A Language-Based Approach to Indexing Heterogeneous Multimedia Lifelog

Peng-Wen Cheng, Snehal Chennuru, Senaka Buthpitiya, Ying Zhang
Carnegie Mellon Silicon Valley
NASA Ames Research Park
Moffett Feild, CA, USA
{pengwen.chen, snehal.chennuru, senaka.buthpitiya, joy.zhang}@sv.cmu.edu

ABSTRACT

Lifelog systems, inspired by Vannevar Bush's concept of "MEMory EXtenders" (MEMEX), are capable of storing a person's lifetime experience as a multimedia database. Despite such systems' huge potential for improving people's everyday life, there are major challenges that need to be addressed to make such systems practical. One of them is how to index the inherently large and heterogeneous lifelog data so that a person can efficiently retrieve the log segments that are of interest. In this paper, we present a novel approach to indexing lifelogs using activity language. By quantizing the heterogeneous high dimensional sensory data into text representation, we are able to apply statistical natural language processing techniques to index, recognize, segment, cluster, retrieve, and infer high-level semantic meanings of the collected lifelogs. Based on this indexing approach, our lifelog system supports easy retrieval of log segments representing past similar activities and generation of salient summaries serving as overviews of segments.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

activity languages

Keywords

Indexing Heterogeneous Multimedia, Lifelogs,

1. INTRODUCTION

Our life experiences are fleeting and immaterial. Memory, our ability to store, retain and recall information, is crucial for day to day life. However memories, unlike digital recordings, fade and are forgotten. Memory problems can be very serious for people who have suffered brain injuries or have memory diseases like the Alzheimer's. As of September 2009, more than 35 million people around the world have been diagnosed with Alzheimer's disease for which episodic memory impairment (EMI) is the main sympton [12]. And according to [3], the prevalence of Alzheimer's is expected to reach approximately 107 million people by 2050.

In 1945, Vannevar Bush proposed a prototype computer system named MEMEX, whose main functionality is to share people's burden in remembering. Such a system has great potential in a variety of applications and is particularly useful for people who suffer from EMI. To assist the ever growing population of EMI sufferers, Bush's MEMEX concept seems to be a promising solution and thus gives birth to the personal lifelog research area. To fulfill Bush's vision, a personal lifelog system must be able to 1) store a large volume of personal multimedia data and 2) efficiently retrieve the relevant data based on user requests.

Technologies today make it possible to capture life experiences in digital format. Advancements in mobile sensors and prevailing computing allows LifeLog systems to record almost every aspect of a person's life [7]. While recording and storing all sensor information in a database poses an engineering challenge, indexing them is key to making lifelog systems useful. An index allows retrieving important pieces of memory from the lifelogs. We cannot expect users to annotate and create an entire index. Yet understanding the semantics of images, audio and video is still an open research problem.

In this paper, we present a novel approach to indexing lifelog data using activity languages. In this ongoing research, we convert the ambulatory sensor inputs such as accelerometer readings into an *activity language* and using it as the main index of multimedia lifelogs. This approach enables the use of statistical natural language processing methods to index, retrieve, cluster, summarize lifelogs which are not easy or possible for images, audio and video information.

The rest of the paper is organized as follows. We first describe our LifeLogger system in Section 2. Section 3 introduces the concept of "Activity as Language" where we quantize the sensory input and convert it into a text representation to interpret the underlying meaning of lifelogs. In Section 3.2, 3.3, 3.4 and 4, we present algorithms and preliminary results on activity recognition, similar activity retrieval and geo-trace pattern recognition. Finally, we conclude our findings and discuss the future works.

2. OUR LIFELOG SYSTEM

Our lifelog system consists of four major parts, namely,

ICMI-MLMI '10, November 8-12, 2010, Beijing, China.

Figure 1: Sensor helmet for collecting activity data.

the Lifelogger Mobile Client, the Lifelogger Server Application, the Similar Activity Retrieval Service, and the Geo Trace Service.

2.1 Lifelogger Mobile Client

Thanks to the rapid advancement of commercial mobile devices, we are able to build our Lifelogger client devices from off-the-shelf products. Our LifeLogger Mobile Client is a helmet mounted with two Nokia N95 phones (Figure 1). The client records various types of sensory data including GPS coordinates for outdoor locations, gyroscope readings for rotation, microphone recordings for sound, camera images for pictures, and WiFi signal strength for indoor locationing. All data is collected with timestamps in JSON format and transmitted to the LifeLogger Server Application using the HTTP protocol via wireless connections.

