Mobile Information Systems 9 (2013) 209–224
DOI 10.3233/MIS-130158
IOS Press

A mobile picture tagging system using tree-structured layered Bayesian networks

Young-Seol Lee and Sung-Bae Cho*
Soft Computing Laboratory, Department of Computer Science, Yonsei University, Seoul, Korea

Abstract. Advances in digital media technology have increased in multimedia content. Tagging is one of the most effective methods to manage a great volume of multimedia content. However, manual tagging has limitations such as human fatigue and subjective and ambiguous keywords. In this paper, we present an automatic tagging method to generate semantic annotation on a mobile phone. In order to overcome the constraints of the mobile environment, the method uses two layered Bayesian networks. In contrast to existing techniques, this approach attempts to design probabilistic models with fixed tree structures and intermediate nodes. To evaluate the performance of this method, an experiment is conducted with data collected over a month. The result shows the efficiency and effectiveness of our proposed method. Furthermore, a simple graphic user interface is developed to visualize and evaluate recognized activities and probabilities.

Keywords: Bayesian network, smartphones, picture tagging

1. Introduction

As digital media technologies have improved, a large amount of media content has been produced [1]. Multimedia digital content such as images and video are shared in online communities and social networks [13]. There are many content sharing web sites like Flickr, Picasa, and YouTube, as well as social networks like Facebook. These services have significant problems in search and retrieval of the shared content because the size of the content increases explosively. Flickr have more than 6 billion photos since August 2011, and Facebook stores roughly 100 billion pictures.

Tagging is an effective method to manage a great volume of multimedia content. Tagging assigns textual annotations to content to describe the content for other users. Tags are essential for access, search, and management of multimedia content. Many social network services such as Flickr, YouTube, Delicious, and Facebook have used semantic tags for the multimedia content to provide additional information about that content.

There are three types of tagging approaches [9]. The first type is traditional manual annotation. In this approach, users annotate content manually [12]. However, it is impractical to annotate a large amount of pictures because of human fatigue. In many cases, manual annotation may be too subjective and ambiguous. The second approach is content-based annotation using low-level features like color, shape, and texture [3,20]. However, there is a significant gap between low-level features and semantic concepts

*Corresponding author: Sung-Bae Cho, Soft Computing Laboratory, Department of Computer Science, Yonsei University, 134 Shinchon-Dong, Seodaemoon-Gu, Seoul 120-749, Korea. Tel.: +82 2 2123 2720; Fax: +82 2 365 2579; E-mail: sbcho@cs.yonsei.ac.kr.

used by humans to interpret images. Moreover, general users have difficulties to use this method due to unfamiliarity with the features. The third approach is automatic annotation technique, which builds semantic concept models and uses the concept models to label images [23,29].

Mobile phones are a suitable platform for collecting raw metadata (location, geo-coordinates, time, etc.) for multimedia contents because current mobile devices include built-in sensors such as accelerometers, digital compasses, gyroscopes, GPS receivers, microphones, and cameras. It can also provide high level metadata like a user's current status in addition to low-level features [33]. However, it is not easy to extract high-level metadata such as semantic information because mobile data are uncertain, incomplete, multimodal, and high dimensional.

This paper is an updated and extended version of [38] by including learning tree structures of Bayesian networks. In our previous work, the tree structures of Bayesian networks were designed by human, but this paper presents a method for learning the tree structures and inference using a two-layered Bayesian network using mobile data. This study handles the uncertainty of mobile data using a probabilistic model. The model attempts to infer semantic information such as human activities based on the probabilistic model with tree structure and intermediate nodes. This method helps us overcome the limitations of the mobile environment and generate semantic annotations effectively. The rest of the paper is organized as follows. Section 2 introduces related work regarding annotating multimedia content. Section 3 describes learning Bayesian network structure and intermediate nodes to reduce the cost of inference of semantic annotation. In Section 4 we introduce a two-layered Bayesian model to increase the accuracy of the annotation. Section 5 shows experiments and the results to evaluate the approach. Finally, Section 6 summarizes this paper and presents future works.

2. Related works

Annotation methodologies for media contents can be generally classified into three types [9]. The first is a manual annotation, in which users provide semantic information manually to explain documents, photographs, and video data. Manual annotation guarantees appropriate semantic information of the content based on the user's subjective view. However, this method requires a great deal of time and effort. Furthermore, human fatigue can reduce the reliability of the semantic information. Users may generate personalized and ambiguous annotation [31] difficult for other people to understand.

On the other hand, automatic annotation technologies reduce manual labor and provide annotations based on a fixed standard. Such automated technologies can provide useful semantic information such as activity and emotion, as well as low-level features such as time, date, and GPS coordinates for the content. Automatic annotation has two types of main streams: the low-level feature-based approach and high-level feature-based approach. The low-level feature-based approach can provide unambiguous and accurate annotation. For example, information such as GPS coordinates, time, or other context can be provided as metadata for a photograph. Existing context-based methodologies generate additional information like Exchangeable Image File Format (EXIF) information [15], time, and so on. However, such low level features do not provide more semantic and intuitive information for contents. A well-known semantic gap remains between the low-level computational representation and the high-level conceptual understanding of the same information [5]. A more intelligent semantic-driven approach is necessary for multimedia contents.

The high-level metadata-based approach is suitable for tagging multimedia contents. Many researchers have attempted to make annotations using data from various information sources. Most previous research used various statistical analysis and machine learning techniques like probabilistic models, similarity

Table 1
Summary of recent works

Author	Year	Data	Method
Zhang et al. [36]	2012	Visual features, user information	Using graph-based reinforcement algorithm for inter-related multi-type objects
Kelm et al. [26]	2011	User contributed metadata, visual features	Assigning geo-tags of visually similar images within geographical boundaries
Dong et al. [17]	2011	Tagged images, visual similarity	Semi-automatic tagging process in an incremental manner
Yu et al. [24]	2011	Annotated dataset, visual feature	Image segmentation and similarity calculation for annotation
Sevillano et al. [34]	2012	Textual metadata, visual similarity (L2-distance)	Extracting geographical relevance from textual metadata and a geo-tagged database
Larson et al. [19]	2011	Features derived only from the video metadata.	Using Markov chain based algorithm
Zhang et al. [35]	2012	Co-occurring tag set	Automatic tagging algorithm based on the posterior probability computation using KL divergence and tag-to-set correlation analysis
Li et al. [30]	2011	Visual features (the variety of color value) for each region	Calculating visual weight for image region and similarity comparison with the content of the region
Jesus et al. [27]	2011	Visual features and temporal features	Using Regularized Least Squares (RLSC) to detect a concept in an image and semi-automatic tagging based on game modules
Sawant et al. [25]	2011	Visual features, EXIF metadata, social comments	Using co-occurrence statistics, clustering, nearest neighbor approaches, and Bayesian classifiers
Badii et al. [1]	2011	Video data and visual features (color, texture, edge, shape, etc.)	Semi-automatic labeling using keywords extracted from visual features and topic-map-based interface
Yang et al. [39]	2011	Visual features (color histogram, color correlogram, etc.), tag sets	Image tagging approach based on near-duplicate image content and collective multi-tag association mining

