, N. and Iwasaki, Y.: HASC challenge: Gathering large scale human activity corpus for the real-world activity understandings, *Proc. 2nd Augmented Human International Conference*, No.27 (2011), available from ⟨http://doi.acm.org/10.1145/1959826.1959853⟩.
[26] Kelder, T., Conklin, B.R., Evelo, C.T. and Pico, A.R.: Finding the right questions: Exploratory pathway analysis to enhance biological discovery in large datasets, *PLoS Biology*, Vol.8, pp.11–12 (online), DOI: 10.1371/journal.pbio.1000472 (2010).
[27] Khoshgoftaar, T.M., Golawala, M. and Hulse, J.V.: An Empirical Study of Learning from Imbalanced Data Using Random Forest, *19th IEEE International Conference on Tools with Artificial Intelligence* (*ICTAI 2007*), pp.310–317 (online), DOI: 10.1109/ICTAI.2007.46 (2007).
[28] Khoshgoftaar, T.M., Van Hulse, J. and Napolitano, A.: Comparing boosting and bagging techniques with noisy and imbalanced data, *IEEE Trans. Systems, Man, and Cybernetics Part A: Systems and Humans*, Vol.41, pp.552–568 (online), DOI: 10.1109/TSMCA.2010.2084081 (2011).
[29] Kim, E., Helal, S. and Cook, D.: Human Activity Recognition and Pattern Discovery, *IEEE Pervasive Computing*, Vol.9, pp.48–53 (online), DOI: 10.1109/MPRV.2010.7 (2010).
[30] Klinger, R.: Classical Probabilistic Models and Conditional Random Fields, *Entropy*, Vol.51, pp.282–289 (2007).
[31] Kwapisz, J.R., Weiss, G.M. and Moore, S.A.: Activity Recognition using Cell Phone Accelerometers, *Human Factors*, Vol.12, pp.74–82 (online), DOI: 10.1145/1964897.1964918 (2010).
[32] Lafferty, J., McCallum, A. and Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. 18th International Conference on Machine Learning*, *ICML '01*, pp.282–289 (online), DOI: 10.1038/nprot.2006.61 (2001).
[33] Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T. and Campbell, A.T.: A survey of mobile phone sensing, *IEEE Communications Magazine*, Vol.48, pp.140–150 (online), DOI: 10.1109/MCOM.2010.5560598 (2010).
[34] Lemmens, L., Van Zelm, R., Vanhaecht, K. and Kerkkamp, H.: Systematic review: Indicators to evaluate effectiveness of clinical pathways for gastrointestinal surgery, Journal of Evaluation in Clinical Practice, Vol.14, No.5, pp.880–887 (2008).
[35] Lymberopoulos, D., Bamis, A. and Savvides, A.: Extracting spatiotemporal human activity patterns in assisted living using a home sensor network, *Universal Access in the Information Society*, Vol.10, No.2, pp.125–138 (2011).
[36] Mannini, A. and Sabatini, A.M.: Machine learning methods for classifying human physical activity from on-body accelerometers, *Sensors*, Vol.10, pp.1154–1175 (online), DOI: 10.3390/s100201154 (2010).
[37] McQueen, J.M.: Segmentation of continuous speech using phonotactics, *Journal of Memory and Language*, Vol.39, pp.21–46 (online), DOI: 10.1006/jmla.1998.2568 (1998).
[38] Mohri, M., Moreno, P. and Weinstein, E.: Discriminative Topic Segmentation of Text and Speech, *Artificial Intelligence*, Vol.9, pp.533–540 (2010).
[39] Morán, E.B., Tentori, M., González, V.M., Favela, J. and Martínez-Garcia, A.I.: Mobility in hospital work: Towards a pervasive computing hospital environment, *International Journal of Electronic Healthcare*, Vol.3, No.1, pp.72–89 (online), DOI: 10.1504/IJEH.2007.011481 (2007).
[40] Naya, F., Ohmura, R., Takayanagi, F., Noma, H. and Kogure, K.: Workers' Routine Activity Recognition using Body Movements and Location Information, *2006 10th IEEE International Symposium on Wearable Computers*, pp.105–108 (online), DOI: 10.1109/ISWC.2006.286351 (2006).
[41] Nohara, Y., Inoue, S., Nakashima, N., Naonori, U. and Kitsuregawa, M.: Large-scale Sensor Dataset in a Hospital, *International Workshop on Pattern Recognition for Healthcare Analytics*, Tsukuba, Japan, 4 pages (online), available from ⟨http://sozolab.jp/publications/176⟩ (2012).
[42] Orwat, C., Graefe, A. and Faulwasser, T.: Towards pervasive computing in health care – A literature review, *BMC Medical Informatics and Decision Making*, Vol.8, p.26 (online), DOI: 10.1186/1472-6947-8-26 (2008).
[43] Osmani, V., Balasubramaniam, S. and Botvich, D.: Human activity recognition in pervasive health-care: Supporting efficient remote collaboration, *Journal of Network and Computer Applications*, Vol.31, No.4, pp.628–655 (online), DOI: 10.1016/j.jnca.2007.11.002 (2008).
[44] Panella, M., Marchisio, S. and Di Stanislao, F.: Reducing clinical variations with clinical pathways: Do pathways work?, *International Journal for Quality in Health Care*, Vol.15, pp.509–521 (online), DOI: 10.1093/intqhc/mzg057 (2003).
[45] Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Forster, K., Troster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., Doppler, J., Holzmann, C., Kurz, M., Holl, G., Chavarriaga, R., Sagha, H., Bayati, H., Creatura, M. and Millan, J.d.R.: Collecting complex activity datasets in highly rich networked sensor environments, *2010 7th International Conference on Networked Sensing Systems* (*INSS*) (online), DOI: 10.1109/INSS.2010.5573462 (2010).
[46] Roggen, D., Troster, G., Lukowicz, P., Ferscha, A., Millan, J.D.R. and Chavarriaga, R.: Opportunistic human activity and context recognition, *Computer*, Vol.46, pp.36–45 (online), DOI: 10.1109/MC.2012.393 (2013).
[47] Rotter, T., Kinsman, L., James, E., Machotta, A., Gothe, H., Willis, J., Snow, P. and Kugler, J.: Clinical pathways: Effects on professional practice, patient outcomes, length of stay and hospital costs, *The Cochrane Database of Systematic Reviews*, CD006632 (online), DOI: 10.1002/14651858.CD006632.pub2 (2010).
[48] Sánchez, D., Tentori, M. and Favela, J.: Activity recognition for the smart hospital, *IEEE Intelligent Systems*, Vol.23, No.2, pp.50–57 (online), DOI: 10.1109/MIS.2008.18 (2008).
[49] Saraiya, P., North, C. and Duca, K.: Visualizing biological pathways: Requirements analysis, systems evaluation and research agenda, *Information Visualization*, Vol.4, pp.191–205 (online), DOI: 10.1057/palgrave.ivs.9500102 (2005).
[50] Sun, X., Kashima, H., Tomioka, R. and Ueda, N.: Large scale real-life action recognition using conditional random fields with stochastic training, *Advances in Knowledge Discovery and Data Mining*, *LNCS*, Vol.6635, pp.222–233 (2011), available from ⟨http://link.springer.com/chapter/10.1007/978-3-642-20847-8_19⟩.
[51] Tang, K., Wang, R. and Chen, T.: Towards Maximizing the Area Under the ROC Curve for Multi-Class Classification Problems, *AAAI*, No.Elkan 2001, pp.483–488 (2011), available from