2.2 Lifelogger Application Server

The LifeLogger Application Server is responsible for storing, pre-processing, and modeling recorded data as an activity language. It is also in charge of supplying the user interface for users to interact with their lifelogs. This server-side module is implemented as a web application so that users can access it easily through web browsers. The application provides an easy-to-use interface for end users to browse, annotate, and search their lifelogs. To make it easy for users to recall their past living experiences, we fit the collected images, audio, and GPS location data into one screen (Figure 2) to let users intuitively combine these memory clues. Moreover, since all sensory data are associated with synchronized timestamps, users can navigate through the data set by simply dragging the time-line at the center of the screen, and the three types of data would be updated simultaneously. Users can also annotate a selected segment of lifelog by providing a short text description. Such text descriptions will be used to learn the association between natural language queries and the stored lifelog data. If the user is interested in a specific part of the lifelog and would like to find all lifelog segments that are similar to it, all the user need to do is first select that specific lifelog part using the time-line control and then click the "Find Similar Activities" button beside it. The similar lifelog segments will be returned and listed in a table at the bottom of the screen (Figure 2) and are sorted based on their relevance.

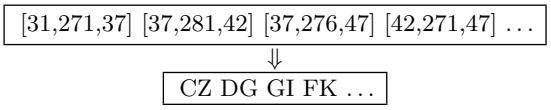

Figure 3: An example of quantizing accelerometer readings to activity language representation.

2.3 Similar Activity Retrieval and Geo Trace Services

The LifeLog system supports similar activity retrieval on two levels of time granularity. At the micro or more fine-grained level, accelerometer readings are used to identify the user's physical activity at a given moment. And at the macro level, i.e., at a higher granularity, GPS coordinates are used to identify patterns and in the user's daily movements. Similar activity detection at higher granularity levels increase in importance when dealing large collections of data pertaining long periods of a person's life. The importance of being able to detect similar activities at high granularity is seen by two features that it enables in lifelog systems,

- Automatic classification of activities that are similar in nature to previously manually labeled activities.
- Irregular activity detection, i.e., detecting activities that may be of special interest to the user.

3. LANGUAGE-BASED INDEXING

3.1 Main Idea

One of challenges in indexing, retrieving and interpreting lifelogs is that a lifelog is a collection of heterogeneous sensory information and each sensory data type requires a special method to process the raw input. In most existing lifelog applications, raw input from sensors is classified into predefined classes by trained classifiers for further processing, usually limiting the scope of the systems to those predefined activities.

In this paper, we propose a novel method of representing sensory input as "activity language" through quantizing raw sensory input. Here we use *motion* information as an example. To record users' motion, we use 3-axis accelerometers. For the accelerometers used in our experiments, the raw readings for each axis ranges from -360 to 360 which translates into 373,248,000 different (a_x, a_y, a_z) combinations. We quantize the raw accelerometer reading into V groups using the K-Means clustering algorithm. Once the clustering algorithm converges resulting in V cluster centroids, we label each cluster arbitrarily (e.g. - "D", "GC" and "DFR"). We label each accelerometer sample by assigning it the nearest cluster centroid's label, thus converting the ambulatory activity into "activity text" (Figure 3).

The main benefits from quantizing the raw sensory input into an "activity language" representation are,

- Dimension reduction of sensory input - Higher dimensional input data is reduced into a single dimension reducing the computation complexity. In the case of accelerometer readings, the original 3-dimension input of (a_x, a_y, a_z) is now reduced to one dimension.
- Efficient indexing and searching of lifelogs - Searching an indexed text corpus is much easier than searching