measures, and social tags. Some research was based on annotation with mobile contextual information. Davis et al. tried to capture temporal, spatial, and social contextual annotation to help manage consumer multimedia content using a camera phone [18]. They automatically applied collaborative filtering techniques to contextual annotations to infer likely sharing recipients for photos captured on camera phones. Sarvas et al. developed a metadata creation system [28] to automate the creation of image content annotation by automatically leveraging available contextual metadata on mobile phones in order to use similarity processing algorithms for reusing shared metadata and images on a remote server. The system developed by Sarvas et al. can interact with mobile phone users during image capture to confirm and augment the system-supplied metadata. The VTT Technical Research Center proposed a video management system [14] comprised of a video server and a mobile camera-phone application named MobiCon. This system allows users to capture videos, annotate them with metadata, specify digital rights management (DRM) settings, upload the videos over the cellular network, and share them with others. EXTENT [7] was developed as an image annotation system that combines context and content information to annotate images with metadata that cannot be reliably inferred from either the context or the content alone. Table 1 summarizes recent works for annotating multimedia contents using high level features. These studies used various sensing information as well as visual features.

Many recent works have focused on the visual features of multimedia content. In this approach, it is assumed that there is already a database of multimedia content with accurate and suitable annotation. In order to create a new tag, we compare the visual similarity of the contents and select similar tags for new content. However, if there are no contents with similar features in our dataset, it cannot generate new tags for it.

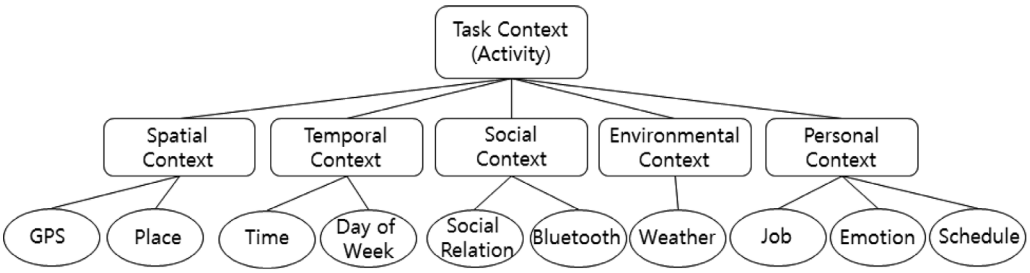

Fig. 1. Context model for activity recognition.

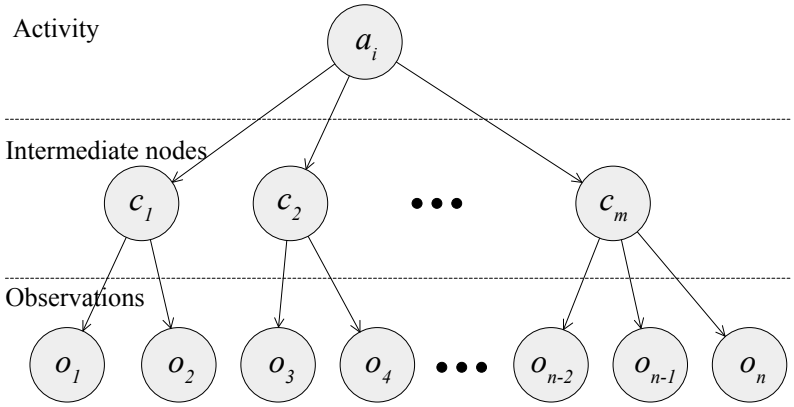

Fig. 2. Bayesian tree structure with intermediate nodes.

In this paper, we propose a method using tree-structured layered Bayesian network to annotate multimedia content automatically. This method can create semantic annotations for pictures without the comparison of similarity of the pictures. Its key features are:

- The use of raw data to generate a user's activity as high level annotation from a mobile phone.
- The use of a tree structured Bayesian network with intermediate nodes to reduce computational complexity and to increase the accuracy of generating semantic annotation.
- A two-layered Bayesian network structure to provide more accurate and broad-coverage annotation.

3. Activity inference using tree structured Bayesian network

The model has to satisfy two conditions to infer a user's activity on a mobile phone. First, the model can handle probabilistic values to cope with incomplete and uncertain real-world sensing information. Second, it is necessary to minimize the computational time due to the limitations of the mobile environment.

We used a Bayesian network as a model to handle uncertainty. The Bayesian network is a directed acyclic graphical model that was developed to represent probabilistic dependencies among random variables [16]. It relies on Bayes' rule Eq. (1) and conditional independence to estimate the distribution over variables.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1}$$

In this case, the nodes in the Bayesian network represent a set of observations (e.g., locations, day of week, etc.) denoted as $O = \{o_1, o_2, o_3, \ldots, o_n\}$, and corresponding activities (e.g., walking, studying, etc.) denoted as $A = \{a_1, a_2, a_3, \ldots, a_m\}$. These nodes, along with the connectivity structure imposed by directed edges between them, define the conditional probability distribution $P(A|O)$ over the target activity A. Equation (2) shows that a_i is a specific activity to recognize, and it can be inferred from observations.

$$P\,(A|O) \rightarrow P(a_i|o_1, o_2, o_3, \ldots, o_n) \tag{2}$$

In order to reduce the computational complexity of the Bayesian network, we used Bayesian network with a tree structure. A general Bayesian network takes a long time to perform the inference process because of the junction tree algorithm, which builds a tree structure from the original network structure [6]. If we can eliminate that process, it will decrease the time and memory required for inference dramatically.

In our previous work, we designed the tree structure by referring to Activity Theory, which describes a valuable framework for modeling human activities [21,22]. It can be used to build a context taxonomy with a tree structure with five context categories: environmental, personal, social, task, and spatio-temporal context [4] as shown in Fig. 1. The model allows us to reduce the complexity of inferring a user's activity by using intermediate nodes. However, it required a lot of effort and time to design the tree structure of a Bayesian network. Moreover, it cannot guarantee the optimal structure of a Bayesian network.

This paper presents a method to learn the tree structure with intermediate nodes from data. The basic Bayesian network includes three types of nodes: observation nodes, intermediate nodes, or activity nodes.

- *Observation nodes* are input for a Bayesian network that is extracted from pre-processed raw sensor data – GPS coordinates, call history, SMS history, etc. The raw data sequences in a time window of fixed length are transformed into observation.
- *Intermediate nodes* are used to increase the accuracy of the inference. If an optimal Bayesian network structure is not a tree structure because some evidence nodes are linked, we can make an optimal tree structure linking the evidence nodes by inserting an intermediate node.
- *Activity nodes* represent a user's activities which can be inferred from mobile sensing information such as temporal, spatial, social, personal, and environmental contexts.