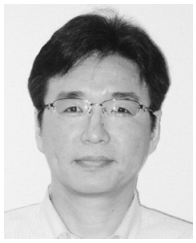
⟨http://www.aaai.org/ocs/index.php/aaai/aaai11/paper/download/3485/3882⟩.
[52] Tapia, E.M., Intille, S.S., Haskell, W., Larson, K., Wright, J., King, A. and Friedman, R.: Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor, *Proc. International Symposium on Wearable Computers*, *ISWC*, pp.37–40 (online), DOI: 10.1109/ISWC.2007.4373774 (2007).
[53] Tentori, M. and Favela, J.: Monitoring behavioral patterns in hospitals through activity-aware computing, *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*, pp.173–176, IEEE (online), DOI: 10.1109/PCTHEALTH.2008.4571062 (2008).
[54] Vail, D., Veloso, M. and Lafferty, J.: Conditional random fields for activity recognition, *Proc. 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, Vol.5, pp.1–8 (2007), available from ⟨http://portal.acm.org/citation.cfm?id=1329409⟩.
[55] Vail, D.L., Lafferty, J.D. and Veloso, M.M.: Feature selection in conditional random fields for activity recognition, *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.3379–3384 (online), DOI: 10.1109/IROS.2007.4399441 (2007).
[56] Vinh, L.T., Lee, S., Le, H.X., Ngo, H.Q., Kim, H.I., Han, M. and Lee, Y.K.: Semi-Markov conditional random fields for accelerometer-based activity recognition, *Applied Intelligence*, Vol.35, pp.226–241 (online), DOI: 10.1007/s10489-010-0216-5 (2011).
[57] Wang, D., Lu, L. and Zhang, H.-J.: Speech segmentation without speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP '03*), Vol.1, pp.468–471 (online), DOI: 10.1109/ICASSP.2003.1198819 (2003).
[58] Ward, J.A., Lukowicz, P., Tröster, G. and Starner, T.E.: Activity recognition of assembly tasks using body-worn microphones and accelerometers, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, pp.1553–1566 (online), DOI: 10.1109/TPAMI.2006.197 (2006).
[59] Zhan, K., Faux, S. and Ramos, F.: Multi-scale Conditional Random Fields for first-person activity recognition, *2014 IEEE International Conference on Pervasive Computing and Communications* (*PerCom*) (2014), available from ⟨http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6813944⟩.
[60] Zhang, M. and Sawchuk, A.A.: Motion primitive-based human activity recognition using a bag-of-features approach, *Proc. 2nd ACM SIGHIT International Health Informatics Symposium*, *IHI '12*, pp.631–640 (2012), available from ⟨http://dl.acm.org/citation.cfm?id=2110433⟩.
[61] Zhang, M. and Sawchuk, A.A.: A feature selection-based framework for human activity recognition using wearable multimodal sensors, *Int. Conf. Body Area Networks*, pp.92–98 (online), DOI: 10.4108/icst.bodynets.2011.247018 (2011).