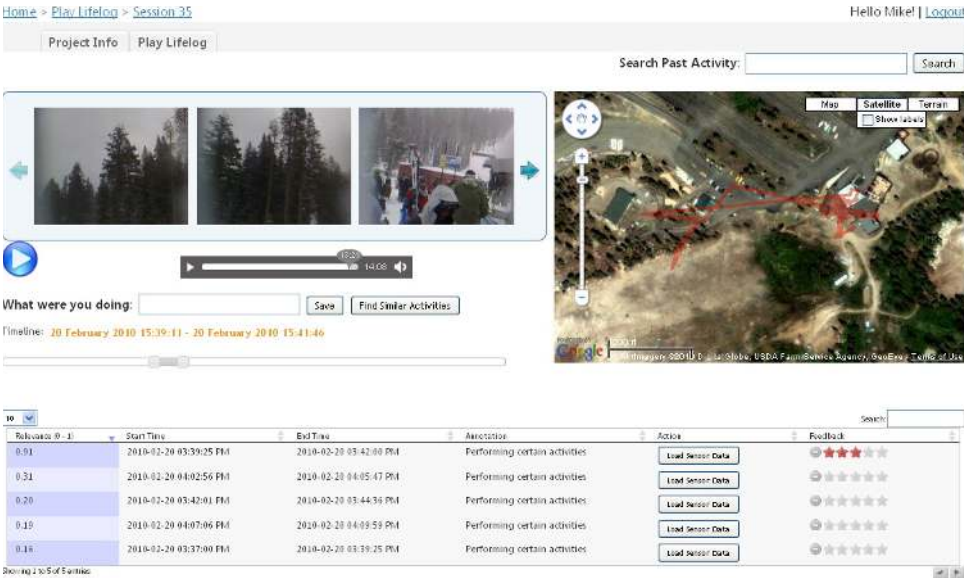

Figure 2: Web interface of browsing/searching/annotating life logs.

a database of real numbers. Compared to the infinite real-number space, the limited "vocabulary" size of the text representation allows the search algorithm to be much more efficient. Index and search algorithms such as the Inverted Index and Suffix Arrays [11] developed for strings can be applied on the "activity language" representations. This is more straight forward than searching the lifelogs based on the the cosine or Euclidean distances between the query and the logged activities.

- Uniformed representation of heterogeneous input - By converting various sensory input into a single type of "activity language" representation, we can develop and apply the same "activity language processing" algorithms on different types of data. In this paper, we demonstrate that motion from accelerometer readings and geo-trace from GPS recordings can all be processed by n-gram language model algorithms.

We call this representation "activity language" based on the analogy between human activity and natural languages. The similarity between human activity and language has been articulated by Burke [4] and Wertsh [19]. Based on the "principle of language as action", natural languages and human activities indeed share some important properties. For instance, they are both "mediational means" or tools by which we achieve our ends. Additionally, they both exhibit structure and satisfy "grammars".

Table 1 illustrates that human ambulatory activities share a lot in common with natural languages at all levels. Anatomical limitations of the human body allows only a limited set of *atomic movements*(e.g. - "turn upper body left" is possible, "jump up at 10g acceleration" is not possible). The set of possible atomic movements form the vocabulary of the activity language. A sequence of atomic movements performed in a meaningful order creates a *movement* such as an *action* of "standing up". *Actions* such as "climbing up stairs" are created by performing actions in a particular order, analogous

Natural Language	Activity Language	Example
Word	Atomic Movement	Turn upper body left
Phrase	Movement	Stand up
Sentence	Action	Climb up stairs
Paragraph	Activity	Enter building, climb up stairs and walk into office
Document	Event	Left home and cycled to campus, arrived at my office on 2nd floor

Table 1: Activity as language at different levels.

to creating a sentence with words. A sequence of actions builds up an *activity*. The higher level concept *event* is composed of a series of activities, as documents are composed of sentences.

3.2 Activity Recognition

Before applying the activity language concept in our lifelog system, we performed a series of predefined-activity recognition experiments to justify the benefits of modeling human activities as a language. In our approach, we view the labeled data for each activity a_i as the training corpus and train a smoothed n-gram language model over the converted activity language text using the SRI language model toolkit [17]. For each testing "activity sentence" t, we input it to each of the pre-built language models to calculate the probability of t being generated by activity a_i and we predict the activity of the testing sentence to be i^* such that,

$$i^* = \arg\max_i P(t|a_i) \quad (1)$$

An issue of using language models for activity recognition is that language model probabilities are not directly comparable if their respective training data have different

	Predicted Activity					
	walking		running		cycling	
walking	94%	95%	3%	1%	3%	4%
running	6%	4%	92%	94%	2%	2%
cycling	8%	2%	0%	0%	92%	98%

Table 2: Classification accuracy on corpus with vocabulary = 100 and with vocabulary = 200 respectively.

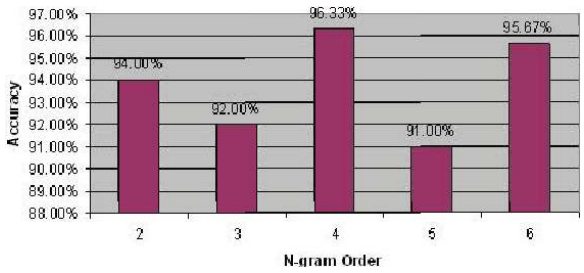

Figure 4: Recognition accuracy vs. n-gram order.

vocabulary sizes. To solve this problem, each training data set is augmented with a universal vocabulary list built from all training data sets. As a result, all our activity language models have the same vocabulary size, and their generated probabilities are comparable.