It is possible to develop a Bayesian network that can maintain a tree structure and avoid the computational complexity by inserting intermediate nodes as illustrated in Fig. 2. The idea is to add intermediate nodes which reflect the influences among observation nodes. a is the activity node, and is also the root node in the tree structure. Each observation node o can have an intermediate parent i.

The joint distribution represented by the structure is defined as Eq. (3).

$$P(o_1, \ldots, o_n, a) = P(a) \prod_{i=1}^{n} P(o_i|i, a) \tag{3}$$

In order to determine inserting intermediate nodes among observation nodes effectively, we must know whether the observation nodes are dependent or not. We used conditional mutual information which is defined as Eq. (4).

$$I_P(O_i; O_j|A) = \sum_{o_i, o_j, a} P(o_i, o_j, a) \log \frac{P(o_i, o_j|a)}{P(o_i|a)P(o_j|a)} \tag{4}$$

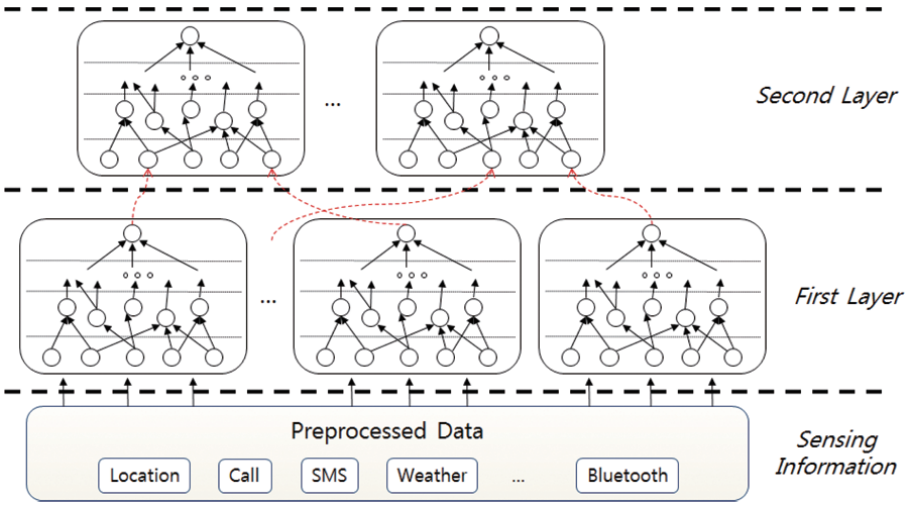

Fig. 3. Two-layered inference process using virtual linking technique.

where o_i and o_j are state values of observation nodes O_i and O_j, and a is a state value of an activity node A respectively. After I_P between each pair of observation nodes is computed to build a complete undirected weighted graph, the weight of an edge connecting O_i to O_j is set to I_P. Then we build a maximum weighted spanning tree from the weighted pairs of observation nodes. A maximum spanning tree is a spanning tree of a weighted graph having maximum weight. In the tree structure, observation nodes are linked with tree-like structure. The tree is divided into sub-trees with depth length 1, and some intermediate nodes are inserted to create a tree structure with depth length 2.

The whole process to create a Bayesian network with tree structure and intermediate nodes is as follows:

- *Computing local dependence* $I_P(O_i; O_j|A)$ which is conditional mutual information.
- *Building a complete undirected graph* G in which the nodes are observation nodes with the weight of an edge connecting O_i and O_j by $I_P(O_i; O_j|A)$.
- *Building a maximum weighted spanning tree* T from G.
- *Dividing the undirected tree* T into sub trees $t_1, t_2, t_3, \ldots$ with depth 1.
- *Constructing a tree structured Bayesian network* by inserting intermediate nodes for each sub tree t and adding a directional edge from an intermediate node to A.

4. Two-layered reasoning for activity inference

The tree structured Bayesian network in Section 3 is used to infer semantic annotation from mobile sensing information. The Bayesian model can effectively reduce the computational complexity of performing inference on a mobile phone. However, in some cases, the model may generate inappropriate annotations because it forces a tree structure to reduce the cost of inference.

In order to enhance the accuracy of the inferred annotation, the proposed method includes two layered inference. This second inference uses the result of the first layers as input evidence in order to cover broader annotation, as shown in Fig. 3. A virtual linking technique is utilized to better reflect the inferred evidence. This technique adds the virtual nodes and regulates their conditional probability values (CPVs)

Table 2
Contextual information from smartphone

Data	Attributes
Location	GPS coordinates (Latitude, Longitude), Speed, Altitude,
Music	MP3 title, start time, end time, singer
Photographs	Time, captured object
Call history	Start time, end time, person (friend/lover/family/others)
SMS history	Time, person (friend/lover/family/others)
Battery	Time, whether charging or not (Yes/No), charge level (%)

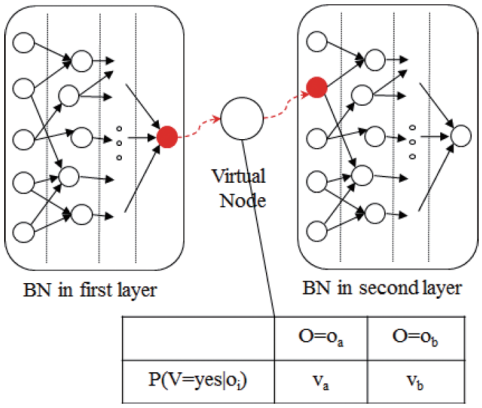

Fig. 4. Virtual linking method using a virtual node.

in order to apply the probability of the evidence [16]. The red dotted lines in Fig. 3 represent the virtual linking to transfer the inferred result into the Bayesian networks in the second layer.

The virtual linking method enables each Bayesian Network (BN) in the first layer to transfer its internal inference results to the BN in the upper layer. The BNs are virtually linked when BN_i in the first layer and BN_j in the second layer have a sharing node S. The inference results of BN_i, $\text{Bel}(S_i) = \{p_1, p_2, \ldots, p_n\}$, are transferred to BN_j by forcing the initial probability of a node S_j of BN_j in accordance with $\text{Bel}(S_i)$. The initial probability of a shared node is controlled by utilizing a virtual node.

A virtual node is an auxiliary node that is added to a network [16]. It sets virtual evidence, which has uncertainty associated with it. Virtual node V is linked to a target node T as a child node, a causality from T to S. The probability of V is set according to the target probability of T. The local inference result of the BN in the first layer can be passed to the upper layer by utilizing a virtual node. Thus, a key problem of the virtual linking method is calculating the probability of a virtual node given target probabilities of a shared node. In this case, it is assumed that a shared node has two states, as shown in Fig. 4.