Sozo Inoue is an associate professor in Kyushu Institute of Technology, Japan. Inoue has a Ph.D of Engineering from Kyushu University in 2003. He was an assistant professor in the Faculty of Information Science and Electrical Engineering in 2003 and an associate professor in the Research Department at the Kyushu University Library in 2006. Since 2009, he is appointed as an associate professor in the Faculty of Engineering at Kyushu Institute of Technology, Japan. Meanwhile, he was a visiting professor at Karlsruhe Institute of Technology, Germany, in 2014. He is a member of the IEEE Computer Society, the ACM, the IPSJ, the IEICE, the Japan Society for Fuzzy Theory and Intelligent Informatics, the Japan Association for Medical Informatics (JAMI), and the Database Society of Japan (DBSJ).

Naonori Ueda received his B.S., M.S., and Ph.D. degrees in Communication Engineering from Osaka University, Osaka, Japan, in 1982, 1984, and 1992, respectively. In 1984, he joined the Electrical Communication Laboratories, NTT, Japan. His current research interests include parametric and non-parametric Bayesian approach to machine learning, pattern recognition, data mining, signal processing, and cyber-physical systems. He was a director of NTT Communication Science Laboratories (April, 2010–March, 2013). Currently, he is a head of Ueda Research Laboratory (NTT Fellow) and a director of Machine Learning and Data Science Center. He is a member of the Infomation Processing Society of Japan (IPSJ), a fellow of the Institute of Electronics, Information, and Communication Engineers in Japan (IEICE), and a senior member of IEEE.



Yasunobu Nohara is a research assistant professor in Kyushu University Hospital, Japan. Nohara has a Ph.D. of Engineering from Kyushu University. His research interests include RFID systems and healthcare systems. He is a member of the IPSJ, the Japan Association for Medical Informatics (JAMI), and IEEE Computer Society.



Naoki Nakashima (M.D. Ph.D.) is the professor and director of the Medical Information Center/the director of the international patient support center in Kyushu University Hospital. He is a councilor member of Japanese Society of Diabetes Mellitus and Japan Association for Medical Informatics. He focuses on the disease management methodology to prevent chronic diseases as diabetes mellitus and complications.