Our preliminary results of using smoothed n-gram language model for activity recognition demonstrated an average accuracy rate of 94% in distinguishing among basic activities such as walking, running, and cycling. Table 2 compares the recognition accuracy of language models trained over corpora of vocabulary size 100 and 200 words. With a larger vocabulary size (i.e. - more atomic movement types) the activity language has greater discriminative power to differentiate human activities. Figure 4 shows the average activity recognition accuracy vs. the order of n in language model training. Overall, for this basic activity recognition task, the order of history does not play a significant role. The promising results of these experiments increased our confidence in using the activity language to improve indexing in lifelog systems.

3.3 Similar Activity Retrieval

In many cases, we want to find out information about activities that are not predefined such as "how many tennis games did I play in the past two months?" or "how much time did I spend sitting in front of TV last week?". It is not possible to enumerate all possible activities and train Hidden Markov Models ahead of time to answer such questions. In our approach, we convert the lifelogs, in particular, the main indexing sensory information into a text representation. This allows us to apply Information Retrieval techniques to "retrieve' relevant activities from the past logs to answer user queries.

In our implementation, a user can select a segment from his/her lifelogs on the web interface and indicate that he/she may want to find similar activities from the past logs. The highlighted segment does not need to be annotated by natural language descriptions such as "playing tennis". The

n	n-gram	in P	in Q	min	Prec_n
1	NB	4	4	4	
	P	6	6	6	10/10=1.0
2	NB NB	2	1	1	
	NB P	1	3	1	
	P P	5	3	3	6/9 = 0.67
	P NB	1	2	1	
3	NB NB P	1	1	1	
	NB P P	1	2	1	
	P NB NB	1	1	1	5/8 = 0.63
	P P NB	1	1	1	
	P P P	4	1	1	
4	NB NB P P	1	0	0	
	NB P P P	1	1	1	
	P P NB NB	1	1	1	2/7 = 0.29
	P P P NB	1	0	0	
	P P P P	3	0	0	
5	NB NB P P P	1	0	0	
	NB P P P P	1	0	0	
	P P P NB NB	1	0	0	0/6 = 0.0
	P P P P NB	1	0	0	
	P P P P P	2	0	0	
					$S(P,Q) = 0.52$

Table 3: Calculating the similarity between two activity sentences using averaged n-gram precision.

system will search past lifelogs and return the most relevant segments for the user to review.

The key here is to calculate "similarity" between two lifelog segments. Inspired by the BLEU metric [15] where averaged n-gram precision is used to measure the similarity between a machine translation hypothesis and human generated reference translations, we use *averaged n-gram precision* to estimate the similarity between two lifelog segments.

Assuming that P and Q are two activity language sentences of the same length l. P is the sequence of P_1, P_2, ..., P_L and Q is the sequence of Q_1, Q_2, ..., Q_L. Denote the *similarity* between P and Q as $S(P,Q)$. Define the n-gram precision between P and Q as $\text{Prec}_n(P,Q) =$

$$\frac{\sum_{\tilde{p}\in\{\text{All } n\text{-gram types in } P\}} \min(freq(\tilde{p},P), freq(\tilde{p},Q))}{\sum_{\tilde{p}\in\{\text{All } n\text{-gram types in } P\}} freq(\tilde{p},P)}, \quad (2)$$

and the similarity between P and Q is defined as:

$$S(P,Q) = \frac{1}{N}\sum_{n=1}^{N} \text{Prec}_n(P,Q) \quad (3)$$

$\text{Prec}_n(P,Q)$ calculates the percentage of n-grams in P that can also be found in Q and $S(P,Q)$ averages the precision over 1-gram, 2-gram and up to N-gram. In our experiments, we empirically set $N = 5$.

Table 3 shows an example of calculating the similarity between activity sentence P ("NB NB P P P P P P NB NB") and Q ("NB P P NB NB P NB P P P").