When a sharing node S between two modules has only two states (*yes* or *no*), the probability of a virtual node $P(V = yes|s_i) = \{v_a, v_b\}$ is computed as shown in Eq. (5), enforcing a hard evidence on *yes*.

$$\frac{v_a}{v_b} = \frac{P(V = yes|S = s_a)P(S = s_b)}{P(V = yes|S = s_b)P(S = s_a)} \tag{5}$$

where $P(V = yes|S = s_a)$ and $P(V = yes|S = s_b)$ are the target probabilities of S, and $P(S = s_a)$ and $P(S = s_b)$ are the initial probabilities of S. The result of (X) is a just ratio rather than probability value, so it needs to be modulated by multiplying the modulation coefficient α until both have values ranging from 0 to 1.

The structures between the activities in the first layer and second layer are defined referring to General Social Survey (GSS) on Time Use, which was a statistical survey conducted by Statistics Canada to gather data on social trends in order to monitor changes in living conditions over time [11]. This survey divides all activities into ten categories, and each category is subdivided into several subcategories as the characteristics. It has a three-level hierarchy, and the total number of activities is 177. The main categories are “Paid work and related activities,” “Household work and related activities,” “Social support,

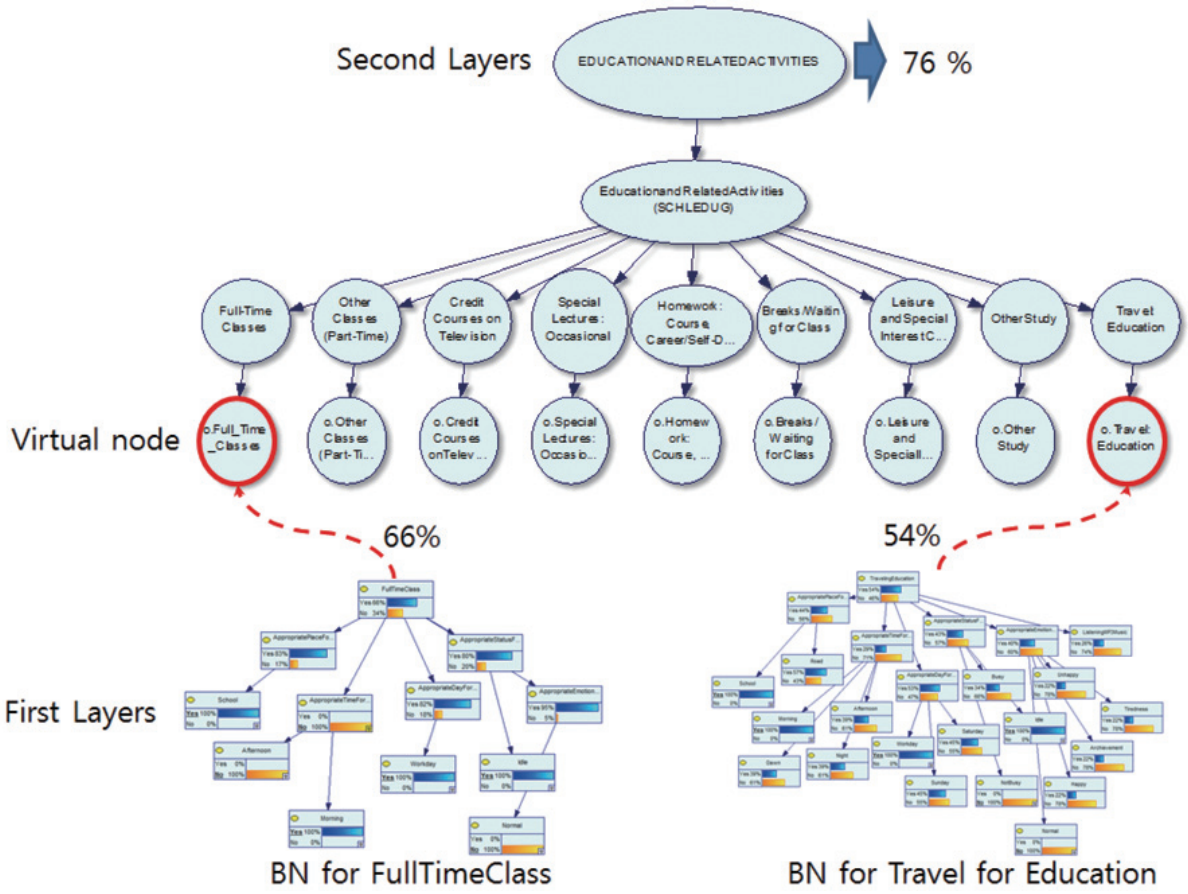

Fig. 5. Example of two-layered Bayesian inference.

civic and voluntary activity," "Education and related activities," "Socializing," "Television, reading and other passive leisure," "Sports, movies and other entertainment events," "Active leisure," and "Residual."

Figure 5 shows an example of two-layered inference. The probability of "Education and related activities" is calculated from the result of two BNs (FullTimeClass and Travel for Education) in the first layer.

5. Experimental result

5.1. Experimental setting

A Samsung m4650 (using Windows Mobile 6.0) smartphone is used to collect mobile data. Participants are 12 undergraduate students who are required to report their activities for a month. The whole system is developed with C# in Visual Studio 9 on a PC running Windows XP Professional. Smile/Genie [10] is used as an inference engine for the Bayesian network. The collected data types are summarized in Table 2.

Dimensionality reduction is necessary to perform inference on a mobile phone because the collected data are high dimensional. It is difficult to process the data because of the large amount of processing

Table 3
Comparison of BN accuracy

	Learned BN	Designed Tree BN	Learned Tree BN
ArmyTraining	0.904999004	0.768132204	0.925247295
Examination	0.753311589	0.795660036	0.773547228
Fetstival	0.718866667	0.795692831	0.721866667
FullTimeClass	0.782334156	0.750334971	0.787995355
GroupStudy	0.628391473	0.592837274	0.617248062
LookingForWork	0.767262063	0.775426969	0.789789389
MealsAtHome	0.819924137	0.913309	0.829773075
MealsAtSchool	0.631202691	0.678232583	0.631342865
MealWithFriends	0.807815417	0.289076	0.810816218
MusicConcerts	0.999138674	0.999337836	0.999138674
Naps	0.910187579	0.838838839	0.939418042
NightSleep	0.791156024	0.773599386	0.791156024
OperaBalletTheatre	0.81675948	0.8681635	0.880999735
ReadingBooks	0.777540678	0.502095	0.789877301
Relaxing	0.555734717	0.541547278	0.567722939
RepairServices	0.789190981	0.758679682	0.754708223
RestaurantMeals	0.709333693	0.653618031	0.71911249
ShoppingEverydayGoods	0.719146945	0.745892305	0.721353122
SocialGaterings	0.800757374	0.918025	0.800890247
SocializingAtBars	0.839157109	0.828775768	0.841157642
SocializingWithFriends	0.837156575	0.810194417	0.836156308
Sports	0.705152038	0.596292482	0.714978091
Study	0.665118605	0.656697819	0.665118605
TravelingEducation	0.651288446	0.65214948	0.6586312
WalkingAndRunning	0.593116554	0.994687	0.603322072
WatchingMovie	0.953394325	0.915715327	0.965990453

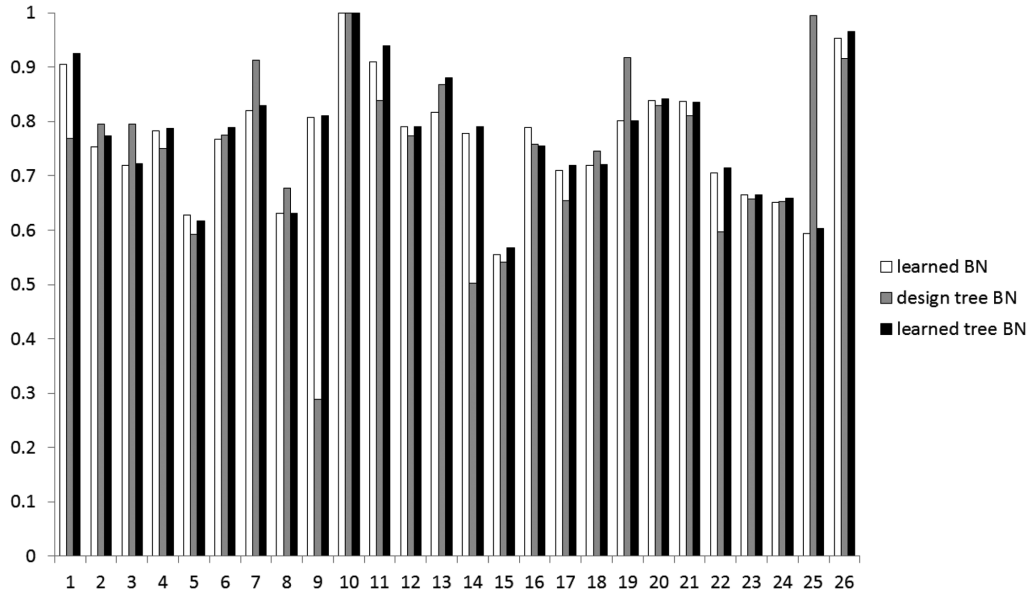

Fig. 6. Accuracy of the proposed method and learned BN.

time required and lack of memory. Therefore, we classified mobile data into three types (useful, partially useful, and useless) in order to annotate a user's activity. As useless context is ignored, the probabilistic models can be built more efficiently. In this case, the four kinds of approaches shown are used to analyze the usefulness of the data.

Using prior knowledge is an easy and effective method for analyzing and classifying useful data. However, it is difficult to discover novel and interesting patterns from a common sense standpoint. Weka and SQL query can be used to visualize data distribution. Weka provides visualization of specific properties among preprocessing functions. If a context is independent of a user's activity, the context is useless because it cannot reflect the user's activity. In addition, SQL query has the advantages in examining the distribution of a large amount of data. Bayesian network learning without prior knowledge helps us examine associations among data.

According to our analysis, time and location are the most important factors to estimate the probability of each activity. As a result, useful context is distinguished to annotate activities. Additionally, time window size can be determined to generate more accurate activity annotation. Activities can be divided into two types with temporal length: long-term activity and short-term activity. Long-term activity such as night sleep lasts is maintained for a long time. On the while, short-term activity such as sending an SMS often ends quickly. The purpose of the discrimination is to determine the time window size according to the duration of activities. The goal of our system is automatic activity annotation from daily situations. If we use too short of a window, false positives may be detected due to incomplete data. On the other hand, too large of a window may ignore short-term activities that occur quickly.

A long time window causes false negatives by recognizing short-term activities. A suitable time window size has to be determined in order to detect each activity. In our experiments, we empirically determined the window size to detect a person's activities by testing diverse window sizes for each activity.

5.2. Performance evaluation of BNs

In order to evaluate the proposed BN with tree structure and intermediate nodes, we compare the inferred activities by the proposed BN with the BN learned from the data. The performance of the BN can be evaluated by accuracy.

$$Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)} \tag{6}$$

where *FP* is the number of false positives, *FN* is the number of false negatives, *TP* is the number of true positives, and *TN* is the number of true negatives.

Figure 6 and Table 3 summarize the results of accuracy evaluation for each activity (1: ArmyTraining, 2: Examination, 3: Fetstival, 4: FullTimeClass, 5: GroupStudy, 6: LookingForWork, 7: MealsAtHome, 8: MealsAtSchool, 9: MealWithFriends, 10: MusicConcerts, 11: Naps, 12: NightSleep, 13: OperaBalletTheatre, 14: ReadingBooks, 15: Relaxing, 16: RepairServices, 17: RestaurantMeals, 18: ShoppingEverydayGoods, 19: SocialGaterings, 20: SocializingAtBars, 21: SocializingWithFriends, 22: Sports, 23: Study, 24: TravelingEducation, 25: WalkingAndRunning, 26: WatchingMovie). In most cases, the proposed tree structured BNs show better performance than learned BNs and designed (tree structured) BNs. On the while, there are some special cases that the designed BNs show best performance than learned BNs and the proposed BNs. If the data are not enough to find optimal structure of BNs, the designed BNs with fixed structure will show good performance thanks to domain knowledge. For instance, it is difficult to find optimal strcturue of BN for "SocialGaterings" which is a rare event, and the accuracy of the designed BN is better than a learned BN.

Table 4
Comparison of BN inference time

	Learned BN	Designed Tree BN	Learned Tree BN
ArmyTraining	0.2783	0.0059	0.0102
Examination	0.0908	0.0013	0.0032
Fetstival	0.0622	0.0012	0.002
FullTimeClass	0.0516	0.0039	0.0023
GroupStudy	0.3225	0.0108	0.002
LookingForWork	0.0394	0.0004	0.0018
MealsAtHome	0.342	0.0052	0.0019
MealsAtSchool	0.1212	0.0031	0.0017
MealWithFriends	1.03	0.0271	0.0022
MusicConcerts	0.0371	0.0005	0.0013
Naps	0.0819	0.0010	0.0018
NightSleep	0.2466	0.0058	0.0018
OperaBalletTheatre	0.0401	0.0014	0.0023
ReadingBooks	1.1062	0.0257	0.0021
Relaxing	0.0447	0.0004	0.0016
RepairServices	0.3099	0.0077	0.0019
RestaurantMeals	0.1974	0.0032	0.0019
ShoppingEverydayGoods	0.4824	0.0063	0.0023
SocialGaterings	0.6747	0.0075	0.0018
SocializingAtBars	0.3215	0.0096	0.0022
SocializingWithFriends	0.5057	0.0083	0.0017
Sports	0.0355	0.0001	0.0014
Study	0.1601	0.0038	0.0016
TravelingEducation	0.1202	0.0071	0.0017
WalkingAndRunning	1.0106	0.0088	0.0018
WatchingMovie	0.166	0.0067	0.0021