Given a query sentence of l words, we assume that similar activities in the lifelog should also be of length l. This assumption makes the retrieval algorithm easier to implement as varied length activity retrieval would require activity segmentation. For a lifelog with G words, there are $G - l$ different strings of l words long. In our current setting, a 24 hours lifelog contains about 200 million activity words. Cal-

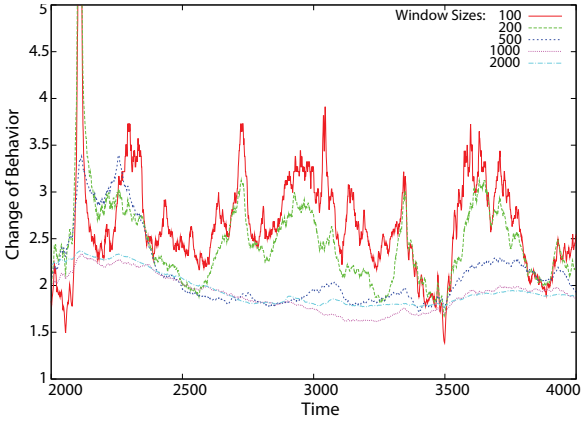

Figure 5: Activity changes calculated by different size of sliding windows.

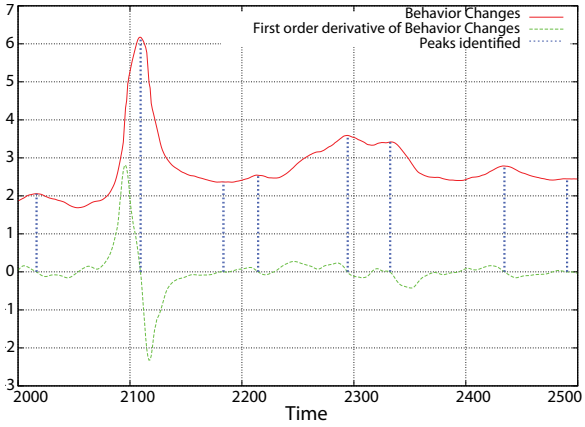

Figure 6: Identifying peaks in the behavior-change-curve.

culating the similarity between each of the $G-l$ strings with the query can be computationally expensive. To speed up retrieval, we use suffix arrays to pre-select strings in the corpus that have high order n-gram matches with the query and calculate $S(P,Q)$ scores for those strings only. The observation is that if a string in the lifelog is similar to the query, then it should have many high order n-grams matched with those in the query string.

Top R similar activity segments are returned to the user on the web interface (as shown in lower panel in Figure 2). The user can load each segment to "play" the corresponding lifelog and for our ongoing experiments evaluate if the segment is truly "similar" to the query.

3.4 Hierarchical Segmentation of Lifelogs

Lifelog records a user's daily life as a continuous sequence of sensory data. After converting the sensory data to activity language text, a lifelog is now a long string of text. Just as we need punctuations, sentence boundaries and paragraph boundaries in written text, it would make lifelogs more readable if we could automatically segment the data based on user activities.

The underlining assumption of our segmentation algorithm is that when a user switches his/her activity at time t, the similarity between string $[t-w,t-1]$ and $[t,t+w]$ should be much lower than if t is inside the same activity for a window of size w. For a window size w, we define the "change of activity" at time t as:

$$H(t,w) = -log(0.00001 + S([t-w,t-1],[t,t+w-1])). \quad (4)$$

The higher the value of $H(t,w)$, the more likely a change of activity at time t would have occurred. Figure 5 shows the H value at each data point given different window sizes for segmenting a lifelog.

It can be noticed that: (1) peaks of activity change identified by larger windows are also peaks identified by smaller windows but not vice versa; and (2) activity changes over larger windows are smoother than smaller windows. Intuitively, larger window sizes capture changes of larger-scale activities whereas smaller window captures changes of smaller activities. Based on this finding, we first segment the lifelog data using larger window sizes and then recursively segment the data using smaller windows. This results in a hierarchical segmentation of lifelogs which allows user to efficiently

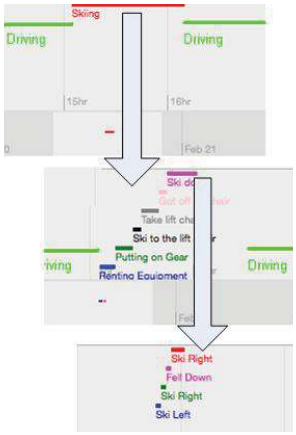

Figure 7: Hierarchically segmented lifelog. Activity boundaries detected automatically by the system and descriptions are added by the user

browse through the lifelog instead of playing the whole lifelogs. (Figure 7).