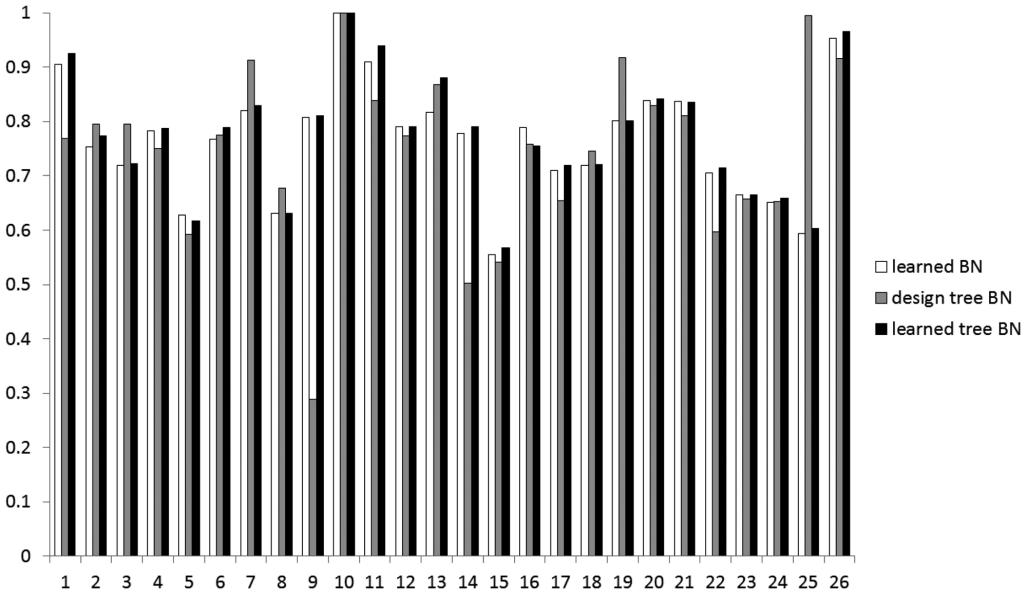

Fig. 7. Time to perform BN inference (unit: second).

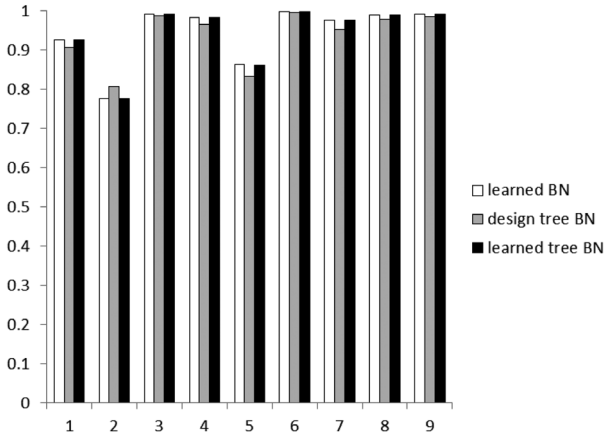

Fig. 8. Accuracy of BNs in second layer.

The inference time of a BN increases in proportion to the complexity of the BN. The tree structured BN shows a drastic reduction of the inference time, as shown in Fig. 7. It indicates that the proposed tree structured BN structure is suitable for use in mobile devices. Table 4 shows that the proposed BN structure is more efficient in inference time. In many cases, the inference time of the designed BNs is shortest but the accuracy of BNs cannot guarantee the good performance than learned BNs.

The proposed tree structured BNs in the first layer can reduce the computational cost effectively, but may be less accurate than designed tree structured models or learned models. In the second layer, the accuracy of the inferred annotation can be enhanced by providing broad-coverage annotations from the result of the first layers. Figure 8 shows the inferred results of BNs in the second layer (1: Active leisure, 2: Education and related activities, 3: Household work and related activities, 4: Passive leisure, 5: Entertainment events, 6: Socializing, 7: social support and voluntary activity, 8: sleep, meals and related activities, 9: Work and related activities).

5.3. Usability test

In order to evaluate the usefulness of the annotation, we built an interface for visualization of photographs and inferred annotation. The interface is implemented using C# and designed by referring to Head's evaluation interface criteria [2]. The overview of the interface is composed of pull-down menus, icon-menus, a list view for displaying data, tab button for photos, and buttons to annotate global activities and detail activities. Figure 9 illustrates the overview of the interface.

Once a user's name and date are selected with a pull-down menu, diverse information in the date is displayed in the main list view in chronological order. Location, status, emotion, and activities are highlighted with different colors in the list in order to improve usability. Each row in the list represents contextual information collected from a mobile device for 30 minutes. Figure 10 displays an example of the buttons, which display three global and detail activities which have a high probability.

Evaluation for usability refers to Head's criteria. The subjects for the assessment are 14 graduate students who understand the usage of the interface. The score of the evaluation is 1 to 5, where 1 is *never agree* and 5 is *strongly agree*. The graphs presented in Figs 11–13 show the results of the evaluation. The interface received high scores in terms of usability, task support, etc. This implies that users can apply semantic annotation through the interface.

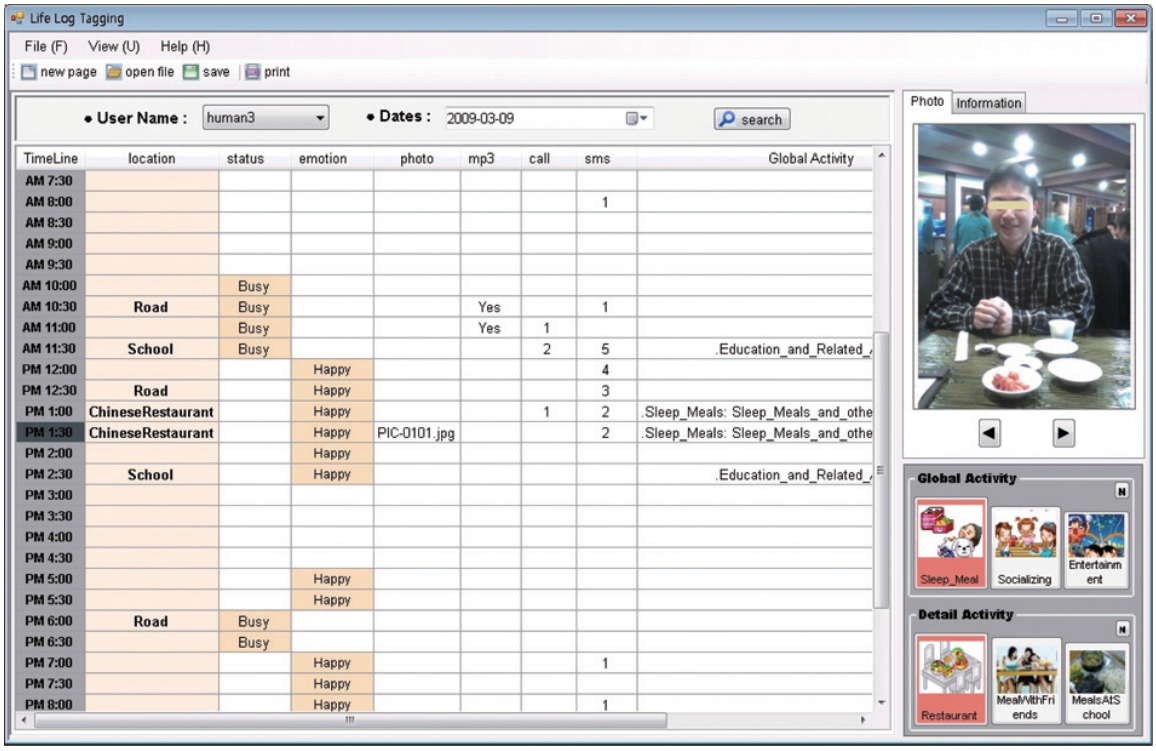

Fig. 9. Visualization interface overview.

Fig. 10. Visualization of activities.

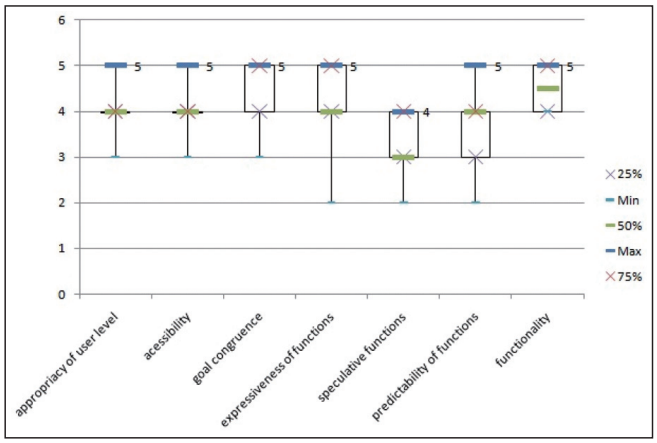

Fig. 11. Results of task support test (1).

6. Summary and discussion

In this paper, we propose tree-structured layered Bayesian networks to generate semantic annotations from mobile data. To create semantic annotations efficiently, the proposed method uses Bayesian probabilistic models with tree structures and intermediate nodes. The tree strictures allow inference process to skip junction tree algorithm which requires a large amount of time. The constrained structure may

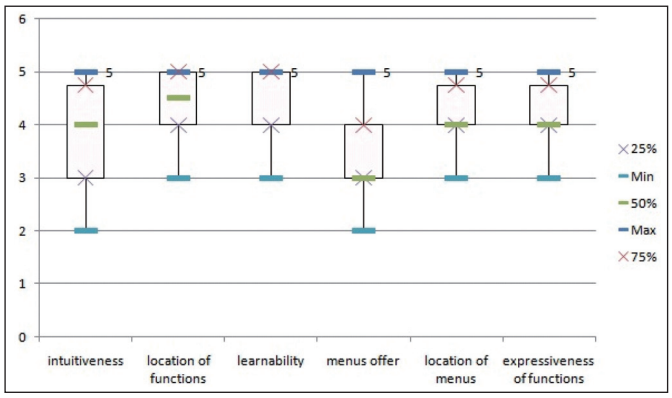

Fig. 12. Results of task support test (2).

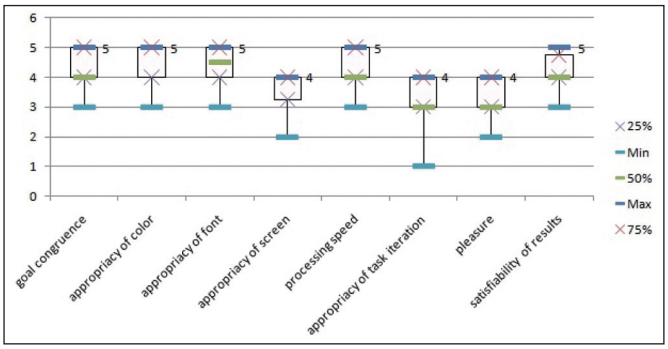

Fig. 13. Results of task support test (3).

cause accuracy degradation, but the intermediate nodes can reduce the degradation of the tree structure. This method is suitable to infer high level context in mobile environment because it is faster than other methods using junction tree algorithm. It also supports two-layered inference to enhance the accuracy of the annotation. The proposed model provides suitable annotation in the mobile environment using a small amount of inference time. The performance of the annotation is evaluated by undergraduate students during their daily activities. In addition, the usability of the annotations is evaluated by building an interface for effective visualization and annotating images.

In future research, online communities will be built to share the annotation and multimedia contents. The types of content will be extended to audio and video data as well as images. It will be necessary to consider more effective combinations and visualizations of annotations and multimedia contents. In addition, the integration of the learned and designed tree structured BNs will be considered. In some cases with insufficient data, domain knowledge helps us construct better BNs. The use of knowledge from human experts is effective for accurate inference.

References

[1] A. Badii, C. Lallah, M. Zhu and M. Crouch, Semi-automatic knowledge extraction, representation and context-sensitive intelligent retrieval of video content using collateral context modelling with scalable ontological networks, *Image Com-*