4. GEO-TRACE PATTERN RECOGNITION

To detect similar activities at macro time scales we use a language model based approach. Training a language model on GPS coordinates to model a user's daily movement requires the conversion of the GPS coordinates into "words". The successive GPS coordinates, logged by a user's position, form "sentences" relating to his daily routine. To perform the conversion of GPS coordinates into "words" in the LifeLog system two approaches were considered, 1) dividing the spacial domain covered by the coordinate system into cells and assigning each cell a unique word, or 2) to perform clustering over logged GPS coordinates and assign each cluster centroid a unique word. The first approach has been implemented and tested showing promising results. The second approach is also being considered as an improvement for the existing system.

In our implementation of the first approach, division of spacial regions covered by the GPS coordinate system is performed by considering only the longitudinal and latitudinal

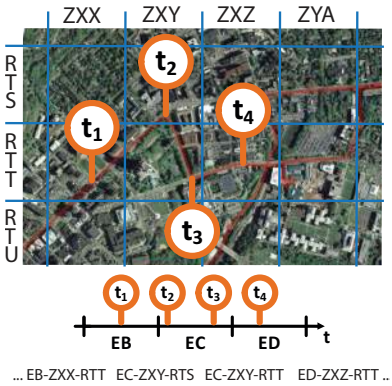

Figure 8: Quantizing GPS coordinates with time information into text representation.

values and ignoring the altitude values. The longitudinal-latitudinal coordinates are separated into regions of roughly equally size across the surface of the globe, creating a two dimensional grid which is augmented with a third dimension by incorporating time (time-of-day), forming a lattice. Therefore the "word" of a GPS sample is formed using longitude, latitude, and timestamp of the reading (Figure 8).

Various aspects of the detection mechanism were tested in a series of experiments described below. As experimental data the location tracks of ten people collected over a period of four weeks were used. This constitutes 40 week-long location tracks. For each test the system is trained using a week-long location track of a person. The testing data is created from the remaining 39 week-long tracks, but dividing each track into segments. Each of the segments are presented to the model learned by the system, which classifies whether or not it falls into the daily routine of the person. Ground truth for the experiments were established using the LifeLogger system to manually classify each segment of the location tracks using the visual, topographical and audio hints provided by the systems interface.

In the first experiment we maintain the granularity of spacial and time divisions, and segment sizes constant. In this experiment the language models are varied, using respectively a unigram model, a bigram model, up to a 10-gram model. For each model the prediction accuracy is measured. The results from this experiment are shown in Table 4 and Figure 9.

The results show that prediction accuracy increases as the number of past states considered by the language model increase. The increasing trend levels-off after the number of past states considered is greater than five.

In the second experiment we employ a 10-gram language model while varying the granularity of spacial division. The segment sizes and time divisions are kept constant in this experiment. For each granularity level, vocabulary size and measured prediction accuracy are noted. From the results of this experiment (depicted in Figure 10) it can be seen that accuracy is very dependent on the granularity level. Extremely high granularity levels could push the algorithm into a hyper sensitive state where noise from the GPS device and minute changes in a person's routine heavily affects the results. While low granularity levels cause the algorithm to loose important features required for an effective classification.

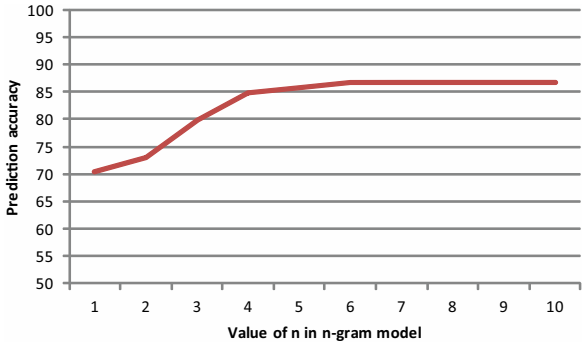

Figure 9: Variation of prediction accuracy with various language models.

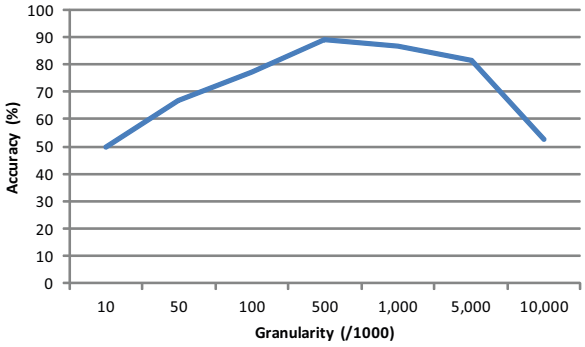

Figure 10: Variation of prediction accuracy with granularity of spacial "word" grid.

In the third experiment we employ a 10-gram language model and maintain fixed spacial and temporal granularity. Spacial granularity is maintained at one million divisions each for longitudinal and latitudinal coordinates. Temporal granularity is maintained at 1 hour divisions. In this experiment we vary the testing segment size from 1 hour to 12 hours while observing the prediction accuracy. The results from this experiment shown in table 6 and figure 11.

As the size of the segment provided for classification increases in length, thereby increasing the available features for classification, the accuracy of the predictions increase.

5. RELATED WORK

5.1 Activity Recognition

There have been several techniques for recognizing or distinguishing basic human activities. They can be categorized into two flavors: heuristic threshold-based classifiers and pattern recognition techniques such as decision trees, nearest neighbor, Naive Bayes, support vector machines (SVM), neural networks, Hidden Markov Models (HMM) and Gaussian mixture models [13]. For recognizing high-level human activities, several attempts had been made in [1, 16]. Among these techniques, the most popular ones we see so far are those based on HMM. Single-layer HMM-based approaches classify the input sensory information into one of the predefined activities such as walking, running, and standing.

n	1	2	3	4	5	6	7	8	9	10
Prediction accuracy (%)	70.34	73.04	79.95	84.82	85.71	86.61	86.61	86.61	86.61	86.61

Table 4: Variation of prediction accuracy with various language models.

Approx. Cell Size (m x m)	4000	800	400	80	40	8	4
Accuracy (%)	50.34	66.47	77.28	88.89	86.61	81.41	52.47

Table 5: Variation of prediction accuracy with granularity of spacial "word" grid

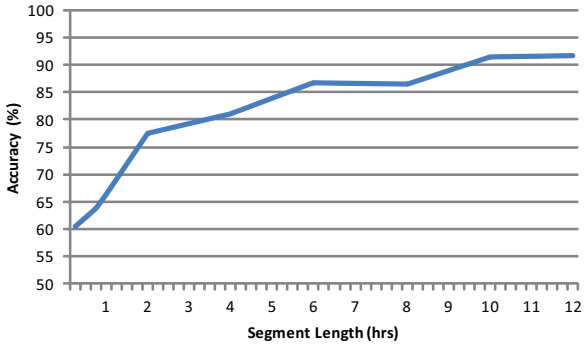

Figure 11: Variation of prediction accuracy with various language models.

However, since HMM assumes the first order Markov chain in the state space and usually does not consider the inherent "grammar" or "structure" of human activities, the activities that can be recognized by HMM are limited to those predefined in the training data, which as a result limits HMM's application in people-centric computing.

Layered probabilistic representations using HMM is much more powerful for sensing, learning and inferencing human activities at multiple levels [5, 14]. Using Multi-level HMM, such as the Hierarchical Hidden Markov Model (HHMM) [6] allows for unsupervised discovery of structures at multiple level in video segmentation and activity recognition [20]. We will compare our approach with the existing multi-level HMM approach by means of system robustness and the cost of training.

5.2 Lifelog System

Different approaches have been used to implement a lifelog system. The MyLifeBits system [7] is designed to store and manage everything in a person's lifetime that can be captured in digital format. Its initial goal was to store all personal information found in PCs such as articles, video, office documents, email, keystrokes, and screen mouse clicks, etc. It then evolved into storing all ambient information of a person's daily life via a specialized camera device named SenseCam. MyLifeBits supports capture, storage, management, and retrieval of many media types, and its sophisticated database design is capable of storing a large volume of multimedia data. However, MyLifeBits only applies a basic metadata-based indexing approach which requires users to manually annotate most of the collected data in order to have meaningful search results. Our works address this issue well by providing a more effective indexing scheme which requires less user involvement and provides more meaningful search results by taking the "meaning" of the collected media into account. Another lifelog system implementation is discussed in [9]. This work focuses on real-time storage and retrieval of lifelog in a ubiquitous environment. The developed system supports semi-automatic activity analysis and provides an intuitive graphical interface for users to browse their lifelogs that correlates the space and temporal information of the displayed sensory data.

In addition to our language approach to indexing lifelog, Kim et al presented a multi-modal sensor fusion technique which supports automatic generation of lifelog's metadata [10]. The key idea is to combine the analysis results of different kinds of low-level sensory data to better infer higher-level context information about the collected lifelog. For example, by combining the analysis of audio, GPS, and accelerometer readings, the system is able to better identify the environment in which the lifelog was taken. Machine learning techniques such as decision tree and Gaussian Mixture Model (GMM) are used to analyze the collected low-level sensory data. Similar techniques to this sensor fusion approach are explored in [2, 18]. The former uses video key frame summarization and conversation scene detection to fulfill efficient lifelog retrieval. The latter proposes an integrated technique to process lifelog data using correlations between different types of the captured data from multiple sensors.