munication **24** (2009), 759–773.
[2] A.J. Head, *Design Wise: A Guide for Evaluating the Interface Design of Information Resources*, Information Today, Inc., 1999.
[3] A.K. Jain and A. Vailaya, Image retrieval using color and shape, *Pattern Recognition* **29** (1996) 1233–1244.
[4] A. Kofod-Petersen and J. Cassens, Using activity theory to model context awareness, in: *Second International Workshop Modeling and Retrieval of Context*, Aug 2005, pp. 1–17.
[5] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, Content-based image retrieval at the end of the early years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12)(2000), 1349–1380.
[6] C. Huang and A. Darwiche, Inference in belief networks: A procedural guide, *International Journal of Approximate Reasoning* **15** (1996) 225–263.
[7] C.-M. Tsai, A. Qamra, E.Y. Chang and Y.-F. Wang, Exten: Inferring image metadata from context and content, in: *IEEE International Conference on Multimedia and Expo (ICME)*, Jul 2005, pp. 1270–1273.
[8] D. Heckerman, A tutorial on learning with Bayesian networks, *Microsoft Research Technical Report*, *MSR-TR-95-06*, 1996.
[9] D. Zhang, M.M. Islam and G. Lu, A review on automatic image annotation techniques, *Pattern Recognition* **45**(1) (2012), 346–362.
[10] GeNie & Smile Web Site.
[11] H.-S. Park and S.-B. Cho, Predicting user activities in the sequence of mobile context for ambient intelligence environment using dynamic bayesian network, in: *2nd International Conference on Agents and Artificial Intelligence (ICAART)*, Jan 2010, pp. 311–316.
[12] H. Tamura and N. Yokoya, Image database systems: A survey, *Pattern Recognition* **17**(1) (1984), 29–43.
[13] I. Ivanov, P. Vajda, L. Jong-Seok and T. Ebrahimi, In tags we trust: Trust modeling in social tagging of multimedia content, *IEEE Signal Processing Magazine* **29** (2012), 98–107.
[14] J. Lahti, M. Palola, J. Korva, U. Westermann, K. Pentikousis and P. Pietarila, A mobile phone-based context-aware video management application, in: *Proceedings of the SPIE, Multimedia on Mobile Devices II*, Feb 2006, pp. 204–215.
[15] J. Tesic, Metadata practices for consumer photos, *IEEE MultiMedia* **12**(3) (2005), 86–92.
[16] K.-S. Hwang and S.-B. Cho, Landmark detection from mobile life log using a modular Bayesian network model, *Expert Systems with Applications* **36** (2009), 12065–12076.
[17] L. Dong, W. Meng, H. Xian-Sheng and Z. Hong-Jiang, Semi-automatic tagging of photo albums via exemplar selection and tag inference, *IEEE Transactions on Multimedia* **13**(2011), 82–91.
[18] M. Davis, N.V. House, J. Towle, S. King, S. Ahern, C. Burgener, D. Perkel, M. Finn, V. Viswanathan and M. Rothenberg, MMM2: mobile media metadata for media sharing, in: *CHI '05 extended abstracts on Human factors in computing systems*, Apr 2005, pp. 1335–1338.
[19] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman and G.J.F. Jones, Automatic tagging and geotagging in video collections and communities, in: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, Apr 2011, pp. 1–8.
[20] M.M. Islam, Z. Dengsheng and L. Guojun, Automatic categorization of image regions using dominant color based vector quantization, in: *Digital Image Computing: Techniques and Applications (DICTA)*, Dec 2008, pp. 191–198.
[21] M. Kaenampornpan and E.O. Neill, Integrating history and activity theory in context aware system design, *Cognitive Science Research Paper*, University of Sussex CSRP, **577** (2005), 87.
[22] M. Kaenampornpan and E. O'neill, Modelling context: An activity theory approach, in: *Ambient Intelligence: Second European Symposium (EUSAI)*, Nov 2004, pp. 367–374.
[23] M. Szummer and R.W. Picard, Indoor-outdoor image classification, in: *IEEE International Workshop on Content-Based Access of Image and Video Database*, Jan 1998, pp. 42–51.
[24] M.T. Yu and M.M. Sein, Automatic image captioning system using integration of N-cut and color-based segmentation method, in: *Proceedings of SICE Annual Conference (SICE)*, Sep 2011, pp. 28–31.
[25] N. Sawant, J. Li and J.Z. Wang, Automatic image semantic interpretation using social action and tagging data, *Multimedia Tools and Applications* **51**(1) (2011), 213–246.
[26] P. Kelm, S. Schmiedeke, K. Clver and T. Sikora, Automatic Geo-referencing of flickr videos, *NEM Summit* (Sep 2011).
[27] R. Jesus, A.J. Abrantes and N. Correia, Methods for automatic and assisted image annotation, *Multimedia Tools and Applications* **55** (2011), 7–26.
[28] R. Sarvas, E. Herrarte, A. Wilhelm and M. Davis, Metadata creation system for mobile images, in: *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, Jun 2004, pp. 36–48.
[29] R. Zhao and W.I. Grosky, From features to semantics: Some preliminary results, in: *IEEE International Conference on Multimedia and Expo (ICME)*, Jul 2000, pp. 679–682.
[30] S.-H. Li, C.-M. Gao and H.-W. Pan, Automatic image tagging based on regions of interest, in: *Proceedings of the Third international conference on Artificial intelligence and computational intelligence – Volume Part I*, Sep 2011, pp. 300–307.

[31] T. Volkmer, J. Thom and S. Tahaghoghi, Modeling human judgment of digital imagery for multimedia retrieval, *IEEE Transactions on Multimedia* **9**(5) (2007), 967–974.
[32] U.B. Kjærulff and A.L. Madsen, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, Springer, 2008.
[33] V. Könönen, J. Mäntyjärvi, H. Similä, J. Pärkkä and M. Ermes, Automatic feature selection for context recognition in mobile devices, *Pervasive and Mobile Computing* **6** (2010), 181–197.
[34] X. Sevillano, T. Piatrik, K. Chandramouli, Q. Zhang and E. Izquierdoy, Geo-tagging online videos using semantic expansion and visual analysis, in: *13th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, May 2012, pp. 1–4.
[35] X. Zhang, Z. Huang, H.T. Shen, Y. Yang and Z. Li, Automatic tagging by exploring tag information capability and correlation, *World Wide Web* **15**(3) (2012), 233–256.
[36] X. Zhang, X. Zhao, Z. Li, J. Xia, R. Jain and W. Chao, Social image tagging using graph-based reinforcement on multi-type interrelated objects, *Signal Processing* (In Press).
[37] Y. Engeström, R. Miettinen and R.-L. Punamäki, *Perspectives on Activity Theory (Learning in Doing: Social, Cognitive and Computational Perspectives)*, Cambridge University Press, 1999.
[38] Y.-S. Lee and S.-B. Cho, Automatic image tagging using two-layered Bayesian networks and mobile data from smart phones, in: *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia (MoMM)*, Dec 2012, pp. 39–46.
[39] Y. Yang, Z. Huang, H.T. Shen and X. Zhou, Mining multi-tag association for image tagging, *World Wide Web* **14**(2) (2011), 133–156.

Young-Seol Lee received the B.S. and M.S. degrees in computer science from Yonsei University, Seoul, Korea, in 2006 and 2008, respectively. Since 2008, he is a Ph. D. student in the Department of Computer Science, Yonsei University. His research interests include mobile inference, probabilistic modeling, smart phones, and Bayesian network.

Sung-Bae Cho received the B.S. degree in computer science from Yonsei University, Seoul, Korea, in 1988 and the M.S. and Ph.D. degrees in computer science from KAIST, Taejeon, Korea, in 1990 and 1993, respectively. Since 1995, he has been an associate professor in the Department of Computer Science, Yonsei University. His research interests include neural networks, pattern recognition, intelligent man-machine interfaces, evolutionary computation, and artificial life.