5.3 Geo-trace Pattern Recognition

Pattern detection in location traces is an area of research that has some previous work [8]. The majority of these research work focus on detecting short time duration patterns [8], where as our system uses frequently occurring short-time duration patterns to classify the similarity of larger time segments.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present a novel yet straightforward approach of processing lifelogs of heterogeneous sensory data. We verify the similarity between activity and language by demonstrating Zipf's distribution over our activity language corpus. The experimental results presented in Section 3.2 demonstrate high accuracy of using language models for human activity recognition. Unlike the traditional HMM approach which is limited to activity recognition task, modeling activity as language enables many other applications such as similar activity retrieval, and hierarchical activity segmentation. Besides, this language-based modeling approach can be applied to other types of sensory input such as geo-locations. We will conduct user study to evaluate the effectiveness of our similar activity retrieval service and

Segment length (hours)	0.25	0.50	0.75	1	2	4	6	8	10	12
Prediction accuracy (%)	60.58	62.16	63.78	66.37	77.42	81.09	86.61	86.46	91.43	91.67

Table 6: Variation of prediction accuracy with testing segment length.

extend our work of hierarchical activity segmentation to automatic lifelog summarization.

7. REFERENCES

[1] R. Aipperspach, E. Cohen, and J. Canny. Modeling human behavior from simple sensors in the home. In *Proceedings of IEEE Conf. on Pervasive Computing*, pages 337–348, Dublin, Ireland, April 2006.

[2] K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki. Efficient retrieval of life log based on context and content. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 22–31, New York, NY, USA, 2004. ACM.

[3] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of alzheimer's disease. *Alzheimer's and Dementia*, 3(3):186–191, July 2007. predicted 107 million people will suffer from Alzheimer by 2050.

[4] K. Burke. *Language as Symbolic Action*. University of California Press, 1966.

[5] B. Clarkson and A. Pentland. Unsupervised clustering of ambulatory audio and video. In *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, pages 3037–3040, Washington, DC, USA, 1999. IEEE Computer Society.

[6] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Mach. Learn.*, 32(1):41–62, 1998.

[7] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. Mylifebits: fulfilling the memex vision. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 235–238, New York, NY, USA, 2002. ACM.

[8] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339, New York, NY, USA, 2007. ACM.

[9] I. jae Kim, S. C. Ahn, and H. gon Kim. Personalized life log media system in ubiquitous environment. *Lecture Notes in Computer Science*, 4412, 2006.

[10] I.-J. Kim, S. C. Ahn, H. Ko, and H. G. Kim. Automatic lifelog media annotation based on heterogeneous sensor fusion. In *Proceedings of IEEE International Conference on Multi Sensor fu-sion and Integration for Intelligent systems*, Seoul, Korea, August 20-22 2008.

[11] U. Manber and G. Myers. Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993.

[12] L. Neergaard. Report: 35 million-plus worldwide have dementia. *Associate Press*, Sep. 21 2009.

[13] A. Nguyen, D. Moore, and I. McCowan. Unsupervised clustering of free-living human activities using ambulatory accelerometry. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 4895–4898, Lyon, France, Aug 22-26 2007.

[14] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 3, Washington, DC, USA, 2002. IEEE Computer Society.

[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Division, Thomas J. Watson Research Center, 2001.

[16] D. Patterson, L. Liao, D. Fox, and H. Kautz. Inferring high-level behavior from low-level sensors. In *Proceedings of the Fifth International Conference on Ubiquitous Computing (UbiComp 2003)*, pages 73–89, Seattle, Washington, October 12-15 2003.

[17] A. Stolcke. Srilm – an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver, CO, 2002.

[18] K. Takata, J. Ma, B. O. Apduhan, R. Huang, and Q. Jin. Modeling and analyzing individual's daily activities using lifelog. In *ICESS '08: Proceedings of the 2008 International Conference on Embedded Software and Systems*, pages 503–510, Washington, DC, USA, 2008. IEEE Computer Society.

[19] J. V. Wertsch. *Mind As Action*. Oxford University Press, USA, 1998.

[20] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Unsupervised discovery of multilevel statistical video structures using hierarchical hidden markov models. In *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03)*, pages 29–32, Washington, DC, USA, 2003. IEEE Computer Society.